STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 17, Number 4, December 2016

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Patryk Barszcz, P.Barszcz@stat.gov.pl
Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec, Technical Assistant
Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 585−588

FROM THE EDITOR

This fourth issue of the volume 17 of the *Statistics in Transition new series* completes the annual (quarterly) editorial cycle for the year 2016. This provides an opportunity for the Journal Editor to express deep gratefulness, also on behalf of all the editorial bodies – Editorial Board, Associate Editors, and the Editorial Office – to its all collaborators and contributors: authors and peer-reviewers, technical staff and other supporters for making it – actually a product of joint efforts of all persons involved – to be appearing on the regular basis while playing a role in the advancement of science of data production and use across disciplines and sectors. Special thanks go to the experts who have served as reviewers, devoting generously their time and expertise in order to ensure high scientific and quality standards of the papers published over the passing year - theirs names are listed in the acknowledgements. This is also an opportunity to acknowledge the help and advice we have been obtaining from the members of the aforementioned panels. And to announce an intention to include some new names from among the actively supporting us experts to the panel of Associate Editors, along with an appreciation of cooperation with our journal over many years of few persons who will leave the panel with the new year.

This issue has customary structure, being composed of the three major sections. But I am pleased to announce an innovation to be introduced with the next issue, which will rely on adding a new section 'Research Communicates'. It is envisioned as a possibility for presenting new conceptual or empirical results of an ongoing research work in the form of not fully completed research paper, sometimes even in its embryo stage, however, advanced enough to pass the peer-reviewing process.

The first section, devoted to sampling and estimation issues, contains two papers. It begins with **J. Subramani's** paper ***New Median Based Ratio Estimator For Estimation of the Finite Population Mean***, which deals with a new median based ratio estimator in the absence of an auxiliary variable. The bias and mean squared error of the proposed median based ratio estimator are obtained. The performance of the median based ratio estimator is compared with that of the SRSWOR sample mean, ratio estimator and linear regression estimator for certain natural population. It is shown from the numerical comparisons that the proposed median based ratio estimator outperforms the SRSWOR sample mean, ratio estimator and also the linear regression estimator.

In the second paper **Housila P. Singh** and **Surya K. Pal** propose ***A New Family of Estimators of the Population Variance Using Information on Population Variance of Auxiliary Variable in Sample Surveys***. Starting with

observation that there are several recently proposed classes of estimators (e.g., due to Sharma and Singh (2014) or Singh and Pal (2016)) the authors emphasize the need for checking their properties. They provide asymptotic expressions of bias and mean squared error (*MSE*) of the suggested family of estimators. Asymptotic optimum estimator (*AOE*) in the family of estimators is also identified. Some subclasses of estimators of the proposed family of estimators have been identified and some theoretical comparisons among the estimators are discussed in the paper.

The research articles section starts with **Elzbieta Gołata's** paper ***Shift in methodology and population census quality.*** The shift in methods embraces a move in conducting population census from a conventional enumeration through a sample survey and a mixed approach to administrative data. This paper compares two censuses which were conducted in Poland in 2002 and 2011; each of them presents by itself different case of the traditional method (2002 census) and the combined approach (2011 census), respectively. The quality of census data is discussed with essential aims and objectives to provide reliable information on the population age and sex structure in detailed territorial division. Quality assessment is provided for the whole country and at regional level. It starts with the consideration of coverage errors using multiple sources of data and non-matching methods - in particular: demographic analysis based on previous censuses, vital statistics and a comparison with other existing sources. Different cross-sections by sex, age and place of residence are considered. The questions of adequacy and divergence are discussed in the substantive terms.

Olusanya Elisa Olubusoye, Grace Oluwatoyin Korter and **Afees Adebare Salisu** present their results on ***Modelling Road Traffic Crashes Using Spatial Autoregressive Model with Additional Endogenous Variable.*** The authors construct a model based on a linear cross sectional Spatial Autoregressive (SAR) framework with additional endogenous variables, exogenous variables and SAR disturbances. The focus is on Road Traffic Crashes (RTC) in Oyo state, Nigeria. The number of RTC in each Local Government Area (LGA) of the state is the dependent variable. A weights matrix, travel density, land area and major road length of each LGA were used as exogenous variables and population was the IV. The objective was to determine the hotspots and examine whether the number of RTC cases in a given LGA is affected by the number of RTC cases of neighbouring LGAs and an instrumental variable. The hotspots include Oluyole, Ido, Akinyele, Egbeda, Atiba, Oyo East, and Ogbomosho South LGAs. The study concludes that the number of RTC in a given LGA is affected by the number of RTC in contiguous LGAs. The authors address some policy implication of their results, such as that road safety and security measures must be administered simultaneously to LGAs with high concentration of RTC and their neighbours to achieve significant remedial effect.

Kumar Prabhash, Vijay M Patil, Vanita Noronha, Amit Joshi and **Atanu Bhattacharjee** in the paper ***Bayesian Accelerated Failure Time and its***

Application in Chemotherapy Drug Treatment Trial propose an alternative to the Cox proportional hazards model (CPH), which is normally applied in clinical trial data analysis. Since this model can generate severe problems with breaking the proportion hazard assumption, the authors present an accelerated failure time (AFT) instead. The model can be used through consideration of different covariates of interest and random effects in each section. The model is simple to fit by using OpenBugs software and is revealed to be good for the Chemotherapy data. However, other model comparison tools can be used to compare the models in different computational platforms.

Henryk Gurgul's and **Marcin Suder's** paper ***Calendar And Seasonal Effects of Size of Withdrawals From ATMs Managed by Euronet*** analyses the calendar effects on withdrawals from Automated Teller Machines (ATMs), daily data, managed by the Euronet network for the period from January 2008 to March 2012. They focus on the identification of specific calendar and seasonal effects in the ATM cash withdrawal series of the company in the Polish provinces of Lesser Poland (Małopolska) and Subcarpathian (Podkarpackie). The results of the analysis show that withdrawals depend strongly on the day of the week. On Fridays more cash is withdrawn than on other days, and Saturdays and Sundays are the days of the week with the lowest level of withdrawals. In a month, it can be seen that cash withdrawals take place more often in the second and in the last weeks of the month. This observation suggests that withdrawals from ATMs can be related to the profile of wage withdrawals. In Poland in the public sector wages are paid at the beginning of the month, and in the private sector at the end of the month. The time series of withdrawals also reflect seasonality. The largest amounts are withdrawn in July, August and December. Reason for the increased demand for cash are the summer holidays and the Christmas season. The results reflect consumer habits which show substantial calendar and seasonal effects.

A set of the next three papers is a selection of the articles based on presentations given at the 34th Multivariate Statistical Analysis Conference being held at the University of Lodz in October 2015.

The paper by **Marta Dziechciarz-Duda** and **Józef Dziechciarz**, ***The Identification of Training Needs For Human Capital Quality Improvement in Poland – a Statistical Approach*** addresses the issues of practical importance concerning methods of assessing efficiency of the Competency Development Programme launched recently by the Ministry of Science and Higher Education. It embraces allocation of additional financial means for activities meant to prepare students with the so-called soft skills necessary in scientific careers and on the labour market. Courses developing skills such as team work ability, leadership, creativity, independent thinking and innovative approach to problem solving will be financed. A thorough analysis of needs starts with existing databases describing the quality of human capital in Poland in order to identify those competencies that graduates of universities are missing. The paper discusses possible statistical tools applicable for that purpose, from the simple descriptive statistics to advanced multivariate statistical analysis.

In the next paper, ***Interval Estimation of Higher Order Quantiles. Analysis of Accuracy of Selected Procedures*** by **Dorota Pekasiewicz** selected nonparametric and semiparametric estimation methods of higher orders quantiles are considered. The construction of nonparametric confidence intervals is based on order statistics of appropriate ranks from random samples or from generated bootstrap samples. Semiparametric bootstrap methods are characterized by double bootstrap simulations. The values of bootstrap sample below the prearranged threshold are generated by the empirical distribution and the values above this threshold are generated by the distribution based on the asymptotic properties of the tail of the random variable distribution. The results of the study allow one to draw conclusions about the effectiveness of the considered procedures and to compare these methods.

Grażyna Dehnel's paper ***M-Estimators in business statistics*** discusses one of the techniques that is meant to deal with outlying observations, namely M-estimation, from the evaluative perspective. Until recently the implementation of robust regression methods, such as M-estimation or MM-estimation, was limited due to their iterative nature. With advances in computing power and the growing availability of statistical packages, such as R and SAS, Stata, the applicability of robust regression methods has increased considerably. The M-estimation method is being assessed using data from a survey of small and medium-sized businesses. The comparison involves nine *M-estimators*, each based on a different weighting function. For instance, the largest gain in efficiency and robustness of *M-estimators* was obtained when Talworth's and Tukey's functions were used.

The last article, ***Informative versus Non-Informative Prior Distributions and their Impact on the Accuracy of Bayesian Inference*** by **Wioletta Grzenda** discusses the benefits arising from the use of the Bayesian approach to predictive modelling, along with exemplification using a linear regression model and a logistic regression model. The impact of informative and non-informative prior on model accuracy is systematically examined and compared. The data from the Central Statistical Office of Poland on unemployment by individual districts in Poland are used, and Markov Chain Monte Carlo methods (MCMC) was employed in modelling. These results indicate that the accuracy of models estimated with informative a priori distributions is higher. Therefore, when additional out-of-sample knowledge is available, the appropriate selection of a priori distribution can improve the accuracy of regression and classification models.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 589

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 591–604

A NEW MEDIAN BASED RATIO ESTIMATOR FOR ESTIMATION OF THE FINITE POPULATION MEAN

J. Subramani [1]

ABSTRACT

The present paper deals with a new median based ratio estimator for the estimation of finite population means in the absence of an auxiliary variable. The bias and mean squared error of the proposed median based ratio estimator are obtained. The performance of the median based ratio estimator is compared with that of the SRSWOR sample mean, ratio estimator and linear regression estimator for certain natural population. It is shown from the numerical comparisons that the proposed median based ratio estimator outperforms the SRSWOR sample mean, ratio estimator and also the linear regression estimator.

Key words: bias, linear regression estimator, mean squared error, natural population, simple random sampling.

1. Introduction

Let $U = \{U_1, U_2, \ldots, U_N\}$ be a finite population with N distinct and identifiable units. Let Y be the study variable with value Y_i measured on $U_i, i = 1, 2, 3, \ldots, N$ giving a vector $Y = \{Y_1, Y_2, \ldots, Y_N\}$. The problem is to estimate the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ with some desirable properties like unbiased, minimum variance of the estimators on the basis of a random sample of size n selected from the population U. Suppose that there exists an auxiliary variable X and it is positively correlated with the study variable Y, then one can propose estimators like ratio estimator, linear regression estimator and their modifications, which will perform better than the SRSWOR sample mean for estimating the population mean of the study variable as stated in Cochran (1977) and Murthy (1967). In the absence of the auxiliary variable, the above estimators are not possible. However one may think of getting additional information on the study variable and one can propose ratio and linear regression type estimators to improve the performance of the estimator. The idea of this paper is to use such information, namely the

[1] Department of Statistics, Pondicherry University, R. V. Nagar, Kalapet, Puducherry – 605 014, India. E-mail: drjsubramani@yahoo.co.in.

median of the study variable, in the proposed ratio estimator. It is reasonable to assume that the median of the study variable is known since this parameter does not require the complete information on the population units of the study variable unlike the other parameters like population mean, population variance, etc. In particular, in the surveys involving the estimation of average income, average marks, etc. it is very reasonable to assume that the population mean is unknown whereas the population median is known. For more details, consider the examples given below, in which the interests are to estimate the population mean of the respective populations.

Example 1. In an Indian university 5000 students entered for the university examination. The results are given below. The problem is to estimate the average marks scored by the students (population mean). Here, it is reasonable to assume that the median of the marks is known since we have the following information.

Table 1. Results of University Examination

Passed with	Percentage of marks	Number of students	Cumulative total
Distinction	[75-100]	850	850
First Class	[60-75)	3100	3950
Second Class	[50-60)	600	4550
Failed	[0-50)	450	5000
Total		5000	5000

The median value will be between 60 and 75. Approximately one can assume the population median value as 67.5.

Example 2. In an Indian university 800 faculty members are working in different categories and the basic salary drawn by different categories of the faculty members are given in Table 2. The problem is to estimate the average salary drawn by the faculty members (population mean) per month. Here, it is reasonable to assume that the median of the salary is known based on the information given in Table 2.

Table 2. Salary of University faculty members

Category	Basic Salary in Indian Rupees (IRs) Per month*	Number of faculty members	Cumulative total
Senior Professor	56000+10000**	20	20
Professor - Grade I	43000+10000	40	60
Professor - Grade II	37400+10000	60	120

Table 2. Salary of University faculty members (cont.)

Category	Basic Salary in Indian Rupees (IRs) Per month*	Number of faculty members	Cumulative total
Associate Professor - Grade I	37400+10000	80	200
Associate Professor - Grade II	37400+9000	100	300
Assistant Professor - Grade I	15100+8000	110	410
Assistant Professor - Grade II	15100+7000	140	550
Assistant Professor - Grade III	15100+6000	250	800
Total		800	800

*Actual salary depends on their experience in their designation and other allowances.

**The Basic salary is the sum of the basic (the first value) and the academic grade pay (the second value), which will differentiate people with same designation but different grades.

The population median value will be assumed as IRs. 15100+8000 = IRs. 23100.

Example 3. In the estimation of body mass index (BMI) of the 350 patients of a Hospital, it is reasonable to assume that the population median of the BMI is known based on the information given in Table 3.

Table 3. Body mass index of 350 patients of a hospital

Category	BMI range – kg/m^2	Number of patients	Cumulative total
Very severely underweight	less than 15	15	15
Severely underweight	from 15.0 to 16.0	35	50
Underweight	from 16.0 to 18.5	67	117
Normal (healthy weight)	from 18.5 to 25	92	209
Overweight	from 25 to 30	47	256
Obese Class I (Moderately obese)	from 30 to 35	52	308
Obese Class II (Severely obese)	from 35 to 40	27	335
Obese Class III (Very severely obese)	over 40	15	350
Total		350	350

The median value will be between 18.5 and 25. Approximately one can assume the population median of the BMI value as 21.75.

Example 4. In the problem of estimating the blood pressure of the 202 patients of a hospital, it is reasonable to assume that the median of the blood pressure is known based on the information available in Table 4.

Table 4. Blood pressure of 202 patients of a hospital

Category	Systolic, mmHg	Number of patients	Cumulative No. of patients
Hypotension	< 90	10	10
Desired	90–119	112	122
Pre-hypertension	120–139	40	162
Stage 1 Hypertension	140–159	20	182
Stage 2 Hypertension	160–179	13	195
Hypertensive Emergency	≥ 180	7	202
Total		202	202

The median value will be between 90 and 119. Approximately one can assume the population median value as 104.5.

Before discussing further about the existing estimators and the proposed median based ratio estimator the notations and the formulae to be used in this paper are described below:

- N - Population size
- n - Sample size
- $N_{C_n} = \binom{N}{n}$ - Number of possible samples of size n from the population of size N
- Y - Study variable
- M - Median of the Study variable
- X - Auxiliary variable
- $\overline{X}, \overline{Y}$ - Population means
- $\bar{x}, \bar{y}$ - Sample means
- ρ - Correlation coefficient between X and Y
- β - Regression coefficient of Y on X
- $\overline{M}$ - Average of sample medians of Y
- m - Sample median of Y
- B(.) - Bias of the estimator
- V(.) - Variance of the estimator
- MSE(.) - Mean squared error of the estimator
- $\mathrm{PRE}(e, p) = \frac{\mathrm{MSE}(e)}{\mathrm{MSE}(p)} * 100$ – Percentage relative efficiency of the proposed estimator(p) with respect to the existing estimator (e)

The formulae for computing various measures including the variance and the covariance of the SRSWOR sample mean and sample median are as follows:

$$V(\bar{y}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{y}_i - \bar{Y})^2 = \frac{1-f}{n} S_y^2 \, , V(\bar{x}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{x}_i - \bar{X})^2 = \frac{1-f}{n} S_x^2,$$

$$MSE(m) = V(m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - M)^2,$$

$$Cov(\bar{y}, \bar{x}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{x}_i - \bar{X})\,(\bar{y}_i - \bar{Y}) = \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^{N} (\bar{x}_i - \bar{X})\,(\bar{y}_i - \bar{Y}) \, ,$$

$$Cov(\bar{y}, m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - M)\,(\bar{y}_i - \bar{Y}) \text{ where } f = \frac{n}{N};\; S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2,\; C'_{xx} = \frac{V(\bar{x})}{\bar{X}^2}, C'_{mm} = \frac{V(m)}{M^2}, C'_{ym} = \frac{Cov(\bar{y},m)}{M\bar{Y}}, C'_{yx} = \frac{Cov(\bar{y},\bar{x})}{\bar{X}\bar{Y}}$$

If there is no auxiliary variable available, the simplest estimator of population mean $\bar{Y}$ is the sample mean $\bar{y}$ of size n drawn by using simple random sampling without replacement. The variance of the SRSWOR sample mean $\hat{\bar{Y}}=\hat{\bar{Y}}_r$ is given by

$$V(\hat{\bar{Y}}_r) = \frac{1-f}{n} S_y^2 \tag{1}$$

The ratio estimator for estimating the population mean $\bar{Y}$ of the study variable Y is defined as $\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R}\bar{X}$. The bias and the mean squared error are as given below:

$$B\left(\hat{\bar{Y}}_R\right) = \bar{Y}\{C'_{xx} - C'_{yx}\} \tag{2}$$

$$MSE\left(\hat{\bar{Y}}_R\right) = V(\bar{y}) + R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x}), \text{ where } R = \frac{\bar{Y}}{\bar{X}} \tag{3}$$

The other important and optimum estimator for estimating the population mean $\bar{Y}$ of the study variable Y using the auxiliary information is the linear regression estimator. The linear regression estimator and its variance are given below:

$$\hat{\bar{Y}}_{lr} = \bar{y} + \beta(\bar{X} - \bar{x}) \tag{4}$$

$$V\left(\hat{\bar{Y}}_{lr}\right) = V(\bar{y})(1 - \rho^2) \text{ where } \rho = \frac{Cov(\bar{y},\bar{x})}{\sqrt{V(\bar{x}) * V(\bar{y})}} \tag{5}$$

The ratio estimator has been extended and introduced several modified ratio estimators in the literature if the population parameters like coefficient of variation, skewness, kurtosis, correlation coefficient, quartiles, etc. of the auxiliary variable are known. For further details on the modified ratio estimators with the known population parameters of the auxiliary variable such as coefficient of variation, skewness, kurtosis, correlation coefficient, quartiles and their linear

combinations, the readers are referred to see the following papers: Kadilar and Cingi (2004, 2006a,b, 2009), Koyuncu and Kadilar (2009), Singh and Kakran (1993), Singh and Tailor (2003, 2005), Singh (2003), Sisodia and Dwivedi (1981), Subramani (2013), Subramani and Kumarapandiyan (2012a,b,c, 2013a,b), Tailor and Sharma (2009), Tin (1965), Yan and Tian (2010) and the references cited therein.

In general it has been established that the ratio estimator and the linear regression estimator performs better than the SRSWOR sample mean under certain conditions, of which that provided the auxiliary information $\overline{X}$ is known. It is to be noted that both the ratio and regression estimators use the population mean of the auxiliary variable as auxiliary information. In the absence of the auxiliary variable X, these estimators are not possible. This point is motivated to look for an alternative method for this problem. Further, it is observed from various discussions and studies that the median of the study variable may be known well in advance in several situations. Hence, an attempt is made in this paper to propose a ratio estimator making use of the population median of the study variable as auxiliary information and as a result a new ratio estimator, namely median based ratio estimator has been proposed. The performance of the median based ratio estimator has been compared with that of the SRSWOR sample mean, ratio estimator and linear regression estimator for certain natural populations. It is shown that the proposed median based ratio estimator outperformed not only the SRSWOR sample mean, ratio estimator but also the linear regression estimator. The proposed median based ratio estimator together with the bias and mean squared error are given in section 2.

2. Proposed median based ratio estimator

In this section a new median based ratio estimator for estimating population mean $\overline{Y}$ has been proposed if the median M of the study variable *Y* is known. The proposed median based ratio estimator together with the bias and mean squared error are given below:

$$\hat{\overline{Y}}_M = \bar{y}\left[\frac{M}{m}\right] \tag{6}$$

$$B\left(\hat{\overline{Y}}_M\right) = \overline{Y}\left\{C'_{mm} - C'_{ym} - \frac{Bias(m)}{M}\right\} \tag{7}$$

$$MSE\left(\hat{\overline{Y}}_M\right) = V(\bar{y}) + R'^2 V(m) - 2R'Cov(\bar{y}, m) \text{ where } R' = \frac{\overline{Y}}{M} \tag{8}$$

The detailed derivations of the bias and the mean squared error are given below:

Consider $\hat{\overline{Y}}_M = \bar{y}\left[\frac{M}{m}\right]$

Let $e_0 = \frac{\bar{y}-\overline{Y}}{\overline{Y}}$ and $e_1 = \frac{m-M}{M}$

$$\Longrightarrow E(e_0) = 0$$
$$\Longrightarrow E(e_1) = \frac{\bar{M}-M}{M} = \frac{Bias(m)}{M}$$
$$\Longrightarrow E(e_0^2) = \frac{V(\bar{y})}{\bar{Y}^2}; E(e_1^2) = \frac{V(m)}{M^2}\ ; E(e_0 e_1) = \frac{Cov(\bar{y},m)}{\bar{Y}M}$$

The estimator $\hat{\bar{Y}}_M$ can be written in terms of e_0 and e_1 as

$$\hat{\bar{Y}}_M = \bar{Y}(1+e_0)\left(\frac{M}{M(1+e_1)}\right)$$
$$\Longrightarrow \hat{\bar{Y}}_M = \bar{Y}(1+e_0)\left(\frac{1}{1+e_1}\right)$$
$$\Longrightarrow \hat{\bar{Y}}_M = \bar{Y}(1+e_0)(1+e_1)^{-1}$$

It is worth to mention that the computation of bias and mean squared error is based on the Taylor series of the function $1/(1+x)$, rounded up to second order term and neglecting the terms $-2(\bar{Y}e_0 - \bar{Y}e_1)(\bar{Y}e_0e_1 - \bar{Y}e_1^2)$ and $(\bar{Y}e_0e_1 - \bar{Y}e_1^2)^2$.

We have

$$\hat{\bar{Y}}_M = \bar{Y}(1+e_0)(1-e_1+e_1^2)$$
$$\Longrightarrow \hat{\bar{Y}}_M = \bar{Y} + \bar{Y}e_0 - \bar{Y}e_1 - \bar{Y}e_0e_1 + \bar{Y}e_1^2$$
$$\Longrightarrow \hat{\bar{Y}}_M - \bar{Y} = \bar{Y}e_0 - \bar{Y}e_1 - \bar{Y}e_0e_1 + \bar{Y}e_1^2$$

Taking expectations on both sides we have

$$E(\hat{\bar{Y}}_M - \bar{Y}) = \bar{Y}E(e_0) - \bar{Y}E(e_1) - \bar{Y}E(e_0e_1) + \bar{Y}E(e_1^2)$$
$$\Longrightarrow E\left(\hat{\bar{Y}}_M - \bar{Y}\right) = \bar{Y}\{E(e_1^2) - E(e_0e_1) - E(e_1)\}$$
$$\Longrightarrow Bias(\hat{\bar{Y}}_M) = \bar{Y}\left\{\frac{V(m)}{M^2} - \frac{Cov(\bar{y},m)}{\bar{Y}M} - \frac{\bar{M}-M}{M}\right\}$$
$$\Longrightarrow Bias(\hat{\bar{Y}}_M) = \bar{Y}\left\{C'_{mm} - C'_{ym} - \frac{Bias(m)}{M}\right\}$$

The mean squared error of $\hat{\bar{Y}}_M$ is obtained as given below:

$$MSE\left(\hat{\bar{Y}}_M\right) = E(\hat{\bar{Y}}_M - \bar{Y})^2 = E(\bar{Y}e_0 - \bar{Y}e_1)^2$$
$$\Longrightarrow MSE\left(\hat{\bar{Y}}_M\right) = \bar{Y}^2\{E(e_0^2) + E(e_1^2) - 2E(e_0e_1)\}$$
$$\Longrightarrow MSE\left(\hat{\bar{Y}}_M\right) = \bar{Y}^2\left\{\frac{V(\bar{y})}{\bar{Y}^2} + \frac{V(m)}{M^2} - 2\frac{Cov(\bar{y},m)}{\bar{Y}M}\right\}$$
$$\Longrightarrow MSE\left(\hat{\bar{Y}}_M\right) = V(\bar{y}) + \frac{\bar{Y}^2}{M^2}V(m) - 2\frac{\bar{Y}}{M}Cov(\bar{y},m)$$
$$\Longrightarrow MSE\left(\hat{\bar{Y}}_M\right) = V(\bar{y}) + R'^2V(m) - 2R'Cov(\bar{y},m)$$

Remark 2.1. When $\bar{M} > M$, the use of the average of the sample medians $\bar{M}$ in the median based ratio estimator in the place of population median M reduces the variance of the proposed estimator due to the following reasons:

1. In general, the sample median m is not an unbiased estimator for the population median M, which leads the mean squared error of m and the value is larger than the variance of m.
2. The covariance $\mathrm{Cov}(\bar{y}, m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - \bar{M})(\bar{y}_i - \bar{Y})$ based on $\bar{M}$ and the covariance $\mathrm{Cov}(\bar{y}, m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - M)(\bar{y}_i - \bar{Y})$ based on the population median M are the same.

Remark 2.2. When n = 2 the sample mean becomes the sample median, i.e. $\bar{y} = m$, the mean squared error of $\hat{\bar{Y}}_M = \mathrm{MSE}\left(\hat{\bar{Y}}_M\right) = 0$ and hence the SRSWOR becomes the trend free sampling (see Mukerjee and Sengupta (1990)).

3. Efficiency comparison

In this section we have derived the algebraic conditions for which the proposed median based ratio estimator will have minimum mean squared error compared to the SRSWOR sample mean, ratio estimator and linear regression estimator for estimating the finite population mean.

3.1. Comparison with that of SRSWOR sample mean

From the expressions given in (8) and (1), the conditions for which the proposed estimator $\hat{\bar{Y}}_M$ is more efficient than the existing estimator $\hat{\bar{Y}}_r$ are derived and given below:

$$\mathrm{MSE}\left(\hat{\bar{Y}}_M\right) \leq \mathrm{V}\left(\hat{\bar{Y}}_r\right), \text{if } 2C'_{ym} \geq C'_{mm} \tag{9}$$

The detailed derivation is given below:

Consider $\mathrm{MSE}\left(\hat{\bar{Y}}_M\right) \leq \mathrm{V}\left(\hat{\bar{Y}}_r\right)$

$$\Rightarrow \mathrm{V}(\bar{y}) + R'^2\,\mathrm{V}(m) - 2R'\,\mathrm{Cov}(\bar{y}, m) \leq \mathrm{V}(\bar{y})$$

$$\Rightarrow R'^2\,\mathrm{V}(m) - 2R'\,\mathrm{Cov}(\bar{y}, m) \leq 0$$

$$\Rightarrow R'^2\mathrm{V}(m) \leq 2R'\,\mathrm{Cov}(\bar{y}, m)$$

$$\Rightarrow \mathrm{Cov}(\bar{y}, m) \geq \frac{R'\,\mathrm{V}(m)}{2}$$

$$\Rightarrow \mathrm{Cov}(\bar{y}, m) \geq \frac{\bar{Y} M C'_{mm}}{2}$$

$$\Rightarrow 2C'_{ym} - C'_{mm} \geq 0$$

3.2. Comparison with that of Ratio Estimator

From the expressions given in (8) and (3) the conditions for which the proposed estimator $\hat{\bar{Y}}_M$ is more efficient than the usual ratio estimator $\hat{\bar{Y}}_R$ are derived and given below:

$$\mathrm{MSE}\left(\hat{\bar{Y}}_M\right) \leq \mathrm{MSE}\left(\hat{\bar{Y}}_R\right), \text{if } C'_{mm} - C'_{xx} \leq 2\left\{C'_{ym} - C'_{xy}\right\} \tag{10}$$

The detailed derivation is given below:

Consider $\mathrm{MSE}(\hat{\bar{Y}}_M) \leq \mathrm{MSE}(\hat{\bar{Y}}_R)$

$$\begin{aligned}
&\Rightarrow V(\bar{y}) + R'^2 V(m) - 2R'Cov(\bar{y}, m) \\
&\qquad \leq V(\bar{y}) + R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x}) \\
&\Rightarrow R'^2 V(m) - 2R'Cov(\bar{y}, m) \leq R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x}) \\
&\Rightarrow R'^2 V(m) - R^2 V(\bar{x}) \leq 2R'Cov(\bar{y}, m) - 2RCov(\bar{y}, \bar{x}) \\
&\Rightarrow \frac{\bar{Y}^2}{M^2} V(m) - \frac{\bar{Y}^2}{\bar{X}^2} V(\bar{x}) \leq 2\frac{\bar{Y}}{M} Cov(\bar{y}, m) - 2\frac{\bar{Y}}{\bar{X}} Cov(\bar{y}, \bar{x}) \\
&\Rightarrow \frac{V(m)}{M^2} - \frac{V(\bar{x})}{\bar{X}^2} \leq 2\left\{\frac{Cov(\bar{y}, m)}{\bar{Y}M} - \frac{Cov(\bar{y}, \bar{x})}{\bar{Y}\bar{X}}\right\} \\
&\Rightarrow 2C'_{ym} - C'_{mm} \geq 2C'_{xy} - C'_{xx}
\end{aligned}$$

3.3. Comparison with that of Linear Regression Estimator

From the expressions given in (8) and (5) the conditions for which the proposed estimator $\hat{\bar{Y}}_M$ is more efficient than the usual linear regression estimator $\hat{\bar{Y}}_{lr}$ are derived and given below:

$$\mathrm{MSE}\left(\hat{\bar{Y}}_M\right) \leq V\left(\hat{\bar{Y}}_{lr}\right), \text{if } 2\,C'_{ym} - \,C'_{mm} \geq \frac{\left[C'_{yx}\right]^2}{C'_{xx}} \tag{11}$$

The detailed derivation is given below:

Consider $\mathrm{MSE}(\hat{\bar{Y}}_M) \leq V(\hat{\bar{Y}}_{lr})$

$$\begin{aligned}
&\Rightarrow V(\bar{y}) + R'^2 V(m) - 2R'Cov(\bar{y}, m) \leq V(\bar{y})(1 - \rho^2) \\
&\Rightarrow R'^2 V(m) - 2R'Cov(\bar{y}, m) \leq -V(\bar{y})\left(\frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x}) * V(\bar{y})}\right) \\
&\Rightarrow 2R'Cov(\bar{y}, m) - R'^2 V(m) \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})} \\
&\Rightarrow 2\frac{\bar{Y}}{M}\, Cov(\bar{y}, m) - \frac{\bar{Y}^2}{M^2}\, V(m) \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})} \\
&\Rightarrow 2\bar{Y}^2 C'_{ym} - \bar{Y}^2 C'_{mm} \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})} \\
&\Rightarrow 2\,C'_{ym} - \,C'_{mm} \geq \frac{\left[C'_{yx}\right]^2}{C'_{xx}}
\end{aligned}$$

Remark 3.1. It is well known that the ratio estimator is more efficient than the SRSWOR sample mean if $2C'_{xy} - C'_{xx} \geq 0$. Similar results hold good here. That is, the median based ratio estimator is more efficient than

(i) SRSWOR sample mean if $2C'_{ym} - C'_{mm} \geq 0$

(ii) Ratio estimator if $2C'_{ym} - C'_{mm} \geq 2C'_{xy} - C'_{xx}$

(iii) Linear regression estimator if $2C'_{ym} - C'_{mm} \geq \frac{[C'_{yx}]^2}{C'_{xx}}$

4. Numerical study

In section 3, the conditions are derived for which the proposed median based ratio estimator performed better than the SRSWOR sample mean, ratio estimator and linear regression estimator. However, it has not been proved explicitly by algebraic expressions that the proposed estimators are better than the estimators mentioned above. Alternatively one has to resort for numerical comparisons to determine the efficiencies of the proposed estimators. In this view, three natural populations available in the literature are used for comparing the efficiencies of the proposed median based ratio estimators with that of the existing estimators. The population 1 and 2 are taken from Daroga Singh and Chaudhary (1986, page no. 177) and population 3 is taken from Mukhopadhyay (2005, page no. 96). The populations 1 and 2 pertain to estimate the area of cultivation under wheat in the year 1974, whereas the auxiliary information is the cultivated areas under wheat in 1971 and 1973 respectively. The population 3 pertains to the number of labourers (the auxiliary variable, in thousand) and the quantity of raw materials (study variable, in lakhs of bales) for 20 jute mills. The parameter values and constants computed for the above populations are presented in Table 5; the bias for the proposed modified ratio estimators and the existing estimators computed for the three populations discussed above are presented in Table 6, whereas the mean squared errors are presented in Table 7. It is worth noting that since the computation of mean squared error is involved in computing all possible N_{C_n} samples of size n, the computation requires strong capacity, which, for very large samples, exceeds the possibilities of the computation facilities. Hence only small sample sizes are considered. However, the same results will hold good for large sample sizes too.

Table 5. Parameter values and constants computed from the 3 populations

Para-meters	For sample size $n = 3$			For sample size $n = 5$		
	Popln-1	**Popln-2**	**Popln-3**	**Popln-1**	**Popln-2**	**Popln-3**
N	34	34	20	34	34	20
n	3	3	3	5	5	5
N_{C_n}	5984	5984	1140	278256	278256	15504
$\bar{Y}$	856.4118	856.4118	41.5	856.4118	856.4118	41.5

Table 5. Parameter values and constants computed from the 3 populations (cont.)

Para-meters	For sample size $n = 3$			For sample size $n = 5$		
	Popln-1	**Popln-2**	**Popln-3**	**Popln-1**	**Popln-2**	**Popln-3**
$\bar{M}$	747.7223	747.7223	40.2351	736.9811	736.9811	40.0552
M	767.5	767.5	40.5	767.5	767.5	40.5
$\bar{X}$	208.8824	199.4412	441.95	208.8824	199.4412	441.95
R	4.0999	4.2941	0.0939	4.0999	4.2941	0.0939
R'	1.1158	1.1158	1.0247	1.1158	1.1158	1.0247
$V(\bar{y})$	163356.4086	163356.4086	27.1254	91690.3713	91690.3713	14.3605
$V(\bar{x})$	6884.4455	6857.8555	2894.3089	3864.1726	3849.248	1532.2812
V(m)	101518.7738	101518.7738	26.1307	59396.2836	59396.2836	10.8348
$Cov(\bar{y}, m)$	90236.2939	90236.2939	21.0918	48074.9542	48074.9542	9.0665
$Cov(\bar{y}, \bar{x})$	15061.4011	14905.0488	182.7425	8453.8187	8366.0597	96.7461
ρ	0.4491	0.4453	0.6522	0.4491	0.4453	0.6522

Table 6. Bias of the existing and proposed estimators

Estimators	For sample size $n = 3$			For sample size $n = 5$		
	Popln-1	**Popln-2**	**Popln-3**	**Popln-1**	**Popln-2**	**Popln-3**
$\hat{\bar{Y}}_R$	63.0241	72.9186	0.2015	35.3748	40.9285	0.1067
$\hat{\bar{Y}}_M$	52.0924	52.0924	0.4118	57.7705	57.7705	0.5061

Table 7. Variance / Mean squared error of the existing and proposed estimators

Estimators		For sample size $n = 3$			For sample size $n = 5$		
		Popln-1	**Popln-2**	**Popln-3**	**Popln-1**	**Popln-2**	**Popln-3**
Existing	$\hat{\bar{Y}}_r$	163356.4086	163356.4086	27.1254	91690.3713	91690.3713	14.3605
	$\hat{\bar{Y}}_R$	155579.7064	161801.6355	18.3265	87325.3836	90817.6922	9.7023
	$\hat{\bar{Y}}_{lr}$	130405.9256	130961.3720	15.5873	73195.5841	73508.8959	8.2521
Proposed	$\hat{\bar{Y}}_M$	88379.0666	88379.0666	11.3372	58356.9234	58356.9234	7.1563

The percentage relative efficiencies of the proposed estimator with respect to the existing estimators are also obtained and given in the following table:

Table 8. PRE of the proposed estimator $\hat{\bar{Y}}_M$ with respect to existing estimators

Existing Estimators	For sample size $n = 3$			For sample size $n = 5$		
	Popln-1	**Popln-2**	**Popln-3**	**Popln-1**	**Popln-2**	**Popln-3**
$\hat{\bar{Y}}_r$	184.84	184.84	239.26	157.12	157.12	200.67
$\hat{\bar{Y}}_R$	176.04	183.08	161.65	149.64	155.62	135.58
$\hat{\bar{Y}}_{lr}$	147.55	148.18	137.49	125.43	125.96	115.31

From the Table 8 it is observed that the percentage relative efficiencies of the proposed estimator with respect to existing estimators are in general ranging from 115.31 to 239.26. Particularly, the PRE is ranging from 157.12 to 239.26 for comparison with the SRSWOR sample mean; ranging from 135.58 to 183.08 for comparison with ratio estimator; ranging from 115.31 to 148.18 for comparison with linear regression estimator. This shows that the proposed estimator performs better than the existing SRSWOR sample mean, ratio and linear regression estimator for all the three populations considered here. Further, it is observed from the numerical comparisons that the following inequalities hold good.

$$\mathrm{MSE}(\hat{\bar{Y}}_M) \leq \mathrm{V}(\hat{\bar{Y}}_{lr}) \leq \mathrm{MSE}\left(\hat{\bar{Y}}_R\right) \leq \mathrm{V}\left(\hat{\bar{Y}}_r\right)$$

5. Conclusion

This paper deals with a new median based ratio estimator for estimation of the finite population mean. The conditions are derived for which the proposed estimator is more efficient than the existing estimators. Further, it is shown that the percentage relative efficiencies of the proposed estimators with respect to existing estimators are in general ranging from 115.31 to 239.26 for certain natural populations available in the literature. It is usually believed that the linear regression estimator is the best linear unbiased estimator or the optimum estimator for estimating the population mean whenever there exists an auxiliary variable, which is positively correlated with that of the study variable. However, it is shown that the proposed median based ratio estimator outperformed not only the SRSWOR sample mean, ratio estimator but also the linear regression estimator. Hence, the proposed modified ratio estimators are recommended for the practical applications. Further, it is to be noted that the median based ratio estimator can be easily extended to median based modified ratio estimators in line with the modified ratio estimators available in the literature and one of the author's research students Prabavathy, G is working at present in this direction.

Acknowledgement

The author is very thankful to the Reviewers and the Editor for their constructive comments helpful in improving the presentation of this paper.

REFERENCES

COCHRAN, W. G., (1977). Sampling techniques, Third Edition, Wiley Eastern Limited, USA.

KADILAR, C., CINGI, H., (2004). Ratio estimators in simple random sampling, Applied Mathematics and Computation, 151, pp. 893–902.

KADILAR, C., CINGI, H., (2006a). An improvement in estimating the population mean by using the correlation co-efficient, Hacettepe Journal of Mathematics and Statistics, 35 (1), pp. 103–109.

KADILAR, C., CINGI, H., (2006b). Improvement in estimating the population mean in simple random sampling, Applied Mathematics Letters, 19, pp. 75–79.

KADILAR, C., CINGI, H., (2009). Advances in sampling theory - Ratio method of estimation, Bentham Science Publishers.

KOYUNCU, N., KADILAR, C., (2009). Efficient estimators for the population mean, Hacettepe Journal of Mathematics and Statistics, 38(2), pp. 217–225.

MUKERJEE, R., SENGUPTA, S., (1990). Optimal estimation of a finite population mean in the presence of linear trend, Biometrika, 77, pp. 625–630.

MUKHOPADHYAY, P., (2005). Theory and methods of survey sampling, PHI Learning, 2nd edition, New Delhi.

MURTHY, M. N., (1967). Sampling theory and methods, Statistical Publishing Society, Calcutta, India.

SINGH, G.N., (2003). On the improvement of product method of estimation in sample surveys, Journal of Indian Society of Agricultural Statistics, 56 (3), pp. 267–265.

SINGH, D., CHAUDHARY, F. S., (1986). Theory and analysis of sample survey designs, New Age International Publisher, New Delhi.

SINGH, H. P., KAKRAN, M., (1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character, Revised version submitted to Journal of Indian Society of Agricultural Statistics.

SINGH, H. P., TAILOR, R., (2003). Use of known correlation coefficient in estimating the finite population means, Statistics in Transition, 6 (4), pp. 555–560.

SINGH, H. P., TAILOR, R., (2005). Estimation of finite population mean with known co-efficient of variation of an auxiliary variable, Statistica, anno LXV, 3, pp. 301–313.

SISODIA, B. V. S., DWIVEDI, V. K., (1981). A modified ratio estimator using co-efficient of variation of auxiliary variable, Journal of the Indian Society of Agricultural Statistics, 33 (2), pp. 13–18.

SUBRAMANI, J., (2013). Generalized modified ratio estimator of finite population mean, Journal of Modern Applied Statistical Methods, 12 (2), pp. 121–155.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012a). Estimation of population mean using coefficient of variation and median of an auxiliary variable, International Journal of Probability and Statistics, 1 (4), pp. 111–118.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012b). Modified ratio estimators using known median and coefficient of kurtosis, American Journal of Mathematics and Statistics, 2 (4), pp. 95–100.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012c). Estimation of population mean using known median and coefficient of skewness, American Journal of Mathematics and Statistics, 2 (5), pp. 101–107.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2013a). Estimation of population mean using deciles of an auxiliary variable, Statistics in Transition-New Series, 14 (1), pp. 75–88.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2013b). A new modified ratio estimator of population mean when median of the auxiliary variable is known, Pakistan Journal of Statistics and Operation Research, Vol. 9 (2), pp. 137–145.

TAILOR, R., SHARMA, B., (2009). A modified ratio-cum-product estimator of finite population mean using known coefficient of variation and coefficient of kurtosis, Statistics in Transition-New Series, 10 (1), pp. 15–24.

TIN, M., (1965). Comparison of some ratio estimators, Journal of the American Statistical Association, 60, pp. 294–307.

YAN, Z., TIAN, B., (2010). Ratio method to the mean estimation using coefficient of skewness of auxiliary variable, ICICA 2010, Part II, CCIS 106, pp. 103–11.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 605–630

A NEW FAMILY OF ESTIMATORS OF THE POPULATION VARIANCE USING INFORMATION ON POPULATION VARIANCE OF AUXILIARY VARIABLE IN SAMPLE SURVEYS

Housila P. Singh[1], **Surya K. Pal**[2]

ABSTRACT

This paper proposes a family of estimators of population variance S_y^2 of the study variable *y* in the presence of known population variance S_x^2 of the auxiliary variable *x*. It is identified that in addition to many, the recently proposed classes of estimators due to Sharma and Singh (2014) and Singh and Pal (2016) are members of the proposed family of estimators. Asymptotic expressions of bias and mean squared error (*MSE*) of the suggested family of estimators have been obtained. Asymptotic optimum estimator (*AOE*) in the family of estimators is identified. Some subclasses of estimators of the proposed family of estimators have been identified along with their properties. We have also given the theoretical comparisons among the estimators discussed in this paper.
ASM Classification: 62D05.

Key words: Auxiliary variable, Study variable, Bias, Mean squared error, Efficiency comparison.

1. Introduction

The problem of estimating the population variance assumes importance in various fields such as industry, agriculture, medical and biological sciences etc. In sample surveys, auxiliary information on the finite population under investigation is quite often available from previous experience, census or administrative databases. It is well known that the auxiliary information in the theory of sampling is used to increase the efficiency of the estimators of the parameters such as mean or total, variance, coefficient of variation etc. Out of many, ratio and regression methods of estimation are good examples in this context. In many

[1] School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India.
[2] School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India.
E-mail: suryakantpal6676@gmail.com

situations of practical importance, the problem of estimating the population variance S_y^2 of the study variable *y* deserves special attention. When the population parameters such as population mean, variance, coefficients of skewness and kurtosis of the auxiliary variable are known, several authors including Das and Tripathi (1978), Srivastava and Jhajj (1980), Isaki (1983), Prasad and Singh (1990, 1992), Kadilar and Cingi (2006), Shabbir and Gupta (2007), Gupta and Shabbir (2008), Singh and Solanki (2013a, b), Solanki and Singh(2013), Singh et al. (2013, 2014), Hilal et al. (2014), Sharma and Singh (2014), Solanki et al. (2015), Yadav et al. (2015) and Singh and Pal (2016) etc. have suggested various estimators and studied their properties.

The principal aim of this paper is to suggest a new family of estimators of the population variance S_y^2 of the study variable *y* using information on population variance S_x^2 of the auxiliary variable *x* along with its properties under large sample approximation.

Consider a finite population $U=\{U_1,U_2,...,U_N\}$ of *N* units. Let *y* and *x* be the study and auxiliary variates respectively. We define the following parameters of the variates *y* and *x*:

Population mean: $\bar{Y}=N^{-1}\sum_{i=1}^{N}y_i$,

Population mean: $\bar{X}=N^{-1}\sum_{i=1}^{N}x_i$,

Population variance /mean square: $S_y^2=(N-1)^{-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$,

Population variance /mean square: $S_x^2=(N-1)^{-1}\sum_{i=1}^{N}(x_i-\bar{X})^2$.

Q_i is the i^{th} quartile (*i*=1, 2, 3) of the auxiliary variable x,

$Q_r=Q_3-Q_1$: the population inter-quartile range of the auxiliary variable x,

$Q_d=(Q_3-Q_1)/2$: the population semi-quartile range of the auxiliary variable x,

$Q_a=(Q_3+Q_1)/2$: the population semi-quartile average of the auxiliary variable x.

It is desired to estimate the population variance S_y^2 of the study variable *y* when the population variance S_x^2 of the auxiliary variable *x* is known. For estimating population variance S_y^2, a simple random sample (*SRS*) of size *n* is drawn without replacement (*WOR*) from the population *U*. The conventional

unbiased estimators of the population parameters $(\bar{Y}, S_y^2, \ \bar{X}, S_x^2)$ are respectively defined by

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i \,, s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \,, \bar{x} = n^{-1}\sum_{i=1}^{n} x_i$$

$$\text{and } s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,.$$

When the population variance S_x^2 of the auxiliary variable *x* is known, Isaki (1983) suggested a ratio-type estimator for estimating population variance S_y^2 defined by

$$t_1 = s_y^2\left(\frac{S_x^2}{s_x^2}\right). \tag{1.1}$$

Upadhyaya and Singh (1986) suggested an alternative estimator for S_y^2 as

$$t_2 = s_y^2\left(\frac{s_x^{*2}}{S_x^2}\right), \tag{1.2}$$

where $s_x^{*2} = (NS_x^2 - ns_x^2)/(N-n)$

$$= \{(1+g)S_x^2 - gs_x^2\}$$

and $g = n/(N-n)$.

Das and Tripathi (1978) suggested a difference-type estimator for the population variance S_y^2 as

$$t_3 = s_y^2 + d(S_x^2 - s_x^2)\,, \tag{1.3}$$

where '*d*' is suitable chosen constant.

Shabbir (2006) suggested a class of estimators of S_y^2 as

$$t_4 = \eta s_y^2 + (1-\eta)s_y^2\left(\frac{s_x^{*2}}{S_x^2}\right), \tag{1.4}$$

where η being suitable chosen constant.

It is well known that the estimator $t_0 = s_y^2$ is an unbiased estimator of S_y^2. The variance *MSE* of s_y^2 under *SRSWOR* to the first degree of approximation is given by

$$Var(t_0) = MSE(t_0) = f\, S_y^4(\lambda_{40} - 1), \tag{1.5}$$

where $\lambda_{40} = \mu_{40}/\mu_{20}^2$, $\mu_{40} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^4$,

$\mu_{20} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $f = \big((1/n) - (1/N)\big)$.

The biases and mean squared errors of the Isaki's (1983) estimator t_1 and Upadhyaya and Singh's (1986) estimator t_2 to the first degree of approximation are, respectively, given by

$$B(t_1) = S_y^2 f\,(\lambda_{40} - 1)(1 - C)\,, \tag{1.6}$$

$$B(t_2) = -S_y^2 g\; f\,(\lambda_{04} - 1)C\,, \tag{1.7}$$

$$MSE(t_1) = S_y^4 f[(\lambda_{40} - 1) + (\lambda_{04} - 1)(1 - 2C)]\,, \tag{1.8}$$

and

$$MSE(t_2) = S_y^4\; f\,[(\lambda_{40} - 1) + g(\lambda_{04} - 1)(g - 2C)]\,, \tag{1.9}$$

where $\lambda_{pq} = \mu_{pq}/(\mu_{20}^{p/2}\mu_{02}^{q/2})$, $\mu_{pq} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^p (x_i - \bar{X})^q$, ($p$, q) being non negative integers; and $C = (\lambda_{22} - 1)/(\lambda_{04} - 1)$.

It is easy to verify that

$$E(t_3) = S_y^2$$

which shows that the difference estimator t_3 is unbiased for S_y^2.

The variance of the difference estimator t_3 to the first degree of approximation is given by

$$MSE(t_3) = Var(t_3) = f\, S_y^4[(\lambda_{40} - 1) + d(\lambda_{04} - 1)(d - 2C)] \tag{1.10}$$

which is minimum when

$$d = C\,. \tag{1.11}$$

Thus the resulting minimum *MSE* of t_3 to the first degree of approximation is given by

$$\begin{aligned} \min.MSE(t_3) &= f\, S_y^4[(\lambda_{40} - 1) - (\lambda_{04} - 1)C^2] \\ &= f\, S_y^4(\lambda_{40} - 1)(1 - \rho^{*2})\,, \end{aligned} \tag{1.12}$$

where $$\rho^* = \frac{Cov(s_y^2, s_x^2)}{\sqrt{Var(s_y^2)Var(s_x^2)}} = \frac{(\lambda_{22} - 1)}{\sqrt{(\lambda_{40} - 1)(\lambda_{04} - 1)}}\,.$$

To the first degree of approximation, the bias and *MSE* of the Shabbir's (2006) estimator t_4 are respectively given by

$$B(t_4) = -S_y^2 g\; f\,(1 - \eta)(\lambda_{04} - 1)C\,, \tag{1.13}$$

$$MSE(t_4) = S_y^4 f[(\lambda_{40} - 1) + g(1 - \eta)(\lambda_{04} - 1)\{g(1 - \eta) - 2C\}]\,. \tag{1.14}$$

The $MSE(t_4)$ is minimum when

$$\eta_{opt} = \left(1 - \frac{C}{g}\right). \tag{1.15}$$

Thus the resulting minimum *MSE* of Shabbir (2006) estimator t_4 is given by

$$\min.MSE(t_4) = f\, S_y^4(\lambda_{40} - 1)(1 - \rho^{*2}) \tag{1.16}$$

which equals to the minimum *MSE* of the difference-type estimator t_3 $[i.e.\ \min.MSE(t_4) = \min.MSE(t_3)]$.

1.1. Sharma and Singh's (2014) estimators

Sharma and Singh (2014) have suggested three classes of estimators of the population variance S_y^2 as:

$$t_5 = w_1 s_y^2 + w_2(S_x^2 - s_x^{*2}), \tag{1.17}$$

$$t_6 = k_1 s_y^2 + k_2(S_x^2 - s_x^{*2})\left[2 - \left(\frac{s_x^{*2}}{S_x^2}\right)\right], \tag{1.18}$$

and

$$t_7 = m_1 s_y^2\left(\frac{s_x^{*2}}{S_x^2}\right) + m_2(S_x^2 - s_x^{*2}), \tag{1.19}$$

where (w_1, w_2),$(k_1, k_2)$and (m_1, m_2) are suitable chosen scalars such that mean squared errors of t_5, t_6 and t_7 are respectively minimum. We note here that the minimum mean squared errors of the estimators t_5, t_6 and t_7 obtained by Sharma and Singh (2014) are incorrect. Therefore the first objective of the authors of the present paper is to give the correct expressions of the minimum mean squared errors of the estimators t_5, t_6 and t_7 proposed by Sharma and Singh (2014).The derivation of the correct expressions of the minimum mean squared errors of the estimators t_5, t_6 and t_7 proposed by Sharma and Singh (2014) are given in the following theorems.

Theorem 1.1. (a): The bias and *MSE* of the estimator t_5 to the first degree of approximation, are respectively given by

$$B(t_5) = (w_1 - 1)S_y^2, \tag{1.20}$$

$$MSE(t_5) = S_y^4[1 + w_1^2\{1 + f(\lambda_{40} - 1)\} + w_2^2 r^2 g^2 f(\lambda_{04} - 1) + 2w_1 w_2 grf(\lambda_{22} - 1) - 2w_1], \tag{1.21}$$

where $r = S_x^2 / S_y^2$ is the ratio of two variances.

Proof: To obtain the bias and *MSE* of t_5, we write

$$s_y^2 = S_y^2(1+e_0)\ ,\ s_x^2 = S_x^2(1+e_1)$$

such that

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

and to the first degree of approximation,

$$E(e_0^2) = f(\lambda_{40}-1)\,, E(e_1^2) = f(\lambda_{04}-1) \text{ and } E(e_0e_1) = f(\lambda_{22}-1)\,.$$

Expressing t_5 in terms of e's we have

$$\begin{aligned} t_5 &= w_1S_y^2(1+e_0) + w_2\{S_x^2 - (1+g)S_x^2 + g(1+e_1)S_x^2\} \\ &= w_1S_y^2(1+e_0) + w_2gS_x^2e_1 \end{aligned}$$

or

$$(t_5 - S_y^2) = w_1S_y^2(1+e_0) + w_2gS_x^2e_1 - S_y^2$$

or

$$(t_5 - S_y^2) = S_y^2[w_1(1+e_0) + w_2gre_1 - 1]\,. \tag{1.22}$$

Taking expectation of both sides of (1.22) we get the bias of t_5 to the first degree of approximation as

$$B(t_5) = (w_1-1)S_y^2\,. \tag{1.23}$$

Squaring both sides of (1.22) we have

$$\begin{aligned} (t_5 - S_y^2)^2 &= S_y^4[1 + w_1^2(1 + 2e_0 + e_0^2) + w_2^2g^2r^2e_1^2 \\ &+ 2w_1w_2gr(e_1 + e_0e_1) - 2w_1(1+e_0) - 2w_2gre_1]\,. \end{aligned} \tag{1.24}$$

Taking expectation of both sides of (1.24) we get the *MSE* of t_5 to first degree of approximation as

$$MSE(t_5) = S_y^4[1 + w_1^2\{1 + f(\lambda_{40}-1)\} + w_2^2g^2r^2f(\lambda_{04}-1) + 2w_1w_2grf(\lambda_{22}-1) - 2w_1] \tag{1.25}$$

which proves the Theorem 1.1(a).

Theorem 1.1. (b): The optimum values of w_1 and w_2 that minimize the $MSE(t_5)$ at (1.25) are respectively given by

$$w_{10} = [1 + f(\lambda_{40}-1)(1-\rho^{*2})]^{-1}, \tag{1.26}$$

$$w_{20} = -\frac{C}{gr[1 + f(\lambda_{40}-1)(1-\rho^{*2})]}, \tag{1.27}$$

and the resulting minimum $MSE(t_5)$ is given by

$$\min.MSE(t_5) = \frac{S_y^4 f(\lambda_{40}-1)(1-\rho^{*2})}{[1+f(\lambda_{40}-1)(1-\rho^{*2})]}, \tag{1.28}$$

$$= \frac{\min.MSE(t_3)}{\left[1+\frac{\min.MSE(t_3)}{S_y^4}\right]}, \quad \text{[from (1.12)]}$$

Proof: Proof is simple so omitted.

Theorem 1.2. (a): The bias and *MSE* of the estimator t_6 to the first degree of approximation, are respectively given by

$$B(t_6) = S_y^2[k_1 + k_2 g^2 rf(\lambda_{04}-1) - 1] \tag{1.29}$$

and

$$MSE(t_6) = S_y^4[1 + k_1^2\{1+f(\lambda_{40}-1)\} + k_2^2 r^2 g^2 f(\lambda_{04}-1)$$
$$+ 2k_1k_2 grf(\lambda_{04}-1)(g+C) - 2k_1 - 2k_2 g^2 rf(\lambda_{04}-1)]. \tag{1.30}$$

Proof: Expressing the estimator t_6 in terms of e's we have

$$t_6 = k_1 S_y^2(1+e_0) + k_2 S_x^2 g e_1[2-(1+g)+g(1+e_1)]$$

$$= k_1 S_y^2(1+e_0) + k_2 S_x^2 g e_1(1+ge_1)$$

or

$$(t_6 - S_y^2) = S_y^2[k_1(1+e_0) + k_2 gr(e_1 + ge_1^2) - 1]. \tag{1.31}$$

Taking the expectation of both sides of (1.31) we get the bias of t_6 to the first degree of approximation as

$$B(t_6) = S_y^2[k_1 + k_2 g^2 rf(\lambda_{04}-1) - 1]. \tag{1.32}$$

Squaring both sides of (1.31) and neglecting terms of e's having power greater than two we have

$$(t_6 - S_y^2)^2 = S_y^4[1 + k_1^2(1+2e_0+e_0^2) + k_2^2 g^2 r^2 e_1^2$$

$$+ 2k_1k_2 gr(e_1 + e_0e_1 + ge_1^2) - 2k_1(1+e_0) - 2k_2 gr(e_1 + ge_1^2)]. \tag{1.33}$$

Taking expectation of both sides of (1.33) we get the *MSE* of t_6 to first degree of approximation as

$$MSE(t_6) = S_y^4[1+k_1^2\{1+ f(\lambda_{40}-1)\}+k_2^2 g^2 r^2 f(\lambda_{04}-1)$$
$$+2k_1k_2 grf(\lambda_{04}-1)(g+C)-2k_1-2k_2 g^2 rf(\lambda_{04}-1)] \quad (1.34)$$

which proves the Theorem1.2(a).

Theorem 1.2. (b): The optimum values of k_1 and k_2 that minimizes the $MSE(t_6)$ are respectively given by

$$k_{10} = \frac{[1-gf(g+C)(\lambda_{04}-1)]}{[1+ f\{(\lambda_{40}-1)-(\lambda_{04}-1)(g+C)^2]}, \quad (1.35)$$

$$k_{20} = \frac{[gf(\lambda_{40}-1)-C]}{gr[1+ f\{(\lambda_{40}-1)-(\lambda_{04}-1)(g+C)^2\}]} \quad (1.36)$$

and the resulting minimum $MSE(t_6)$ is given by

$$\min.MSE(t_6) = \frac{S_y^4 f(\lambda_{40}-1)[1-g^2 f(\lambda_{04}-1)-\rho^{*2}]}{[1+ f\{(\lambda_{40}-1)-(\lambda_{04}-1)(g+C)^2]}. \quad (1.37)$$

Proof: Differentiating (1.34) partially with respect to k_1, k_2 and equating to zero we get the following equations:

$$k_1\{1+ f(\lambda_{40}-1)\}+k_2 grf(\lambda_{04}-1)(g+C)=1, \quad (1.38)$$

$$k_1(\lambda_{04}-1)(g+C)+k_2 gr(\lambda_{04}-1)=g(\lambda_{04}-1). \quad (1.39)$$

Solving (1.38) and (1.39) for k_1 and k_2, we get the optimum values of k_1 and k_2, respectively as given by (1.35) and (1.36).

Substituting the values of k_{10} and k_{20}, from (1.35) and (1.36) in (1.34) we get the resulting minimum *MSE* of t_6 given by (1.37).

This proves the Theorem 1.2(b).

Theorem 1.3. (a): The bias and *MSE* of the estimator t_7 to the first degree of approximation, are respectively given by

$$B(t_7) = S_y^2[m_1\{1-gf(\lambda_{22}-1)\}-1], \quad (1.40)$$

$$MSE(t_7) = S_y^4[1+m_1^2\{1+ f[(\lambda_{40}-1)+g(\lambda_{04}-1)(g-4C)]\}]$$
$$+m_2^2 g^2 r^2 f(\lambda_{04}-1)+2m_1 m_2 grf(\lambda_{04}-1)(C-g)-2m_1\{1-gf(\lambda_{22}-1)\}]. \quad (1.41)$$

Proof: Expressing the estimator t_7 in terms of e's we have

$$t_7 = S_y^2(1+e_0)[1+g-g(1+e_1)]+m_2[S_x^2-(1+g)S_x^2+gS_x^2(1+e_1)]$$
$$= S_y^2(1+e_0)(1-ge_1)+m_2 gS_x^2 e_1$$

or

$$(t_7 - S_y^2) = S_y^2[m_1(1+e_0-ge_1-ge_0e_1)+m_2 gre_1-1]. \quad (1.42)$$

Taking expectation of both sides of (1.42) we get the bias of t_7 to the first degree of approximation as

$$B(t_7) = S_y^2[m_1\{1-gf(\lambda_{22}-1)\}-1]. \quad (1.43)$$

Squaring both sides of (1.42) and neglecting terms of e's having power greater than two we have

$$(t_7 - S_y^2)^2 = S_y^4[1+m_1^2(1+2e_0-2ge_1+e_0^2-4ge_0e_1+g^2e_1^2)+m_2^2 g^2 r^2 e_1^2$$
$$+2m_1 m_2 gr(e_1+e_0e_1-ge_1^2)-2m_1(1+e_0-ge_1-ge_0e_1)-2m_2 gre_1]. \quad (1.44)$$

Taking expectation of both sides of (1.44) we get the *MSE* of t_7 to first degree of approximation as

$$MSE(t_7) = S_y^4[1+m_1^2\{1+ f[(\lambda_{40}-1)\}+g(\lambda_{04}-1)(g-4C)]\}$$
$$+m_2^2 g^2 r^2 f(\lambda_{04}-1)+2m_1 m_2 grf(\lambda_{04}-1)(C-g)-2m_1\{1-gf(\lambda_{22}-1)\}].$$

This is same as given in (1.40). Thus the theorem is proved.

Theorem 1.3. (b): The optimum values of m_1 and m_2 that minimizes the $MSE(t_7)$ given by

$$m_{10} = \frac{[1-gf(\lambda_{04}-1)C]}{[1+f\{(\lambda_{40}-1)-(\lambda_{04}-1)C(2g+C)\}]}, \quad (1.45)$$

$$m_{20} = -\frac{[1-gf(\lambda_{04}-1)C](C-g)}{gr[1+f\{(\lambda_{40}-1)-(\lambda_{04}-1)C(2g+C)\}]} \quad (1.46)$$

and the resulting minimum $MSE(t_7)$ is given by

$$\min.MSE(t_7) = \frac{S_y^4 f(\lambda_{40}-1)[1-\rho^{*2}\{1+g^2 f(\lambda_{04}-1)\}]}{[1+f\{(\lambda_{40}-1)-(\lambda_{04}-1)C(2g+C)\}]}. \qquad (1.47)$$

Proof: Differentiating (1.40) partially with resects to m_1 , m_2 and equating to zero we get the following equations:

$$\begin{bmatrix} \{1+f[(\lambda_{40}-1)+g(\lambda_{04}-1)(g-4C)]\} & grf(\lambda_{04}-1)(C-g) \\ grf(\lambda_{04}-1)(C-g) & g^2r^2 f(\lambda_{04}-1) \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 1-gf(\lambda_{04}-1)C \\ 0 \end{bmatrix}. \qquad (1.48)$$

Solving (1.48) we get the optimum values of m_1 and m_2 as given in (1.44) and (1.45) respectively. Substituting the optimum values of m_{10} and m_{20} of m_1 and m_2 respectively in the $MSE(t_7)$ at (1.41), we get the minimum *MSE* of t_7 as given by (1.47).

Thus the theorem is proved.

1.2. Efficiency comparison

This section compares some existing known estimators of the population variance S_y^2.

From (1.5), (1.8), (1.9), (1.12) and (1.16) we have

$$Var(t_0) - \min.Min(t_j) = f\,S_y^4(\lambda_{40}-1)\rho^{*2}\ , \qquad (1.49)$$
$$j = 3,4 \qquad \geq 0,$$

$$MSE(t_1) - \min.Min(t_j) = f\,S_y^4[\rho^*\sqrt{(\lambda_{40}-1)} - \sqrt{(\lambda_{04}-1)}]^2 \qquad (1.50)$$
$$j = 3,4 \qquad \geq 0,$$

$$MSE(t_2) - \min.Min(t_j) = f\,S_y^4[\rho^*\sqrt{(\lambda_{40}-1)} - g\sqrt{(\lambda_{04}-1)}]^2 \qquad (1.51)$$
$$j = 3,4 \qquad \geq 0.$$

It follows from (1.49) to (1.51) that the difference estimator t_3 [Das and Tripathi (1978)] and Shabbir (2006) estimator t_4 (at optimum condition) are better than the usual unbiased estimator s_y^2, Isaki's (1983) estimator t_1 and Upadhyaya and Singh's (1986) estimator t_2.

Now, we present the comparison of the estimators t_5, t_6 and t_7 due to Sharma and Singh (2014) with that of Das and Tripathi's (1978) difference estimator t_3 and Shabbir (2006) estimator t_4. From (1.12), (1.16), (1.28), (1.37) and (1.47) we have

$$\min.MSE(t_j) - \min.Min(t_5) = f\,S_y^4(\lambda_{40}-1)(1-\rho^{*2})\left[1-\frac{1}{\{1+f\,(\lambda_{40}-1)(1-\rho^{*2})\}}\right] \tag{1.52}$$

$$j = 3,4 \qquad \geq 0,$$

$$\min.MSE(t_j) - \min.Min(t_6) = f\,S_y^4(\lambda_{40}-1)\left[(1-\rho^{*2})\left(1-\frac{1}{D}\right)+\frac{g^2 f\,(\lambda_{04}-1)}{D}\right] \tag{1.53}$$

$$j = 3,4 \qquad \geq 0,$$

$$\min.MSE(t_j) - \min.Min(t_7) = f\,S_y^4(\lambda_{40}-1)\left[(1-\rho^{*2})\left(1-\frac{1}{D^*}\right)+\frac{\rho^{*2} g^2 f\,(\lambda_{04}-1)}{D^*}\right] \tag{1.54}$$

$$j = 3,4 \qquad \geq 0,$$

where

$$D = [1+f\{(\lambda_{40}-1)-(\lambda_{04}-1)(g+C)^2\}] \quad \text{and}$$
$$D^* = [1+f\{(\lambda_{40}-1)-(\lambda_{04}-1)C(2g+C)\}].$$

It follows from (1.52) to (1.54) that the estimators t_5, t_6 and t_7 due to Sharma and Singh (2014) are better than Das and Tripathi's (1978) difference estimator t_3 and Shabbir (2006) estimator t_4 and hence better than the usual unbiased estimator s_y^2, Isaki's (1983) estimator t_1 and Upadhyaya and Singh's (1986) estimator t_2.

2. The suggested class of estimators for the population variance S_y^2

Keeping the form of Das and Tripathi's (1978) difference type estimator, Isaki's (1983) ratio-type estimator, Upadhyaya and Singh's (1986) estimators , Singh et al.'s (1988) estimator, Shabbir's (2006) estimator, Kadilar and Cingi's (2006, 2007) estimators, Shabbir and Gupta's (2007) estimator, Singh and

Solanki's (2013a, b) estimator, Solanki and Singh (2013) estimator, Singh et al.'s (2013, 2014) estimator, Sharma and Singh's (2014), Solanki et al. (2015) estimator and Singh and Pal (2016) estimators in view, we define a generalized class of estimators for S_y^2 as:

$$t_{SP}=\left[\phi_1 s_y^2\left\{\xi+(1-\xi)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{\alpha}\lambda\right\}+\phi_2(S_x^2-s_{x(a,b)}^2)\left\{\phi+(1-\phi)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{\delta}\right\}\right]$$
$$\times\left\{\theta+(1-\theta)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{p}\right\}\times\exp\left\{\frac{q(S_x^2-s_{x(a,b)}^2)}{(S_x^2+s_{x(a,b)}^2)}\right\}, \qquad (2.1)$$

where $s_{x(a,b)}^2=(aS_x^2+bs_x^2)/(a+b)$, (ϕ_1,ϕ_2) being suitable chosen constants, $(\phi,\xi,\theta,\lambda)$ are suitable chosen scalars such that $0\le(\phi,\xi,\lambda)\le 1$, λ may be equal to $\tau=(1+\psi C_{xy})/(1+\psi C_x^2)$ with $\psi=(1-f)/n$ and $f=n/N$; (α,δ,p,q) are scalars taking real values to generate ratio and product- type acceptable estimators; and (a,b) are either real numbers or the functions of the known parameters of the study variable *y* such as C_y coefficient of variation, (see Searls (1964), Lee (1981) and Singh (1986)), $\beta_2(y)\,(=\lambda_{40})$ (coefficient of kurtosis of the study variable *y* see Singh et al. (1973) and Searls and Intarapanich (1990)), coefficient of skewness $\beta_1(y)\,(=\lambda_{30}^2)$ of y, $\Delta(y)=(\beta_2(y)-\beta_1(y)-1)$ (see, Sen (1978), Upadhyaya and Singh (1984) and Singh and Agnihotri (2008)) or the functions of auxiliary variable *x* such as population mean $\bar{X}$, coefficient of variation C_x, coefficients of skewness $\beta_1(x)\,(=\lambda_{03}^2)$ and kurtosis $\beta_2(x)\,(=\lambda_{04})$ and the parameter $\Delta(x)=(\beta_2(x)-\beta_1(x)-1)$ or the population correlation coefficient ρ between the study variable *y* and the auxiliary variable *x*.

We would like to remark that for various values of the parameters in (2.1), we get some existing known estimators as shown in Table 2.1. Many other estimators can also be generated from the proposed family of estimators t_{SP} for suitable values of scalars (ϕ_1, ξ, α, λ, a, b, ϕ_2, ϕ, δ, θ, p, q).

Table 2.1. Some known members of proposed class of estimators

Values of the constants	**Estimator**
(λ, ϕ_1, ϕ_2, θ, ξ, ϕ, α, δ, p, q, a, b)	
(1, 1, 0, -, -, -, 0, -, 0, 0, -, -)	$t_{SP(1)}=s_y^2$
(-, 1, d, -, 1, -, -, 0, 0, 0, 0, -)	$t_{SP(2)}=s_y^2+d(S_x^2-s_x^2)$ Das and Tripathi's (1978) estimator

(1, 1, 0, -, 0, -, $-\alpha$, -, 0, 0, 0, -)	$t_{SP(3)} = s_y^2\left(\frac{S_x^2}{s_x^2}\right)^{\alpha}$ Das and Tripathi's (1978) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, 1-b, b)	$t_{SP(4)} = s_y^2\left(\frac{S_x^2}{s_x^2 + b(s_x^2 - S_x^2)}\right)$ Das and Tripathi's (1978) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, 0, -)	$t_{SP(5)} = s_y^2(S_x^2 / s_x^2)$ Isaki's (1983) ratio estimator
(1, 1, 0, -, 0, -,1, -, 0, 0, $(1+g)$, $-g$)	$t_{SP(6)} = s_y^2(s_x^{*2} / S_x^2)$ Upadhyaya and Singh's (1986) estimator
(1, ϕ_1, ϕ_2, -, -, -, 0, 0, 0, 0, 0, -)	$t_{SP(7)} = \phi_1 s_y^2 + \phi_2(S_x^2 - s_x^2)$ Singh et al.'s(1988) estimator
(1, ϕ_1, 0, -, 0, -, -1, -, 0, 0, 0, -)	$t_{SP(8)} = \phi_1 s_y^2(S_x^2 / s_x^2)$ Prasad and Singh's (1990) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{\beta_2(x)}{S_x^2}$, 1)	$t_{SP(9)} = s_y^2\left(\frac{S_x^2 + \beta_2(x)}{s_x^2 + \beta_2(x)}\right)$ Upadhyaya and Singh's (1999) estimator
(1, 1, 0, -, A, -, -1, -, 0, 0, $\frac{\beta_2(x)}{S_x^2}$, 1)	$t_{SP(10)} = As_y^2 + (1-A)s_y^2\left(\frac{S_x^2 + \beta_2(x)}{s_x^2 + \beta_2(x)}\right)$ Chandra and Singh's(2001) estimator
(1, 1, 0, -, $1+w$, -,1, -, 0, 0, $\frac{\beta_2(x)}{S_x^2}$, 1)	$t_{SP(11)} = (1+w)s_y^2 - ws_y^2\left(\frac{s_x^2 + \beta_2(x)}{S_x^2 + \beta_2(x)}\right)$ Chandra and Singh's(2001) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $-\frac{C_x}{S_x^2}$, 1)	$t_{SP(12)} = s_y^2\left(\frac{S_x^2 - C_x}{s_x^2 - C_x}\right)$ Kadilar and Cingi's (2005) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $-\frac{\beta_2(x)}{S_x^2}$,1)	$t_{SP(13)} = s_y^2\left(\frac{S_x^2 - \beta_2(x)}{s_x^2 - \beta_2(x)}\right)$ Kadilar and Cingi's (2005) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $-\frac{C_x}{S_x^2}$, $\beta_2(x)$)	$t_{SP(14)} = s_y^2\left(\frac{S_x^2\beta_2(x) - C_x}{s_x^2\beta_2(x) - C_x}\right)$ Kadilar and Cingi's (2005) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $-\frac{\beta_2(x)}{S_x^2}$, C_x)	$t_{SP(15)} = s_y^2\left(\frac{S_x^2 C_x - \beta_2(x)}{s_x^2 C_x - \beta_2(x)}\right)$ Kadilar and Cingi's (2005) estimator
(1, 1, 0, -, ξ, -1, -, 0, 0, $(1+g)$, $-g$)	$t_{SP(16)} = \xi s_y^2 + (1-\xi)s_y^2\left(\frac{s_x^{*2}}{S_x^2}\right)$ Shabbir's (2006) estimator
(τ, 1, 0, -, ξ, -, -1, -, 0, 0, 0, -)	$t_{SP(17)} = \xi s_y^2 + (1-\xi)s_y^2\left(\frac{S_x^2}{s_x^2}\right)\tau$ Kadilar and Cingi's (2006) estimators

(1, ϕ_1, ϕ_2, -, -, -, 0, 0, 0, 1, 0, -)	$t_{SP(18)} = [\phi_1 s_y^2 + \phi_2(S_x^2 - s_x^2)]\exp\left\{\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right\}$ Shabbir and Gupta's (2007) estimator
(1, ϕ_1, ϕ_2, 2, -, -, 0, 0, p, 0, 0, -)	$t_{SP(19)} = [\phi_1 s_y^2 + \phi_2(S_x^2 - s_x^2)]\left[2 - \left(\frac{s_x^2}{S_x^2}\right)^p\right]$ Gupta and Shabbir's (2008) estimator
(1, ϕ_1, ϕ_2, 2, -, -, 0, 0, 1, 0, 0, -)	$t_{SP(20)} = [\phi_1 s_y^2 + \phi_2(S_x^2 - s_x^2)]\left[2 - \left(\frac{s_x^2}{S_x^2}\right)\right]$ Gupta and Shabbir's (2008) estimator
(1, ϕ_1, ϕ_2, 2, -, -, 0, 0, -1, 0, 0, -)	$t_{SP(21)} = [\phi_1 s_y^2 + \phi_2(S_x^2 - s_x^2)]\left[2 - \left(\frac{S_x^2}{s_x^2}\right)\right]$ Gupta and Shabbir's (2008) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{\eta(1-\alpha^*)S_x^2 - \upsilon)}{S_x^2}$, $\alpha^*\eta$)	$t_{SP(22)} = s_y^2 \frac{(\eta S_x^2 - \upsilon)}{\{\alpha^*(\eta s_x^2 - \upsilon) + (1-\alpha^*)(\eta S_x^2 - \upsilon)\}}$ Yadav and Pandey's (2012) type estimator and Singh and Malik's (2014) type estimator
(1, k, 0, -, 0, -, -1, -, 0, 0, $\frac{\eta(1-\alpha^*)S_x^2 - \upsilon)}{S_x^2}$, $\alpha^*\eta$)	$t_{SP(23)} = k s_y^2 \frac{(\eta S_x^2 - \upsilon)}{\{\alpha^*(\eta s_x^2 - \upsilon) + (1-\alpha^*)(\eta S_x^2 - \upsilon)\}}$ Yadav and Pandey 's(2012) type estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_1}{S_x^2}$, 1)	$t_{SP(24)} = s_y^2\left(\frac{S_x^2 + Q_1}{s_x^2 + Q_1}\right)$ Subramani and Kumarapandiyan's (2012b) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_3}{S_x^2}$, 1)	$t_{SP(25)} = s_y^2\left(\frac{S_x^2 + Q_3}{s_x^2 + Q_3}\right)$ Subramani and Kumarapandiyan's (2012b) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_2}{S_x^2}$, 1)	$t_{SP(26)} = s_y^2\left(\frac{S_x^2 + Q_2}{s_x^2 + Q_2}\right)$ Subramani and Kumarapandiyan's (2012b) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_d}{S_x^2}$, 1)	$t_{SP(27)} = s_y^2\left(\frac{S_x^2 + Q_d}{s_x^2 + Q_d}\right)$ Subramani and Kumarapandiyan's (2012b) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_a}{S_x^2}$, 1)	$t_{SP(28)} = s_y^2\left(\frac{S_x^2 + Q_a}{s_x^2 + Q_a}\right)$ Subramani and Kumarapandiyan's (2012b) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{M_d}{S_x^2}$, 1)	$t_{SP(29)} = s_y^2\left(\frac{S_x^2 + M_d}{s_x^2 + M_d}\right)$ Subramani and Kumarapandiyan's (2012a) estimator

(1, w_1, $w_2 \frac{(a+b)}{b}$, 0, -, -, 0, 0, -1, 0, $\frac{h}{S_x^2}$, b)	$t_{SP(30)} = [w_1 s_y^2 + w_2 (S_x^2 - s_x^2)]\left(\frac{bS_x^2 + h}{bs_x^2 + h}\right)$ Singh and Solanki's (2013a) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{C_x}{S_x^2}$, 1)	$t_{SP(31)} = s_y^2\left(\frac{S_x^2 + C_x}{s_x^2 + C_x}\right)$ Singh and Solanki's (2013a) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{C_x}{S_x^2}$, $\beta_2(x)$)	$t_{SP(32)} = s_y^2\left(\frac{S_x^2\beta_2(x) + C_x}{s_x^2\beta_2(x) + C_x}\right)$ Singh and Solanki's (2013a) estimator
(1, 1, 0, -, 0 -, -1, -, 0, 0, $\frac{\beta_2(x)}{S_x^2}$, C_x)	$t_{SP(33)} = s_y^2\left(\frac{S_x^2 C_x + \beta_2(x)}{s_x^2 C_x + \beta_2(x)}\right)$ Singh and Solanki's (2013a) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{Q_3}{S_x^2}$, ρ)	$t_{SP(34)} = s_y^2\left(\frac{S_x^2\rho + Q_3}{s_x^2\rho + Q_3}\right)$ Khan and Shabbir's (2013) estimator
(1, 1, 0, -, 0, -, -1, -, 0, 0, $\frac{\gamma L_i^2}{S_x^2}$, 1) γ being a constant such that $0 \le \gamma \le 1$	$t_{SP(35)} = s_y^2\left(\frac{S_x^2 + \gamma L_i^2}{s_x^2 + \gamma L_i^2}\right)$ i= 1 to 6, $L_1 = Q_1$, $L_2 = Q_2$, $L_3 = Q_3$, $L_4 = Q_r$, $L_5 = Q_d$, $L_6 = Q_a$, where Q_1 (first quartile), Q_2 (second range), Q_3 (third quartile), $Q_d = (Q_3 - Q_1)/2$, $Q_a = (Q_3 + Q_1)/2$. Singh et al.'s (2013) estimator
(1, w_1, w_2, -, -, -, 0, 0, 0, 0, $(1+g)$, $-g$)	$t_{SP(36)} = w_1 s_y^2 + w_2 (S_x^2 - s_x^{*2})$ Sharma and Singh's (2014) estimator
(1, k_1, k_2, -, -, 2, 0, 1, 0, 0, $(1+g)$, $-g$)	$t_{SP(37)} = k_1 s_y^2 + k_2 (S_x^2 - s_x^{*2})[2 - (s_x^{*2}/S_x^2)]$ Sharma and Singh's (2014) estimator
(1, m_1, m_2, -, 0, -, 1, 0, 0, 0, $(1+g)$, $-g$)	$t_{SP(38)} = m_1 s_y^2 (s_x^{*2}/S_x^2) + m_2 (S_x^2 - s_x^{*2})$ Sharma and Singh's (2014) estimator
(1, w_1, $w_2 \frac{(a+b)}{b}$, 0, -, -, 0, 0, -1, 0, $\frac{\eta^* L^2}{S_x^2}$, δ^*)	$t_{SP(39)} = [w_1 s_y^2 + w_2 (S_x^2 - s_x^2)]\left(\frac{\delta^* S_x^2 + \eta^* L^2}{\delta^* s_x^2 + \eta^* L^2}\right)$ Singh and Pal 's (2016) estimator
(1, w_1, $w_2 \frac{(a+b)}{b}$, -, -, -, 0, 0, 0, 1, $\frac{\eta^* L^2}{S_x^2}$, δ^*)	$t_{SP(40)} = [w_1 s_y^2 + w_2 (S_x^2 - s_x^2)]\exp\left(\frac{\delta^* (S_x^2 - s_x^2)}{\delta^* (S_x^2 + s_x^2) + 2\eta^* L^2}\right)$ Singh and Pal's (2016) estimator

where $(A, w, \alpha^*, h, w_1, w_2, k_1, k_2)$ are suitable chosen constants and (δ^*, L) are either real constants or function of known parameter of an auxiliary variable x and η^* being a constant such that $|\eta^*| \le 1$.

2.1. Derivation of the expressions of bias and mean squared error (*MSE*) of the class of estimators t_{SP}

To obtain the bias and *MSE* of the class of estimators t_{SP} at (2.1) in terms of e's we have

$$t_{SP} = [\phi_1 S_y^2 (1+e_0)\{\xi + (1-\xi)(1+b^* e_1)^{\alpha}\lambda\} - \phi_2 b^* S_x^2 e_1 \{\phi + (1-\phi)(1+b^* e_1)^{\delta}\}]$$

$$\times \{\theta + (1-\theta)(1+b^* e_1)^{p}\} \times \exp\left\{-\left(\frac{b^* q}{2}\right) e_1 \left(1+\frac{b^*}{2} e_1\right)^{-1}\right\}, \qquad (2.2)$$

where $b^* = b/(a+b)$.

We assume that $|b^* e_1| < 1$ so that $(1+b^* e_1)^{\alpha}, (1+b^* e_1)^{\delta}$, $(1+b^* e_1)^{p}$ and $\left(1+\frac{b^*}{2} e_1\right)^{-1}$ are expandable. Now expanding the right hand side of (2.2), we have

$$t_{SP} = \left[\phi_1 S_y^2 (1+e_0)\left\{\xi + (1-\xi)\lambda\left[1+\alpha b^* e_1 + \frac{\alpha(\alpha-1)}{2} b^{*2} e_1^2 + \ldots\right]\right\}\right.$$

$$\left. - \phi_2 b^* S_x^2 e_1 \left\{\phi + (1-\phi)\left[1+\delta b^* e_1 + \frac{\delta(\delta-1)}{2} b^{*2} e_1^2 + \ldots\right]\right\}\right]$$

$$\times \left\{\theta + (1-\theta)\left[1 + p b^* e_1 + \frac{p(p-1)}{2} b^{*2} e_1^2 + \ldots\right]\right\}$$

$$\times \left\{1 - \frac{b^* q}{2} e_1 \left(1+\frac{b^*}{2} e_1\right)^{-1} + \frac{b^{*2} q^2}{8} e_1^2 \left(1+\frac{b^*}{2} e_1\right)^{-2} - \ldots\right\}$$

$$= S_y^2 \left[\phi_1 (1+e_0)\left\{\xi_0 + (1-\xi)\lambda\alpha b^* \left(e_1 + \frac{(\alpha-1)b^*}{2} e_1^2 + \ldots\right)\right\}\right.$$

$$\left. - \phi_2 b^* r e_1 \left\{1 + (1-\phi)\delta b^* \left(e_1 + \frac{(\delta-1)b^*}{2} e_1^2 + \ldots\right)\right\}\right] \times \left\{1 + (1-\theta) p b^* \left(e_1 + \frac{(p-1)b^*}{2} e_1^2 + \ldots\right)\right\}$$

$$\times \left\{1 - \frac{b^* q}{2} e_1 + \frac{b^{*2} q(q+2)}{8} e_1^2 - \ldots\right\}$$

$$= S_y^2 \left[\phi_1 \left\{\xi_0 (1+e_0) + (1-\xi)\lambda\alpha b^* (e_1 + e_0 e_1) + \frac{(1-\xi)\lambda\alpha(\alpha-1)b^{*2}}{2} e_1^2 + \ldots\right\}\right.$$

$$-\phi_2 b^* r\{e_1+(1-\phi)\delta b^* e_1^2+...\}]\times\left\{1+(1-\theta)pb^* e_1+(1-\theta)\frac{p(p-1)b^{*2}}{2}e_1^2\right.$$

$$\left.-\left(\frac{b^* q}{2}\right)e_1+\left(\frac{b^{*2}pq(1-\theta)}{2}\right)e_1^2+\frac{b^{*2}q(q+2)}{8}e_1^2+...\right\}$$

$$=S_y^2\left[\phi_1\left\{\xi_0(1+e_0)+(1-\xi)\lambda\alpha b^*(e_1+e_0e_1)+\frac{(1-\xi)\lambda\alpha(\alpha-1)}{2}b^{*2}e_1^2+...\right\}\right.$$

$$\left.-\phi_2 b^* r\{e_1+(1-\phi)\delta b^* e_1^2+...\}\right]\times\left\{1+u_1 b^* e_1+\frac{b^{*2}u_2}{2}e_1^2+...\right\}$$

$$=S_y^2\left[\phi_1\left\{\xi_0(1+e_0)+(1-\xi)\lambda\alpha b^*(e_1+e_0e_1)+\frac{(1-\xi)\lambda\alpha(\alpha-1)}{2}b^{*2}e_1^2+\xi_0 b^* u_1(e_1+e_0e_1)\right.\right.$$

$$\left.\left.+(1-\xi)\lambda\alpha b^{*2}u_1(e_1^2+e_0e_1^2)+\frac{b^{*2}u_2}{2}\xi_0(e_1^2+e_0e_1^2)+...\right\}-\phi_2 b^* r\{e_1+(1-\phi)\delta b^* e_1^2+b^* u_1 e_1^2+....\}\right]$$

$$=S_y^2\left[\phi_1\left\{\xi_0+\xi_0e_0+b^*\xi^* e_1+b^*\xi^* e_0e_1+\left(\frac{b^{*2}\theta^*}{2}\right)e_1^2+b^{*2}\left((1-\xi)\lambda\alpha u_1+\frac{u_2\xi_0}{2}\right)e_0e_1^2+...\right\}\right.$$

$$\left.-\phi_2 b^* r\{e_1+b^*[(1-\phi)\delta+u_1]e_1^2+....\}\right], \qquad (2.3)$$

where

$$\xi_0=[\xi+(1-\xi)\lambda],\ \ \xi^*=[\xi u_1+(1-\xi)\lambda(\alpha+u_1)],\ \ u_1=\left[(1-\theta)p-\frac{q}{2}\right],$$

$$\theta^*=[\alpha(\alpha-1)(1-\xi)\lambda+\xi_0u_2)],\ \ u_2=\left[(1-\theta)p(p-q-1)+\frac{q(q+2)}{4}\right],$$

$$\phi^*=[(1-\phi)\delta+u_1]=[(1-\phi)\delta+(1-\theta)p-(q/2)],\ r=S_x^2/S_y^2,$$

$$b^*=b/(a+b).$$

Neglecting terms of e's in (2.3) having power greater than two, we have

$$t_{SP}\cong S_y^2\left[\phi_1\left\{\xi_0+\xi_0e_0+b^*\xi^* e_1+b^*\xi^* e_0e_1+\left(\frac{b^{*2}\theta^*}{2}\right)e_1^2\right\}-\phi_2 b^* r\{e_1+b^*\phi^* e_1^2\}\right]$$

or

$$(t_{SP}-S_y^2)\equiv S_y^2\left[\phi_1\left\{\xi_0+\xi_0e_0+b^*\xi^* e_1+b^*\xi^* e_0e_1+\left(\frac{b^{*2}\theta^*}{2}\right)e_1^2\right\}-\phi_2 b^* r\{e_1+b^*\phi^* e_1^2\}-1\right], \qquad (2.4)$$

Taking expectation of both sides of (2.4) we get the bias of t_{SP} to the first degree of approximation as

$$B(t_{SP}) = S_y^2\left[\phi_1\left\{\xi_0 + f(\lambda_{04}-1)b^*\left(\frac{b^*\theta^*}{2}+\xi^* C\right)\right\} - \phi_2 r b^{*2}\phi^* f(\lambda_{04}-1) - 1\right]. \tag{2.5}$$

Squaring both sides of (2.4) and neglecting terms of e's having power greater than two, we have

$$\begin{aligned}(t_{SP}-S_y^2)^2 &= S_y^4[1+\phi_1^2\{\xi_0^2(1+2e_0+e_0^2)+2b^*\xi_0\xi^*(e_1+2e_0e_1)+b^{*2}(\xi^{*2}+\theta^*\xi_0)e_1^2\}\\ &+\phi_2^2 r^2 b^{*2}e_1^2 - 2\phi_1\phi_2 b^* r\{\xi_0(e_1+e_0e_1)+b^*(\xi^*+\phi^*\xi_0)e_1^2\} - 2\phi_1\{\xi_0(1+e_0)\\ &+b^*\xi^*(e_1+e_0e_1)+((b^{*2}\theta^*)/2)e_1^2\}+2\phi_2 b^* r(e_1+b^*\phi^* e_1^2)].\end{aligned} \tag{2.6}$$

Taking expectation of both sides of (2.6) we get the *MSE* of t_{SP} to the first degree of approximation as

$$MSE(t_{SP}) = S_y^4[1+\phi_1^2 A_1+\phi_2^2 A_2 - 2\phi_1\phi_2 A_3 - 2\phi_1 A_4 + 2\phi_2 A_5], \tag{2.7}$$

where

$$\begin{aligned}A_1 &= [\xi_0^2 + f\{\xi_0^2(\lambda_{40}-1)+b^*(\lambda_{04}-1)[b^*(\xi^{*2}+\theta^*\xi_0)+4\xi_0\xi^* C]\}],\\ A_2 &= r^2 b^{*2} f(\lambda_{04}-1),\\ A_3 &= b^* r f(\lambda_{04}-1)[b^*(\xi^*+\phi^*\xi_0)+\xi_0 C],\\ A_4 &= [\xi_0 + b^* f(\lambda_{04}-1)((b^*\theta^*/2)+\xi^* C)],\\ A_5 &= b^{*2}\phi^* r f(\lambda_{04}-1),\\ f &= ((1/n)-(1/N)).\end{aligned}$$

Differentiating (2.7) partially with respect to ϕ_1 and ϕ_2 and equating to zero, we have

$$\begin{bmatrix} A_1 & -A_3 \\ -A_3 & A_2 \end{bmatrix}\begin{bmatrix}\phi_1 \\ \phi_2\end{bmatrix} = \begin{bmatrix} A_4 \\ -A_5\end{bmatrix}. \tag{2.8}$$

Solving (2.8) we get the optimum values of ϕ_1 and ϕ_2 as

$$\left.\begin{aligned}\phi_1 &= \frac{(A_2A_4 - A_3A_5)}{(A_1A_2 - A_3^2)} = \phi_{10}(say)\\ \phi_2 &= \frac{(A_3A_4 - A_1A_5)}{(A_1A_2 - A_3^2)} = \phi_{20}(say)\end{aligned}\right\}. \tag{2.9}$$

Putting (2.9) in (2.8) we get the resulting minimum MSE of t_{SP} as

$$\min .MSE(t_{SP}) = S_y^4\left[1-\frac{(A_2A_4^2 - 2A_3A_4A_5 + A_1A_5^2)}{(A_1A_2 - A_3^2)}\right]. \tag{2.10}$$

Thus we established the following theorem.

Theorem 2.1: Up to the first degree of approximation,

$$MSE(t_{SP}) \geq S_y^4\left[1-\frac{(A_2A_4^2 - 2A_3A_4A_5 + A_1A_5^2)}{(A_1A_2 - A_3^2)}\right]$$

with equality holding if

$$\phi_1 = \phi_{10},$$
$$\phi_2 = \phi_{20},$$

where ϕ_{i0} 's (i=1, 2)are defined in (2.9).

2.2. Special Case-I $(\lambda = 1)$

Putting $\lambda = 1$ in (2.1) we get the class of estimators of S_y^2 as

$$t_{SP}^{(1)} = \left[\phi_1 s_y^2\left\{\xi + (1-\xi)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{\alpha}\right\} + \phi_2(S_x^2 - s_{x(a,b)}^2)\left\{\phi + (1-\phi)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{\delta}\right\}\right]$$
$$\times\left\{\theta + (1-\theta)\left(\frac{s_{x(a,b)}^2}{S_x^2}\right)^{p}\right\}\times\exp\left\{\frac{q(S_x^2 - s_{x(a,b)}^2)}{(S_x^2 + s_{x(a,b)}^2)}\right\}. \tag{2.11}$$

Substitution of $\lambda = 1$ in (2.5) and (2.7) yields the bias and *MSE* of the class of estimators to the first degree of approximation, respectively as

$$B(t_{SP}^{(1)}) = -S_y^2\left[1-\phi_1\left\{1+ f(\lambda_{04}-1)b^*\left(\frac{b^*\theta_1^*}{2} + \xi_1^* C\right)\right\} + \phi_2 r b^{*2}\phi^* f(\lambda_{04}-1)\right] \tag{2.12}$$

and

$$MSE(t_{SP}^{(1)}) = S_y^4[1 + \phi_1^2 A_1^* + \phi_2^2 A_2 - 2\phi_1\phi_2 A_3^* - 2\phi_1 A_4^* + 2\phi_2 A_5], \tag{2.13}$$

where

$$A_1^* = [1+ f\{(\lambda_{40}-1) + b^*(\lambda_{04}-1)[b^*(\xi_1^{*2} + \theta_1^*) + 4\xi_1^* C]\}],$$

$$A_3^* = b^* rf(\lambda_{04}-1)[b^*(\xi_1^* + \phi^*) + C],$$
$$A_4^* = [1 + b^* f(\lambda_{04}-1)((b^*\theta_1^*/2) + \xi_1^* C)],$$
$$\xi_0 = 1, \xi_1^* = [u_1 + \alpha(1-\xi)], \theta_1^* = [\alpha(\alpha-1)(1-\xi) + u_2].$$

The $MSE(t_{SP}^{(1)})$ is minimized for

$$\left.\begin{aligned} \phi_1 &= \frac{(A_2 A_4^* - A_3^* A_5)}{(A_1^* A_2 - A_3^{*2})} = \phi_{10}^{(1)} \\ \phi_2 &= \frac{(A_3^* A_4^* - A_1^* A_5)}{(A_1^* A_2 - A_3^{*2})} = \phi_{20}^{(1)} \end{aligned}\right\}. \tag{2.14}$$

Thus the resulting minimum $MSE(t_{SP}^{(1)})$ is given by

$$\min.MSE(t_{SP}^{(1)}) = S_y^4\left[1 - \frac{(A_2 A_4^{*2} - 2A_3^* A_4^* A_5 + A_1^* A_5^2)}{(A_1^* A_2 - A_3^{*2})}\right]. \tag{2.15}$$

Thus we established the following corollary.

Corollary 2.1: Up to the first degree of approximation,

$$MSE(t_{SP}^{(1)}) \geq S_y^4\left[1 - \frac{(A_2 A_4^{*2} - 2A_3^* A_4^* A_5 + A_1^* A_5^2)}{(A_1^* A_2 - A_3^{*2})}\right]$$

with equation holding if

$$\phi_1 = \phi_{10}^{(1)},$$
$$\phi_2 = \phi_{20}^{(1)}.$$

From (2.10) and (2.15) we have

$$\min.MSE(t_{SP}^{(1)}) - \min.MSE(t_{SP})$$
$$= S_y^4\left[\frac{(A_2 A_4^2 - 2A_3 A_4 A_5 + A_1 A_5^2)}{(A_1 A_2 - A_3^2)} - \frac{(A_2 A_4^{*2} - 2A_3^* A_4^* A_5 + A_1^* A_5^2)}{(A_1^* A_2 - A_3^{*2})}\right]$$

which is positive if

$$\frac{(A_2 A_4^2 - 2A_3 A_4 A_5 + A_1 A_5^2)}{(A_1 A_2 - A_3^2)} > \frac{(A_2 A_4^{*2} - 2A_3^* A_4^* A_5 + A_1^* A_5^2)}{(A_1^* A_2 - A_3^{*2})}. \tag{2.16}$$

Thus the proposed family of estimators t_{SP} would be more efficient than the family of estimators $t_{SP}^{(1)}$ as long as inequality (2.16) is satisfied.

2.3. Special Case-II $(\phi_1, \lambda) = (1,1)$

Putting $(\phi_1, \lambda) = (1,1)$ in (2.1) we get class of estimators of S_y^2 as

$$t_{SP}^{(2)} = \left[s_y^2 \left\{ \xi + (1-\xi)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{\alpha} \right\} + \phi_2 (S_x^2 - s_{x(a,b)}^2) \left\{ \phi + (1-\phi)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{\delta} \right\} \right]$$
$$\times \left\{ \theta + (1-\theta)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{p} \right\} \times \exp\left\{ \frac{q(S_x^2 - s_{x(a,b)}^2)}{(S_x^2 + s_{x(a,b)}^2)} \right\}. \qquad (2.17)$$

Inserting $(\phi_1, \lambda) = (1,1)$ in (2.5) and (2.7) yield the bias and *MSE* of $t_{SP}^{(2)}$ to the first degree of approximation, respectively given by

$$B(t_{SP}^{(2)}) = S_y^2 b^* f(\lambda_{04} - 1)\left[\left(\frac{b^* \theta_1^*}{2} \right) + \xi_1^* C - \phi_2 r b^* \phi^* \right] \qquad (2.18)$$

and

$$MSE(t_{SP}^{(2)}) = S_y^4 [1 + A_1^* - 2A_4^* + \phi_2^2 A_2 - 2\phi_2 (A_3^* - A_5)]. \qquad (2.19)$$

The $MSE(t_{SP}^{(2)})$ is minimized for

$$\phi_2 = \frac{(A_3^* - A_5)}{A_2} = \phi_{20}^* (say). \qquad (2.20)$$

Thus the resulting minimum $MSE(t_{SP}^{(2)})$ is given by

$$\min.MSE(t_{SP}^{(2)}) = S_y^4 \left[1 - A_1^* - 2A_4^* - \frac{(A_3^* - A_5)^2}{A_2} \right], \qquad (2.21)$$
$$= f S_y^4 (\lambda_{40} - 1)(1 - \rho^{*2})$$

which equals to the minimum *MSE* of the difference estimator t_3 defined in (1.3).

Now, we state the following corollary.

Corollary 2.2: Up to the first degree of approximation,

$$MSE(t_{SP}^{(2)}) \geq f S_y^4 (\lambda_{40} - 1)(1 - \rho^{*2})$$

with equality holding if

$$\phi_2 = \phi_{20}^*.$$

From (2.15) and (2.21) we have

$$\min.MSE(t_{SP}^{(2)}) - \min.MSE(t_{SP}^{(1)}) = \frac{S_y^4[A_3^*(A_3^* - A_5) - A_2(A_1^* - A_4^*)]^2}{A_2(A_1^* A_2 - A_3^{*2})} \tag{2.22}$$

which is always positive. It follows that the proposed family of estimators $t_{SP}^{(1)}$ is better than the family of estimators $t_{SP}^{(2)}$ and the difference type estimators t_3 in (1.3) at their optimum conditions.

2.4. Special Case-III $(\phi_1 = 1)$

For $\phi_1 = 1$, the suggested class of estimators t_{SP} reduces to the class of estimators

$$t_{SP}^{(3)} = \left[s_y^2 \left\{ \xi + (1-\xi)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{\alpha} \lambda \right\} + \phi_2 (S_x^2 - s_{x(a,b)}^2) \left\{ \phi + (1-\phi)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{\delta} \right\} \right]$$
$$\times \left\{ \theta + (1-\theta)\left(\frac{s_{x(a,b)}^2}{S_x^2} \right)^{p} \right\} \times \exp\left\{ \frac{q(S_x^2 - s_{x(a,b)}^2)}{(S_x^2 + s_{x(a,b)}^2)} \right\}. \tag{2.23}$$

Putting $\phi_1 = 1$ in (2.5) and (2.7) we get the bias and *MSE* of the estimator $t_{SP}^{(3)}$ to the first degree of approximation, respectively given by

$$B(t_{SP}^{(3)}) = -S_y^2 \left[1 - \left\{ \xi_0 + fb^*(\lambda_{04} - 1)\left(\frac{b^*\theta^*}{2} + \xi^* C \right) \right\} + \phi_2 r b^{*2} \phi^* f(\lambda_{04} - 1) \right] \tag{2.24}$$

and

$$MSE(t_{SP}^{(3)}) = S_y^4[1 + A_1 - 2A_4 + \phi_2^2 A_2 - 2\phi_2(A_3 - A_5)]. \tag{2.25}$$

The $MSE(t_{SP}^{(3)})$ is minimized when

$$\phi_2 = \frac{(A_3 - A_5)}{A_2} = \phi_{20}^{*(1)}(say). \tag{2.26}$$

Thus the resulting minimum $MSE(t_{SP}^{(3)})$ is given by

$$\min.MSE(t_{SP}^{(3)}) = S_y^4 \left[1 + A_1 - 2A_2 - \frac{(A_3 - A_5)^2}{A_2} \right]. \tag{2.27}$$

Now, we state the following corollary.

Corollary 2.3. To the first degree of approximation,

$$MSE(t_{SP}^{(3)}) \geq S_y^4\left[1+A_1-2A_2-\frac{(A_3-A_5)^2}{A_2}\right]$$

with equality holding if

$$\phi_2=\phi_{20}^{*(1)}.$$

From (2.10) and (2.27) we have

$$\min.MSE(t_{SP}^{(3)})-\min.MSE(t_{SP})=\frac{S_y^4[A_2(A_1-A_4)-A_3(A_3-A_5)]^2}{A_2(A_1A_2-A_3^2)}$$

$$\geq 0 \qquad (2.28)$$

which clearly indicates that the t_{SP} family of estimators is more efficient than that of the $t_{SP}^{(3)}$ family of estimators .

Concluding remarks

This paper intends to suggest a new family of estimators for the variance S_y^2 of the variable *y* of interest when the population variance S_x^2 of the auxiliary variable *x* is known. The proposed family generalizes that of the several estimators ($t_{SP(i)}$, i= 1 to 38) as listed in Table 2.1. We have obtained the bias and mean squared error (*MSE*) expressions up to first order of approximation in simple random sampling without replacement (*SRSWOR*). From the bias and *MSE* expressions of the suggested family, one can easily derive the bias and *MSE* expressions of existing known estimators as well as those of potential new proposals. The present study unifies several results at one place.

The family is certainly not exhaustive but it can act as different against the proliferation of equivalent proposals that could be appearing in the future. Three subclasses of the proposed family are identified and their properties are studied. We have also given the comparisons among the proposed class of estimators and the three subclasses of estimators. It has been theoretically shown that the proposed class of estimators is more efficient than the difference type estimator t_3 due to Das and Tripathi (1978) and hence the usual unbiased estimator s_y^2 and Isaki (1983) ratio estimator t_1 and several other estimators. This paper also provides the correct *MSE* expressions of the estimators (t_5, t_6, t_7) recently proposed by Sharma and Singh(2014). Indeed, improvement upon the difference type estimators t_3 as well as upon other estimators can be achieved when the theoretical expressions of the minimum mean squared error are considered. These expressions are based on the knowledge of population parameters which can be

obtained either through past data or experience gathered in due course of time. For more discussion on this issue, the reader is referred to Das and Tripathi (1978) and Srivastava and Jhajj (1980). However, more light on this study can be focused if one would have included an empirical study. Overall this study is of academic interest as well as of practical importance, see, Diana et al. (2011), Singh et al. (2013) and Singh and Solanki (2013 a, b), Solanki and Singh(2013), Singh et al. (2013), Singh et al.(2014), Solanki et al. (2015) and Singh and Pal (2016) etc.

Acknowledgements

Authors are thankful both the referees for his valuable suggestions regarding improvement of the paper.

REFERENCES

CHANDRA, P., SINGH, H. P., (2001). Modified estimators for population variance that utilizes the kurtosis of an auxiliary variable in sample surveys. Vikram Math. Jour., 21, 31–36.

DAS, A. K., TRIPATHI, T. P., (1978). Use of auxiliary information in estimating the finite population variance. Sankhya, C, 40 (2), 139–148.

DIANA, G., GIARDAN, M., PERRI, P. F., (2011). An improved class of estimators for the population mean . Statist. Math. Appl., 20(2), 123–140.

GUPTA, S., SHABBIR, J., (2008). Variance estimation in simple random sampling using auxiliary information. Hacett. Jour. Math. Statist., 37 (1), 57–67.

HILAL, A. L., TAILOR, R., SINGH, H. P., VERMA, M. R., (2014). New alternatives to ratio estimators of population variance in sample surveys, Appl. Math. Comp., 247, 255–265.

ISAKI, C. T., (1983). Variance estimation using auxiliary information. Jour. Amer. Statist. Assoc., 78 (381), 117–123.

KADILAR, C., CINGI, H., (2005). A new ratio estimator in stratified random sampling. Commun. Statist. Theo. Meth., 34, 597–602.

KADILAR, C., CINGI, H., (2006). Ratio estimators for the population variance in simple and stratified random sampling. Appl. Math. Comp., 173 (2), 1047–1059.

KADILAR, C., CINGI, H., (2007). Improvement in variance estimation in simple random sampling. Commun. Statist. Theo. Meth., 36, 2075–2081.

KHAN, M., SHABBIR, J., (2013). A ratio type estimator for the estimation of population variance using quartiles of an auxiliary variable. Jour. Statist. Applica. Prob., 2 (3), 319–325.

LEE, K. H., (1981). Estimation of variance of mean using known coefficients of variation. Commun. Statist. Theo. Meth., A10 (5), 503–514.

PRASAD, B., SINGH, H. P., (1990). Some improved ratio-type estimators of finite population variance in sample surveys. Commun. Statist. Theo. Meth., 19 (3), 1127–1139.

PRASAD, B., SINGH, H. P., (1992). Unbiased estimators of finite population variance using auxiliary information in sample surveys. Commun. Statist. Theo. Meth., 21 (5), 1367–1376.

SEARLS, D. T., (1964). The utilization of a known coefficients variance in the estimation procedure. Jour. Amer. Statist. Assoc., 59 (308), 1225–1226.

SEARLS, D. T., INTARAPANICH, P., (1990). A note on an estimator for the variance that utilizes kurtosis. Amer. Statist., 44 (4), 295–296.

SEN, A. R., (1978). Estimation of the population mean when the coefficient of variation is known. Commun. Statist. Theo. Meth., A7 (7), 657–672.

SHABBIR, J., (2006). A new estimator of population mean in stratified sampling. Commun. Statist. Theo. Meth., 35 (7), 1201–1209.

SHABBIR, J., GUPTA, S., (2007). On improvement in variance estimation using auxiliary information. Commun. Statist. Theo. Meth., 36 (12), 2177–2185.

SHARMA, P., SINGH, R., (2014). Improved dual to variance ratio type estimators for population variance. Chilean Jour. Statist., 5 (2), 45–54.

SINGH, H. P., PAL, S. K., SOLANKI, R. S., (2013). Improved estimation of finite population variance using quartiles. Istatistik- Jour. Tur. Stat. Assoc., 6 (3), 166–121.

SINGH, H. P., PAL, S. K., SOLANKI, R. S., (2014). A new procedure for estimation of finite population variance using auxiliary information. Jour. Reliab. Stat. Stud., 7 (2), 149–160.

SINGH, H. P., AGNIHOTRI, N., (2008). A general procedure of estimating population mean using auxiliary information in sample surveys. Statist. Trans. new series, 9 (1), 71–78.

SINGH, H. P., PAL, S. K., (2016). An efficient class of estimators of finite population variance using quartiles. Jour. Appl. Stat., 43 (10), 1945–1958.

SINGH, H. P., SOLANKI, R. S., (2013a). A new procedure for variance estimation in simple random sampling using auxiliary information. Statistical Papers, 54 (2), 479–497.

SINGH, H. P., SOLANKI, R. S., (2013b). Improved estimation of finite population variance using auxiliary information. Commun. Statist. Theo. Meth., 2 (15), 2718–2730.

SINGH, H. P., (1986). A note on the estimation of variance of sample mean using the knowledge of coefficients of variation in natural population. Commun. Statist. Theo. Meth., 15 (12), 3737–3746.

SINGH, H. P., UPADHYAYA, L. N., NAMJOSHI, U. D., (1988). Estimation of finite population variance. Cur. Sci., 57 (24), 1331–1334.

SINGH, J., PANDEY, B. N., HIRANO, K., (1973). On the utilization of a known coefficient of kurtosis in estimation procedure of variance. Amn. Inst. Statist. Math., 25 (1), 51–55.

SINGH, R., MALIK S., (2014). Improved estimation of population variance using information on auxiliary attribute in simple random sampling. Appl. Math. Comp., 235, 43–49.

SOLANKI, R. S., SINGH, H. P., (2013). An improved class of estimators for the population variance. Mod. Assist. Statist. Appl., 8 (3), 229–238.

SOLANKI, R. S., SINGH, H. P., PAL, S. K., (2015). Improved ratio-type estimators of finite population variance using quartiles, Hacettepe Jour. Math. Stat., 44 (3), 747–754.

SRIVASTAVA, S. K., JHAJJ, H. S., (1980). A class of estimators using auxiliary information for estimating finite population variance. Sankhya, C, 42 (1-2), 87–96.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012a). Variance estimation using median of the auxiliary variable. Inter. Jour. Prob. Statist., 1 (3), 62–66.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012b). Variance estimation using quartiles and their functions of an auxiliary variable. Inter. Jour. Statist. Appl., 2 (5), 67–72.

UPADHYAYA, L. N., SINGH, H. P., (1984): On the estimation of the population mean with known coefficient of variation. Biometrical Jour., 26 (6), 915–922.

UPADHYAYA, L. N., SINGH, H. P., (1986). On a dual to ratio estimator for estimating finite population variance. Nepal Math. Sci. Rep., 11 (1), 37–42.

UPADHYAYA, L. N., SINGH, H. P., (1999). Use of transformed auxiliary variable in estimating the finite population mean. Biometrical Jour., 41 (5), 627–636.

YADAV, S. K., PANDEY, H., (2012). Improved family of estimators for the population variance using qualitative auxiliary information. Assam Statist. Rev., 26 (2), 63–70.

YADAV, S. K., KADILAR, C., SHABBIR, J., GUPTA, S., (2015). Improved family of estimators of population variance in simple random sampling. Jour. Statist. Theo. Pract., 9, 219–226.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 631–658

SHIFT IN METHODOLOGY AND POPULATION CENSUS QUALITY

Elżbieta Gołata[1]

ABSTRACT

The article refers to the shift in methods to conduct a population census: from a conventional enumeration through a sample survey and a mixed approach to administrative data, as a new standard in statistics. The paper compares two Polish censuses of 2002 and 2011. It is aimed at quality assessment in the case of both: the traditional method (2002 census) and the combined approach (2011 census).

The quality of census data is discussed with essential aims and objectives to provide reliable information on the population age and sex structure in detailed territorial division. Therefore, quality assessment is provided for the whole country and at regional level. First of all, coverage errors are considered. We use multiple sources of data and non-matching methods, in particular: demographic analysis based on previous censuses, vital statistics and a comparison with other existing sources. Different cross-sections according to sex, age and place of residence are considered. In each separate domain adequacy and divergence assessments are accompanied by an attempt to provide substantive explanations.

Key words: population census quality, register-based census, coverage errors.

1. Introduction

A population census is not only the oldest investigation, best-known, well-formed in terms of methodology, but also an investigation which is widely regarded as the most reliable source of data. As the methods of conducting censuses, especially methods of data collecting, have changed incredibly over the last decades, it is important to address the issue of quality assessment of the population census under the shift in methodology. The purpose of this paper is to discuss the quality of information derived from the 2011 population census in Poland. Special attention is given to a comparison of census data accuracy in view of the conventional versus register-based approach.

[1] Poznań University of Economics. E-mail: elzbieta.golata@ue.poznan.pl.

Modern technologies, their development and application in all spheres of social and economic life influenced also the population census methodology. Huge changes and modifications can be seen at every stage of the census procedure, and, in principle, they are observed in all countries around the world (UN 2012). Some countries have opted for a fundamental change in methodology understood as a new source of data, while the others introduced only some innovations in the technology of data collecting and processing. One of the main reasons for these changes is the need of saving, but also improving census data quality (Longva, Thomsen, Severeide 1998, UN 2010a, 2012, 2013, CSO 2012). As noted by P. Valente (2010), it is essential to count the population, but census taking is costly, and a growing number of people are reluctant to participate. Similar opinions are also presented in the CSO report (2012), which emphasized that organization of the census turned out to be very expensive and laborious. For that reason, Poland decided to give up the traditional census in favour of a mixed method, in the 2011 round. National Statistical Institutes (NSIs) in many countries face growing challenges and difficulties: pressure to make greater use of information available elsewhere, lower public cooperation and participation, changing user demand and the need to control or reduce costs (UN 2011). These issues were discussed during various meetings and seminars and were ideally recognized by T. Holt in his presentation: *The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper* (2007)[2].

According to the UN Recommendations for the 2010 censuses of population and housing (UN 2006), there are three primary approaches[3] to conducting a census, based on the method of data collection. These are: a) the conventional method of universal enumeration based on field operations at a given moment, b) the method of using registers and other administrative sources, and c) a combination of registers and other administrative sources and surveys. In the 2010 round, majority (56%) of the countries applied the traditional method, but the percentage using the register-based approach doubled amounting to 14.5%. However, if the complex method were also taken into account, this proportion would rise from 20% in the 2000 round up to 40% in the 2010 round (UN 2012).

Register-based censuses have already been conducted since the 70s (Statistics Denmark 1995, Statistics Finland 2004, Statistics Netherland 2004). Invaluable in this respect is the experience of the Nordic countries (Statistics Finland 2004, UN 2007, UN 2011). But the 2010 round brought a methodological shift in the way of conducting censuses in many countries (UN 2010a, 2012). The register-based approach and the mixed method greatly expanded. They were applied also by countries of a much more numerous population than Finland, Sweden, Denmark or Norway, and in countries with little experience in the use of administrative data in official statistics, like Poland and Germany. Poland is one of those countries

[2] This is the title of professor D. Tim Holt's invited presentation during a special seminar in celebration of the 100th meeting of Committee on National Statistics of The National Academies.

[3] In addition, two methods might be reported: the so-called "rolling census" carried in France and the traditional enumeration with yearly updates in the United States.

which decided to abandon the traditional method and to turn to the "mixed" one. Public administration registers and information systems were used as the census-data source, but data on different topics was also collected directly from population in a large-scale sample survey.

This study attempts to provide quality assessment of data from the 2011 population census in Poland. There are a number of methods to evaluate censuses including: post enumeration surveys[4], demographic analysis, interpenetrating studies used in conjunction with a current census, record checks and a comparison of census data with results of existing household surveys (Fosu 2001, UN 2009, 2010). Evaluation methods differ with respect to the type of error to be measured (coverage and content error), technical sophistication, data requirements and quality of results (Baldrige, Brown, Jones, Keane 1985). Baldridge et al. (1985) presented a typology of such methods and distinguished methods based on a single source of data and methods based on comparison of data from two or more sources (matching and non-matching studies).

According to UN survey (UN 2013b), demographic analysis was the method used by the greatest number of countries (76 per cent) for the measurement of either undercoverage or overcoverage. Additionally, differences in the methods used by countries conducting different types of census were observed. For a traditional census, a larger proportion of countries used a census coverage survey and demographic analysis. More countries, which adopted the register-based approach, used comparisons with aggregate administrative datasets and comparisons with existing surveys. Also, the majority of countries that implemented the combined method, conducted the comparison with unit level administrative datasets, an analysis of questionnaire return rates and demographic analysis. In all countries, regardless of differences in adopted census methodology, comparisons with existing surveys were the highest or second highest reported method.

The census coverage survey is usually considered as one of the best methods to assess the accuracy of census estimates, particularly in terms of coverage (Baldrige et al., 1985, Kordos, 2007, 2012, Gołata, 2012). Although the coverage survey was conducted in Poland shortly[5] after 2011 census, its results have not been published yet and there is no information on their use in the estimation. Additionally, results of the coverage analysis have not been presented by CSO, nor detailed description of the methodology applied. And CCS data are unavailable for researchers for the purposes of scientific research. However, at the end of 2015 preparations were undertaken by CSO to make individual unidentifiable data from 2011 census available for scientists.

[4] Both surveys: Post Enumeration Survey (PES) and Census Coverage Survey (CCS) are non-demographic methods of post census adjustments for population estimates. Therefore, if there was no explicit reference to a particular study, these terms might be used interchangeably.

[5] The census coverage survey was conducted from 1 to 11 July 2011, while the census data was collected from 1 April until 30 June 2011, as of March 31, hours. 24:00.

In this study we use multiple sources of data and non-matching methods, in particular: demographic analysis based on the previous census, vital statistics, birth and death registers, population register (PESEL) and the comparison with other existing sources. The choice of these methods is a result of the availability of data. It is natural to question quality of data from all sources to assess their reliability. But the process of the comparison analysis included in the combined method of taking censuses and creating the 'gold record' makes it possible to work with quality assessment in a new way (Wallgren & Wallgren, 2013). Polish Population census in 2011 was a combination of registers, other administrative sources and surveys that were subjected to a thorough quality assessment. This allowed for a thesis that the mixed census, as a combination of multiple data sources and as a part of a register-based statistical system, provides estimates that are not of a worse quality than those from a traditional census.

The assessment of the census quality under the new approach involves many methodological and practical issues. Some of them are discussed in the paper. The first group of problems relates to change in the methods in statistical research adopted for conducting the census. In the next section questions on assessing the quality of the census are discussed under essential aims and objectives, to provide reliable information on the population by age and sex in a detailed territorial division. Important issues involving methodological questions, data sources and types of errors are indicated. However, the study is limited to assessing the quality of the census due to coverage errors. The results of an empirical study are presented in the fourth section. Considerations are closed with conclusions and some final remarks.

2. Shift in statistics and population census quality

Currently, we are witnessing a change in the way of conducting statistical surveys (Baffour, King, Valente, 2013, Zhang 2012, UN 2011). Q. A and B. Bakker (2000) and P. van der Laan (2000) define changes in Statistics Netherlands as a process of reorganization of social statistics. To take one example, they present Sociaal Statistisch Bestand (SSB), a micro-data base obtained as a result of record linkage and statistical integration of different administrative records. Among most important reasons for these changes one may indicate: an increase in demand for information, the pressure to improve the efficiency of statistical process to make savings in costs and staff resources, demands to reduce the burden placed on the respondents to statistical surveys, but also the development of computing, data collection methods, data editing and integration and in estimation methodology (Wallgren & Wallgren, 2013, UN 2011). These expectations are often contradictory and force statisticians to consider alternatives to the traditional survey approach. The most natural is to see if usable data already exist elsewhere and may be used for statistical purposes. These data are rarely direct substitutes for those collected via statistical surveys,

but there are many possibilities like a combination of variables from multiple sources to obtain satisfactory results (UN 2011).

In particular, the above changes apply to a population census which is the oldest survey for counting people and recording their characteristics. Censuses are normally carried out once every decade for the whole population residing in the country. For centuries, the census was the most common form of examination the entire population (Bethlehem 2009). Census data constitute one of the most important source of information relating to demographic and socio-economic characteristics, because it provides a broad overview of a country's population to the lowest level of geographical division.

Although it dates back to works presented by John Graunt (1620-1674) and William Petty (1620 – 1683), the idea of studying some representatives instead of the entire population became more popular at the beginning of the XX century. The representative method has been developed by the works of Jerzy Neyman, Karl Pearson and Sir Ronald Fisher. Currently, sample surveys are best known and most commonly used method of conducting statistical surveys. However, due to the growing financial restrictions and an increasing number of data, nowadays a survey would not be carried out automatically. First, one would rather look at registers, administrative records and other existing sources (including the Internet and Big data (Ruggles, 2014)) to learn what information is available. Different data sources, like parish records, or other administrative records were also used in past. The research conducted by J. Graunt is widely known. In *Natural and Political Observations Made upon the Bills of Mortality* (1662), he used the mortality rolls in London to construct first life tables. But this was rare. At present, administrative records are easily available, although they are created for different purposes, there are many possibilities of exploring them for statistics. A new approach to obtain information for statistics is observed (Zhang, 2013, Al, Bakker 2000).

This change applies to the nature of statistics, understood as the whole process of obtaining information which is the basis for further research and analysis. And it is not just the data collecting, but the whole process of statistical survey (Wallgren and Wallgren, 2007, 2014). The increasing use of administrative data for statistical purposes is called by Wallgren and Wallgren (2013) the transition to the register-based statistics production system. This means a shift from a system based on address lists and interviews to the one in which sample surveys become 'register-based'. Registers are not only helpful in updating frames, improving sample scheme and survey design, but are also used in estimation process, provide auxiliary data for estimation or serve for evaluation purposes (UN 2011, Baffour, King, Valente 2013, Zhang 2011, 2012). The administration register data may be combined with other data sources as well as it can be used to improve other surveys in the system.

The shift in the process of statistical research, and particularly in the way of conducting population census, enlarged also the palette of evaluation topics, as many different criteria for a successful census might be listed (UN 2013a). Of

course, at first one would mention the two types of census errors: coverage and content errors (Baldrige et al., 1985). But the 2010 round of population censuses showed that the change in data sources induced a wider use of modern technologies and new methods. Instead of classical enumeration, data were extracted from administrative records, conventional field operations were replaced with Internet transfers. The use of multiple sources of data for the census induced developments in the use of data imputation, record linkage, calibration, estimation using auxiliary variables from external sources. Implementation of each of these projects can be considered as one of the evaluation criteria: improving quality of the registers, accuracy of the estimates, cost reduction, use of modern ICT in data collecting and dissemination. In view of the new methodology, including new data sources, quality assessment might be considered in a structural way. Berka et al. (2012) proposed a three stage approach to derive quality of raw, combined and imputed data in three hyperdimensions (Documentation HD_D, Pre-processing HD_P and External Source HD_E) to satisfy such requirements as transparency, accuracy and feasibility (UN 2012b). Discussing the change in census methodology, some authors indicate the need for conceptualization and measurement of the statistical accuracy in register statistics, which would enable application of rigorous statistical concepts such as bias, variance, efficiency and consistency, as in the case of survey sampling (Zhang 2011).

Wallgren & Wallgren (2013) discuss quality assessment for register-based statistical systems as a process consisting of two parts, each of which has two levels. The first one is to analyse the source itself. It includes a discussion of metadata regarding the analysed source to determine its relevance, and an analysis of microdata from the source to determine its accuracy. The second part is a comparison analysis of the source with its base register and with other sources in the system containing similar variables. Systematic comparisons between surveys and registers in the system give new knowledge of quality in different surveys, and also give new possibilities to redesign surveys to improve their quality.

By 2011, population censuses in Poland were carried out using traditional methods involving census enumerators visiting all inhabited units and noting down information obtained from respondents on census forms (available in hard copy). The 2011 Polish Census of Population and Housing (NSP 2011) was the first census conducted since Poland's accession to the European Union, and it took place in the period from 1 April to 30 June 2011 (as of 31 March 2011, at 00.00). The census was conducted by applying the mixed method with the use of administrative records (full survey - short form), supplemented by information from Internet self-enumeration. Additionally, a sample survey (long form) was carried out on approximately 20% of randomly selected dwellings. Data collected from administrative registers and sample survey formed the so-called golden record. This record was the result of integration of information from all data sources in the environment of Operational Micro-Database. Further processing

allowed for creation of Analytical Micro-Database which was used as the basis for census estimates.

As described in Berka et al. (2012) and Wallgren & Wallgren (2013), during the construction of the golden record, detailed studies and comparative analysis were carried out. In preparation for the 2011 census metadata about 300 various administrative registers were collected and analysed. All variables in those systems were rated with regards to the possibility of obtaining information on population, housing and buildings, in line with the recommendations and classifications of the United Nations Economic Commission for Europe (UNECE) and Eurostat (UN 2006). In preparation for the 2011 census Central Statistical Office (CSO) examined many administrative records and conducted a large-scale research of their conformity for the census as concerns concepts, definitions and classifications (*The report on the work of sub-group for the use* ..., 2007, *Memo from the current state of research* ..., 2007 Dziubiński, 2008 Kobus, Smolka, Nowakowska, 2009, *List of concepts and definitions* ..., 2007, Golata 2009). As a result of a detailed analysis, 28 registers were selected. Among them, as a priority, the following systems should be mentioned: Common Electronic System of Population Register (PESEL), Social Security System (ZUS), the Health Insurance System, Land and Buildings, Register of Territorial Division of the Country, data from the State Fund for Rehabilitation of Persons with Disabilities. Information collected from administration sources, which was properly structured and divided into strata, was also used in creating the frame for the census sample survey.

Social assessment of the new census methodology and attention of the scientific community are diverse: some opinions give full recognition and others are negative. The traditional census was perceived by the public, local government, and also by many scientists, as an indisputable source of 'certain' and unquestionable information (Barwiński, 2014, *Raport* … 2011). However, there are also clear assessments indicating that the previous arrangements were not ideal because of coverage errors (Sakson 2002, Śleszyński 2004, 2005) and due to the fact that the data was not collected directly from the respondents (Paradysz 2002). J. Paradysz (2010) underlines the need for critical evaluation of previous censuses and suggests the usage of all available data sources to improve census estimates. On the other hand, lack of comprehensive information on such topics as families and households, as well as unavailability of data for a detailed territorial division (information that was available only from the sample survey) is often considered as a disadvantage of the mixed census (Gołata 2013).

Coverage errors refer to either an undercount or overcount of units owing to omissions, duplication or erroneous inclusion. In the traditional census, an undercount was a typical situation (Paradysz 2002, 2010). Operational guidelines for conducting Post Enumeration Surveys (UN 2010) illustrate the use of various procedures with results obtained for selected countries, in majority, undercounts. As for the register-based census, Lenk (UN 2012b) underlines the importance of detecting inactive records in the population register and to eliminate them to avoid

overcoverage. Statistics Austria, for example, used the residence analysis, which allowed ensuring that only individuals with a pre-defined number of "signs of life" were counted in the census. All the individuals covered only by the population register, but not by other administrative source, were asked in a written form to confirm their main place of residence. Finally, approximately 0.5 percent of the initial population was not counted[6]. Unfortunately, CSO did not provide any information on applying similar procedure to avoid overcoverage in 2011 census in Poland.

There are a number of non-demographic methods of post census adjustments for population estimates: Post Enumeration Survey (PES), Coverage Surveys and the Reverse Record Check, Dual System Estimation (DSE) or Residents Temporarily Overseas (RTOs), used in many countries, e.g. in Australia, Canada, Japan, New Zealand, United Kingdom or USA (Newell and Smallwood 2010). Dual system estimation is one of the methods that can be used to generate population estimates from census data (Plewis et al. 2011, Brown et al. 2006). This method is based on the assumption of independence between census and census coverage survey data (different data collection methods, different personnel and a different address frame). The method applies matching procedures and detail analysis ensuring that individuals counted by both surveys are correctly allocated by age and sex within each of CCS estimation areas. Then, a two-dimensional distribution table is constructed to allow comparisons of the estimates. Having data from the population register and CCS, it seems possible to apply a similar approach (Tab. 1) to assess coverage.

Table 1. Basic Table for Dual System Estimation

Source of data	Observation	Census Coverage Survey		
		Observed	Not observed	Total
PESEL - Population Register	Observed	n_{11}	n_{12}	$n_{1.}$
	Not observed	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Source: Based on Plewis et al. (2011), Brown et al. (2006)

It is important to put attention to n_{12}, the number of persons observed in the population register, but not in the CCS. Differences in this dimension seem to be possible and significant, as Poland is a country of intensive migration, and people who migrate (irrespective of the length of their stay abroad) are included in the register unless they notify the authorities about leaving the country. Another real problem concerns the people who went abroad and died there. If nobody informs the Polish office of a death, the person may "live in the register" even up to 200 years. As a result, the oldest man of the world lives in Poland, according to the

[6] In Austria, due to the results of the test census, about 80 percent of the non-counted individuals were removed from the residence registers by local municipalities (UN 2012b).

register (Kuc 2014). Another important inconsistency is the number n_{21} , representing people observed in the CCS but not in the PESEL. This number may refer to all foreigners staying in Poland (even for more than 12 months) who do not have legal resident status. This status is associated with the registration for permanent residence, which requires submission of a document confirming the right of permanent residence[7]. In the case of a failure to meet formal requirements for permanent residence, even immigrants living in Poland for over 12 months were not counted as residents in the census.

Under independence assumption ($\theta = \frac{n_{11}n_{22}}{n_{21}n_{12}} = 1$) within a given age-sex group in the analysed territorial unit, we can estimate the census coverage and CCS coverage as well as the unobserved number of persons n_{22}. However, if the independence assumption is not valid then DSE will be biased: when $\theta > 1$, DSE has negative bias and when $\theta < 1$, DSE has positive bias (Brown et al. 2006). The assumptions of homogeneity and independence are very strong, and for several reasons they might not be met (Plewis, Simpson and Williamson 2011). Differences in probabilities of responding to CCS and of 'being included' in the register are possible, particularly by age and sex group, and for different territorial units. Some people may have no chance of being included in the register (for example due to legal regulations or because their propensity to respond to government enquires equals to zero). The matching process might be invalid for different reasons, e.g. migration, due to postponing the registration of newborn children, or inertness of the register in updating the reported changes.

However, not ignoring the importance of CCS in assessing the quality of the census, as there is no access to the data, this study provides census evaluation only in terms of demographic analysis in comparison to previous censuses, administrative data and other existing surveys, mirror statistics. Demographic analysis was carried out for the population of the whole country by sex and age, and with regard to certain aspects of territorial division.

3. Evaluation of 2011 population census in Poland

There is a considerable difficulty in identifying references for assessing accuracy of the estimates of the 2011 population census in Poland. The population register may serve as one of them. Another may be the census sample survey, which was conducted on a random sample of 20% of dwellings on the national scale. These two studies were the primary sources of census data, so they could hardly be considered as reference in assessing the census accuracy.

[7] This documentation differs depending on the immigrant's home country and may involve complex procedures. For a person from a country outside the European Union such documentation includes a permit for a long-term residency in the European Union, the decision to grant the refugee status in the Polish Republic, the award of subsidiary protection or tolerated stay permit in Poland, among other things.

The population register was evaluated during preparations for the census (Józefowski and Rynarzewska-Pietrzak,2010, Paradysz, 2010, Roszka, 2013). Recognizing generally very positive results obtained, the population register was accepted as the basis for the census data system. The census survey was one stage sampling scheme with deep stratification and consisted of more than 2,744 thousand dwellings, out of nearly 13.5 million. Although for all census results precision tables were provided, the original weights had to be adjusted due to 13.7% of non-response. Nevertheless, the analysis of non-response has not been available yet. In turn, Census Coverage Survey did not meet the requirements of an independent survey carried out in a more precise way. It was conducted by CSO using the same frame. A sample of 80 thousand dwellings was drawn out of 2,744 thousand flats drawn earlier to the census sample survey. But the frame was restricted only to flats with at least one person with an assigned phone or mobile phone number, and the survey was performed by CATI. Additionally, it covered all dwellings that took part in self-enumeration by the Internet.

All the above reasons influenced the decision to discuss the census quality in terms of demographic analysis in comparison to several existing data sources including the previous census. Previous traditional censuses in Poland were evaluated mainly by demographers, who used the possibilities of demographic analysis based on other existing data sources. There is quite well documented evidence on coverage errors in Polish censuses (Jończy, 2010, Kordos, 2007, Paradysz, 2010, Sakson, 2002, Śleszyński, 2004, 2005, Zasępa 1993). As concerns the coverage assessment, it is common to have a net census undercount as the number of omissions usually exceeds the number of duplications. Among the biggest coverage errors, J. Paradysz (2002, 2010) indicated a shortage of up to 30% of women with the shortest duration of marriage (1988 Census), omission of 10% of the youngest infants up to 6 months (2002 Census), omission of the population with increased mobility (2002 Census), lack of the elderly aged over 90 (2002 Census).

In the coverage assessment, special attention was paid to population at the age of an increased risk of biased estimates. These age groups were defined on the basis of earlier studies (Paradysz, 2010) and an introductory analysis. Special consideration was also paid to the fact that Poland is a country of intensive emigration, and consequently to the population at the age of particularly intensive migration mobility (as in the classical Rogers and Castro model).

We started the evaluation with a simple survival analysis. This phase consisted of a comparison of the census population in 2002 and 2011 by sex and age, including relevant aging. Survival rates are a basic tool in this case. The compatibility of survival rates was examined at first for: a) census data from 2002 and 2011 (Fig. 1 - thicker lines denoted as SR Census: solid for the entire population, dashed for men and dotted for women), and then also for b) projection of 2002 census data for 2011 (Fig. 1 - thinner lines denoted as SR Projection). Since both censuses, in 2002 and 2011, were carried out in spring (with the critical moments of May 21, 2002 and March 31, 2011), a simplifying assumption that the single age groups overlap was adopted. This means that an additional ageing for 1 month was omitted, and we assumed that a person aged 1 completed

year (according to the 2002 census), at the critical moment of the 2011 census, was aged 10 completed years. The survival coefficients were obtained according to formula (1).

$$SR_{\frac{2002}{2011}}(s,x) = \frac{\mathrm{P}_{2002}(s,x)}{\mathrm{P}_{2011}(s,x+9)} \tag{1}$$

where:

$SR_{\frac{2002}{2011}}(s,x)$ – between 2002 and 2011 censuses survival rate by sex and age

P_t – census population: t=2002 or t=2011

s – sex: *T*- both sexes; *F*- females; *M*- males

x – age

Both estimates of survival rates: (a) based on census data and (b) on demographic projection (of 2002 census data for 2011) are very much in line. This similarity of estimates obtained by using different data and different methodology indicates compatibility of the data. However, some values of the survival rates are cause for concern. The obtained results indicate the existence of such single year age groups, for which survival rates between censuses (SR Census) take values greater than unity. A closer look at the values of the SR Census coefficients allowed us to note that higher values were assigned for women than for men, although similar tendencies were observed. For both sexes the same age groups focus special attention. These are: 9-13 years, 16-19 years, 30 years and 33-35 with survival rates (SR Census) exceeding one. In addition, for the age of 65 years we observe a temporary collapse of the survival rate. It drops for about 5.5% from 0.89 to 0.84, but in the age of 66 years, almost 4% increase in the value (up to 0.88) was observed.

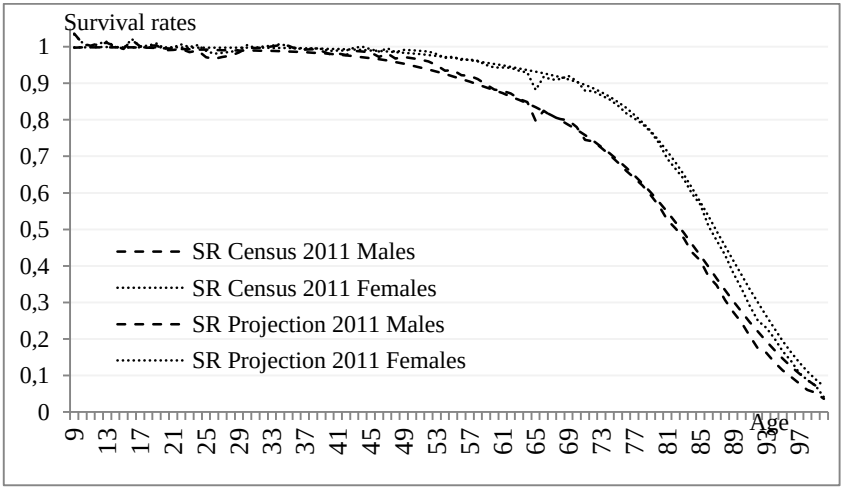


Figure 1. Survival rates for single year intervals between 2002 and 2011 census, Poland

Source: Estimates based on 2002 and 2011 Polish Population Census data and life tables, http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

As between the censuses the survival rates (SR Census) relate to the population of people living at the time of the previous census, they should not exceed unity. Situations when this condition is not met suggest poor quality of the data or intense impact of migration. First, we analysed the impact of migration. The inclusion of migration in the analysis required relevant data that were available only for 5-year age intervals. Estimation for five-year intervals required an appropriate amendment in the formula. Summing the population for 5-year age intervals allowed the smoothing of the survival rates function according to age. The inclusion of migration allowed the elimination of unexpected values of survival rates (Fig. 2 and Tab. 2). The explanation of questionable survival rates was possible as the migration balance indicated a higher number of emigrants than immigrants. Nevertheless, Poland, which is historically a country of emigration, in recent years has served also as the host country for immigrants. However, both numbers of emigrants and immigrants differ significantly. The number of emigrants (according to 2011 census) amounted to over 2 million and was more than 50 times higher than the number of immigrants, which in 2011 census was estimated at approx. 40 thousand.

Statistics relating to migration raises legitimate uncertainty. However, they do not undermine the observed relationship (Kicinger and Koryś 2011, Fihel, Kaczmarczyk, Okólski 2006). Furthermore, recent migration has shifted to be more fluctuating with changing destination due to studies, employment or family reasons, indefinite period of stay and changes of the country of residence. This applies to the Polish migration after accession to the UE in 2004 in particular. Researchers are exploring these new phenomena under the concept of transnationalism (Borket and Penninx 2011).

Table 2. Survival rates for five year intervals between 2002 and 2011 census with and without migration, Poland

Age	Survival Rates					
	Without migration			With migration		
	Total	Males	Females	Total	Males	Females
10-14	**1.006**	**1.006**	**1.006**	0.974	0.974	0.974
15-19	**1.005**	**1.004**	**1.006**	0.979	0.979	0.979
20-24	0.995	0.990	**1.001**	**0.935**	**0.938**	**0.931**
25-29	0.980	0.974	0.986	**0.867**	**0.873**	**0.861**
30-34	**1.001**	0.999	**1.002**	**0.892**	**0.893**	**0.892**
35-39	0.995	0.994	0.996	**0.917**	**0.916**	**0.918**
40-44	0.988	0.983	0.993	0.926	0.921	0.931
45-49	0.983	0.975	0.990	0.930	0.921	0.939
50-54	0.969	0.955	0.983	0.929	0.913	0.944
55-59	0.941	0.918	0.963	0.913	0.889	0.936

Table 2. Survival rates for five year intervals between 2002 and 2011 census with and without migration, Poland (cont.)

Age	Survival Rates					
	Without migration			With migration		
	Total	Males	Females	Total	Males	Females
60-64	0.905	0.869	0.939	0.888	0.852	0.922
65-69	0.860	0.806	0.908	0.849	0.796	0.896
70-74	0.815	0.739	0.877	0.807	0.731	0.868
75-79	0.734	0.640	0.803	0.728	0.635	0.796
80+	0.456	0.392	0.490	0.451	0.389	0.485

Source: Estimates based on 2002 and 2011 Polish Population Census data, http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

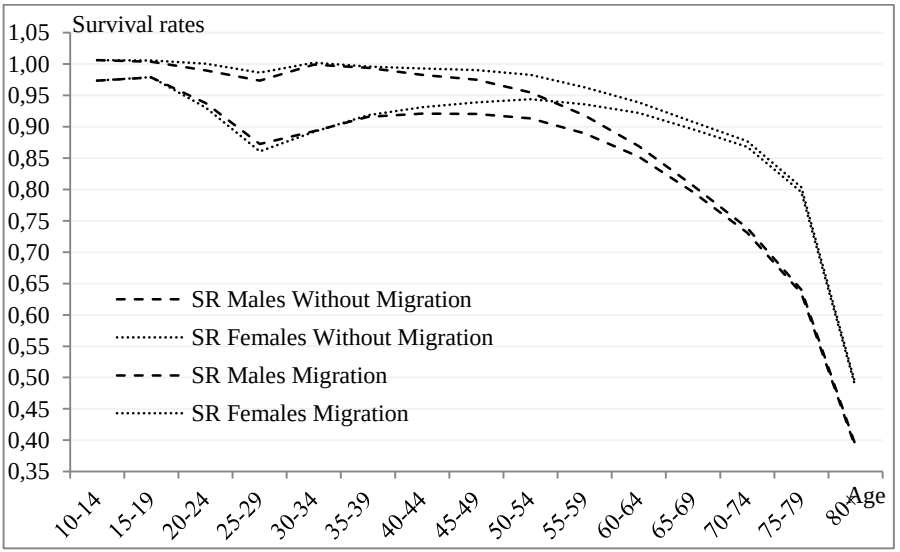


Figure 2. Survival rates for five year intervals between 2002 and 2011 census with and without migration, Poland

Source: Estimates based on 2002 and 2011 Polish Population Census data, http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

Given that Poland is a source country for migrants, a significant decrease in the population was expected, and in fact introducing migration leads to a reduction of survival rates. It is especially pronounced in the age group with the highest mobility: 20-34 years old. For these age groups survival rates decreased by up to 11% (Tab. 2). We observe also a change in the relationship between survival ratios for men and women, reflecting differences in the intensity of migration by sex (Fig. 2). Survival rates without migration are higher for women than for men. This relationship is consistent with the survival probabilities of life

tables. The introduction of migration eliminates differences between the probabilities of the life tables and between censuses survival ratios. In addition, the reduction was so deep that we observe lower survival rates for women than men, as a consequence of migration. For women in the age group of 25-29 years, a drop in the survival ratios amounted nearly to 13%, while for men it was slightly above 10%. For subsequent age groups these differences are becoming less significant. Over 60 years of age, they do not exceed 2% and are rapidly converging to zero.

Returning to the assessment of census data quality, conducted with the use of demographic analysis, we compare 2011 census estimates (denoted as Census 2011 on Fig. 3) with estimates obtained by predicting population from 2002 census (denoted as Projection 2011 in Fig. 3). The population by sex and age of the 2002 census was adopted as a starting point for the projection. Survival probabilities from the life tables for the years 2002-2011 were used to obtain age and sex structure of the population for the following years, similarly as in population predictions. Live births by sex in years 2002-2011 from vital statistics evidence were incorporated and subjected to the ageing procedure. A two-stage approach was applied. The first stage used data from the records of infant's deaths by month. These data allowed estimating the number of children completing the first year of life. Next, the above estimates were subjected to the second stage of the ageing procedure that used survival probabilities of the life tables for subsequent years of age.

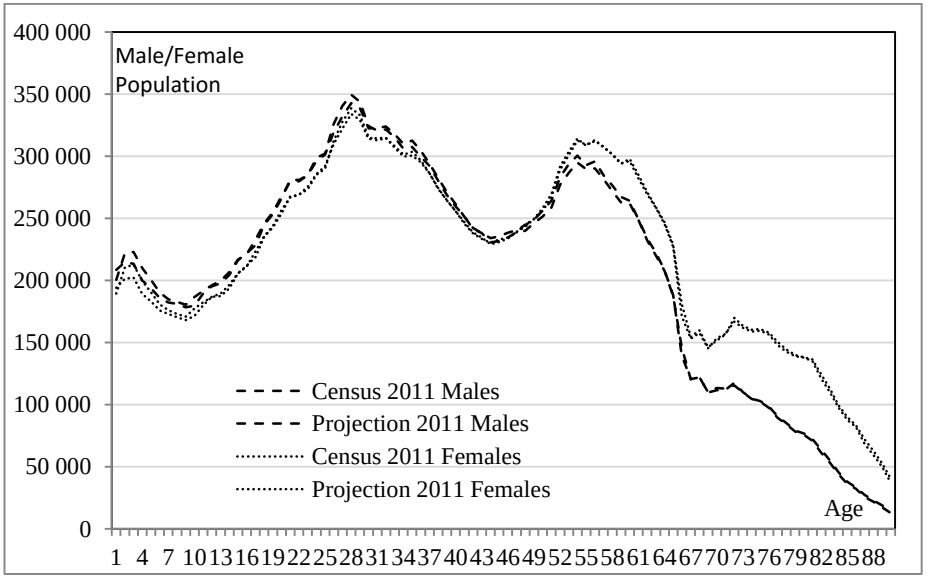


Figure 3. Population by age and sex: 2011 census estimates and projection, Poland

Source: 2002 and 2011 Population Census data, life tables and vital statistics years 2002-2011, http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

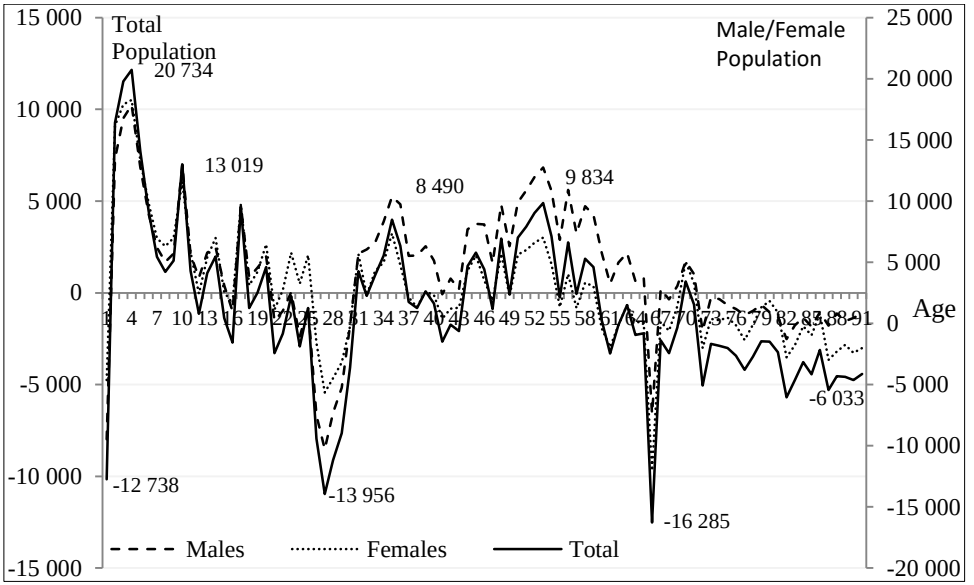


Figure 4. Differences between 2011 census population and projection, Poland

Source: 2002 and 2011 Population Census data, life tables and vital statistics years 2002-2011, http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

Initially, this procedure did not consider migration as data on migration are not available for one age intervals. Additionally, as indicated above, the measurement of migration gives rise to considerable controversy (Fihel, Kaczmarczyk, Okólski 2006). This issue, along with the appropriate analysis requires a separate study. On the other hand, omission of migration allows evaluating the quality of census estimates. Generally, one can expect projection compliance with the 2011census data, except for those years of age in which intense migration was observed. The above presented algorithm for population estimates was applied on the national scale as well as for selected regions.

Comparison analysis of the two population estimates (Census 2011 and Projection 2011, Fig. 3) allows observing good compatibility at national scale. Difference between estimates of the 2011 census and the reference population is small and amounts about 100 thousands in total, which constitutes 0.26%. However, direction of this relation differs by sex. For women underestimation is observed while overestimation for men. This relation varies by age (Fig. 4) allowing identification of those age groups which require special attention: (i) infants and children 0-4 years old, (ii) young people: studying and starting their professional career, (iii) working age population, (iv) the elderly.

While analysing population prediction based on 2002 census compared with 2011 census data, one must constantly bear in mind the methodological differences resulting from the method of the census: traditional and register-

based. However, regardless of the applied methodologies, we assess the compatibility of the final estimates. Note that if the difference between 2011 census and 2011 projection is negative, it means that the administrative registers do not cover a specific group of people - underestimation is observed. The opposite situation, when the 2011 census data are greater than expected from the projection, suggests that census data in fact show a non-existent population.

In each of the described cases, the justification of observed discrepancies would be desirable. In the previous discussion attention has been paid to the impact of migration, which is especially obvious for the age of 20-35 years. In the following discussion we turn attention to the possible explanations of differences observed for infants and children up to 4 years.

3.1. Coverage assessment – infants and children

A comparison of census population aged 0 completed years at the census critical moment (1 April 2011) with data on live births from vital statistics evidence allowed the evaluation of similarities and differences of these two data sources. With the register-based census approach, it is natural to expect full agreement with vital statistics. For the aim of this analysis, detailed data on births by month and sex (for five years before census) was incorporated and subjected to the ageing procedure. As described above, the two-stage approach was applied. Depending on availability of data on infants deaths by age in months and in time (by month of the year), vital statistics data was used to estimate the number of children completing the first year of life. In the analysis for subsequent years of age, more aggregate data was used based on survival probabilities of life tables for single year of age. To provide a comparable assessment for 2002 census, an identical procedure was applied to information from the years 1997-2002.

Data obtained for infants showed that 2011 census population was underestimated by nearly 13 thousand compared with the birth and death registers (Tab. 2). This represents 3.3% of census population. For the total number of children aged up to 4 years, an opposite situation was observed. This group of census population was overestimated by more than 58 thousand compared with the birth register. This result is difficult to explain, since census data showed children not included in the birth register. A common mistake is rather to underestimate the population, whereas the overall overestimation by 2.8% was observed in this case.

An underestimation of the number of infants is often explained by a delayed birth registration. An assessment of the population register, which was carried by Józefowski and Rynarzewska-Pietrzak (2010), indicated at least two-week delay in the transmission of information about new-born children. In addition, the authors raised the problem that the register does not take into account all the events as of the indicated date. This ‘outdate’ of the register means also the omission of infants. And it is particularly important that administrative records include all the events as of the indicated date – the critical moment of the census.

Table 2. Coverage assessment for infants and children: differences between 2011 census data and estimates based on vital statistics

Age	Total		Males		Females		Population aged 0-4 years and the difference between estimates		
	persons	%	persons	%	persons	%	2011 Census	Birth Register	Difference
0	-12 738	-3.3	-7 981	-4.0	-4 757	-2.5	2 057 998	1 999 725	58 273
1	16 414	3.8	7 297	3.3	9 117	4.3			2.83%
2	19 776	4.5	9 511	4.3	10 265	4.8			
3	20 734	5.1	10 229	4.9	10 505	5.3			
4	14 086	3.6	6 857	3.4	7 230	3.8			
5	9 192	2.5	4 304	2.3	4 888	2.7			

Source: Estimates based on 2002 and 2011 Polish Population Census data, life tables and vital statistics (records of births and deaths in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

Table 3. Coverage assessment: differences between the number of children under one year of life according to the birth register and 2002 and 2011 population censuses

Data source	2011 Census			2002 Census		
	Total	Males	Females	Total	Males	Females
Census	389 903	200 592	189 311	351 662	180 116	171 546
Birth Register	402 641	208 573	194 068	357 096	183 440	173 656
Difference	-12 738	-7 981	-4 757	-5 434	-3 324	-2 110
Difference (%)	3.27	3.98	2.51	1.55	1.85	1.23

Source: Estimates based on 2002 and 2011 Polish Population Census data, life tables and vital statistics (records of births in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

Similar differences were also observed earlier, in censuses conducted with the conventional approach (Paradysz, 2010). However, it might be expected that while conducting a census based on a population register, greater compatibility will be observed than in the case of a traditional census with independent field operations. Detailed information on the census of 2002 and 2011 indicates quite the opposite situation (Tab. 3). A discrepancy between the estimates of census and the birth records was identical in direction, but in 2011 it amounted to 3.27% and was higher by more than 100% in comparison to 2002, when it was equal to 1.55. These results imply a need for further work on the quality of the population registry in cooperation between responsible government authorities and public statistics and in consultation with the scientific community.

As already discussed, for children aged 1-4 years an opposite relation was observed. This time it was an overestimation. When we focus on children aged 1 completed year or older, the differences between census estimates, and the ones resulting from the birth register were positive. The biggest difference related to children aged 3 years (completed in 2011) exceeded 20 thousand (more than 5%). The overestimation was slightly higher for girls than for boys. For the total number of population aged 1-4, census population exceeded register data by 71 thousand, that is 4.3%.

The observed discrepancies might be associated with intensive migration and an increasing number of births given by Polish women abroad, especially in the United Kingdom (Janta 2013, Waller *et al.*, 2014, Zumpe *et al.*, 2012). Where are the infants registered as born in Poland, but not enumerated during the census? Where are the children aged 1-4 enumerated by the census, who were not listed in the Polish Birth Register? Answers to these questions are beyond the scope of this paper. Probably, an indefinite life situation might suggest Polish migrants to enumerate their children in the census survey in Poland. Some of them not only decided to enumerate their children in Poland, but also registered them in the home country. This might be logically explained, as children born abroad to Polish parents need to be also registered in Poland in order to gain Polish passports. And a large number of children born in England and Wales to parents of Polish citizenship obtain Polish passport (ONS 2013 p.23). The ONS data for the youngest age group (0-4 years), show difference amounting almost to 50 thousand between Polish-born and Polish nationals. The ONS data show also 74 456 live births in UK to Polish women in 2007-2010 (Zumpe *et al.* 2012 p. 24). An in-depth mirror statistics might reveal some trends, but it is basically impossible to provide exact numbers.

3.2. Coverage assessment – young people

The analysis referring to young people receiving education and starting their professional careers identifies another important problem. As showed above, the census underestimates population of young people aged 25-30. However, the analysis carried out at regional level gives various examples, either confirming or not confirming this observation.

It is worth noting that one of the primary purposes of the census is to provide information on population by age and sex in detailed territorial division. In the census based on administrative records, this task is fulfilled with respect to information that comes from registers. In relation to those characteristics of the population that can be possessed only from a sample survey, the problem of estimating for small areas arises. This means verification of compliance of definitions and classifications, data integration, examining the relation between different characteristics to choose auxiliary variables, methodological studies on estimation for small domains, assessment of consistency, calibration, etc. Bearing in mind all the above problems and their impact and consequences for the assessment of the census quality, we confine ourselves with the coverage analysis.

In regional dimension, the coverage is obviously different from that for the whole state. Poland is a country characterized by large regional differences. Therefore, a comprehensive evaluation of census data by territory will certainly provide a wide variety of information that would be extremely valuable in a regional development strategy. For this reason, the evaluation of regional census data would require a separate study. In this paper we focused on assessing the compatibility of the estimates for young people aged 20-35 years. The overall assessment for the whole country showed a significant underestimation of this group of people. Data at the country level is a balance of regional assessments. There are such territorial units for which the indicated underestimation would be even greater. But there are also such regions where we obtain contrary information, with the case of large cities as an example. Exemplary considerations apply to the population of Poznan - the fifth largest city in Poland with half million inhabitants (554 696). At regional level, one may notice greater discrepancies between 2011 census data and the projection than on the national scale. The total number of residents of the city was underestimated by more than 20 thousand, which gives 3.7% of the 2011 census population.

The differences between 2011 census and the projection are widely disparate according to age (Tab. 4). The biggest underestimate of 11.7 ths. (23.1%) refers to Poznan residents aged 30-34 years. On the other hand, the greatest overestimation of 10.5 ths. (24.1%) was observed for the age of 20-24. The difference between the number of people in a given age group according to 2011 census and corresponding population of respectively younger age group according to 2002 census shows unusual trends. The observed changes do not result from natural demographic processes, births and deaths, but internal migration and the suburbanization process.

Table 4. Differences between 2011 census data and projection based on 2002 census, children and youth, Poznan

Age	2011 census - adequate 2002 census population*		2011 census population – projection based on 2002 census	
	Absolute	Relative (%)	Absolute	Relative (%)
10-14	-2 206	-9.7	-2 360	-11.4
15-19	1 198	4.8	221	0.9
20-24	11 887	37.6	10 471	24.1
25-29	8 982	20.3	4 318	8.1
30-34	-12 595	-19.9	-11 684	-23.1
35-39	-10 957	-21.1	-7 058	-17.2

Note: *Difference between the number of people in a given age group according to 2011 census and corresponding population of respectively younger age group according to 2002 census

Source: Estimates based on 2002 and 2011 Polish Population Census data, life tables and vital statistics (records of births and deaths in years 2002-2011), http://demografia.stat.gov.pl/bazademografia (Accessed 10 May 2014).

Similar discrepancies for the age group 20-35 were observed by T. Józefowski and B. Rynarzewska -Pietrzak (2011), who studied the quality of the population register. They indicated that the reason for these discrepancies (amounting even to 34%) is the relationship between the actual population (census) and permanent residents (register), which results from the fact that Poznan functions as a university centre. As the capital of the region, Poznan is a city of almost 160 thousand students of different forms of studies, including nearly 90 thousand of regular daily students (in the 2011/2012 academic year, there were 88 349 regular daily students in Poznan, CSO 2013). These young people usually are not registered as Poznan citizens. But after graduation they usually decide to stay in the city and take a job there. In the 2011 census a full analogy was observed. Estimates based on the projection for the age group of 20-27 years were much lower than census data, as they refer to indigenous inhabitants of the city, who nine years earlier were 11-18 years old, and they do not include students who came to Poznan from other places.

Current findings seem to be consistent with the analysis conducted for previous census data. But it is worth noting that a distinct change in the relationship between census data and projection estimates was observed for the age of 29 years. It should be emphasized at this point that the projection was made on the basis of the census carried out by traditional method in 2002. This means that the increase in data was due to the number of students studying in Poznan nine years earlier. In 2011 census, the estimates referred to data from the registry, which did not include students. Thus, the drop observed for the age of 29 indicates the 'loss' of the population that had studied at the universities nine years earlier. The deficiencies in the register-based 2011 census suggest that students studying in the city nine years earlier had not decided to stay after graduation. Of course, we do not know whether they returned to their place of residence prior to the studies, or emigrated abroad. Nor do we know whether they were people who were successful in their professional career in Poznan or surroundings, and decided to live in suburban areas rather than in the city centre. The deficiencies of 2-3 thousand people in each of the subsequent year of age show that the city was not able to keep the potential of young and educated people. All in all, this is a group of about 20 thousand people, that is 4% of the city population at the age of the most intense economic, matrimonial and reproductive activities.

This analysis may also indicate another problem. It is the decreasing number of city residents, which is not only related to foreign emigration, but also to the process of suburbanization. Within a radius of 20-25 km around the city, new settlements are created and inhabited mostly by young, educated people, who after graduation and marriage, change a student’s flat in the city for a house near the city (Klimanek, 2012). At the moment, the problem of suburbanization is becoming increasingly important for the development of the city and its surroundings.

4. Conclusion

The preparation and implementation of a new census methodology require time. It is very important to emphasize here the introduction of relevant legislative regulations and the process for reviewing and improving the quality of administrative records. Extensive work on the evaluation of the quality of administrative records and their use by public statistics should be considered as a great achievement of the 2011 census in Poland. Of course, the process of evaluation and improvement of records is a continuous one, and work on this has barely started.

The above discussion is an attempt made to evaluate the results of the census, and also the population register indirectly. Using the methods of demographic analysis as concerns fertility, mortality, migration and projections, a comparative analysis was conducted. Different cross-sections according to sex, age and place of residence were considered. In each of the separate domains, adequacy and divergence assessments were provided and accompanied by substantive explanations. Among the results obtained, we can specify:

The survival rates for young people (age groups: 9-13, 16-19, 30 years and 33-35 years) between censuses were exceeding one.

This situation may suggest that not all the people were enumerated in 2002 census. In particular, the lack of infants in the 2002 census was already observed in earlier studies (Paradysz 2010). The deficiencies observed for people aged 30-35 years may reflect emigrants who were not counted in the census conducted by the conventional method, but were included in the register-based census (as they are not removed from the register). The population register is not free from erroneous enumerations, repetitions and omissions and others. Thus, further work on its improvement is necessary.

For better identification of the usually resident population, it should be considered to extend the residence analysis. Additional methods of assessing compliance with the definition criteria might be introduced, for example by examining activity of individuals in other registers to ensure that only individuals with a pre-defined number of "signs of life" are counted in the census.

1. The lack of children aged 0 completed years observed in the census was most likely due to a delay in the registry. A two-week delay corresponds to about 1/3 of the monthly number of births and the missing number of infants - which is consistent with the statutorily specified time to register one’s child. This shortcoming requires adequate solutions.
2. The surplus of children aged 1-4 years observed in the census can be explained by children born in exile and registered in Poland to obtain Polish citizenship. This is facilitated by a simple registration procedure. This was partially confirmed by mirror statistical analysis for England and Wales.
3. Trends characterizing changes in population aged 20-30-40 years could be explained through an in-depth comparative analysis of data from successive

censuses and current population statistics. Diversity of data sources and the principles of demographic analysis allowed a discussion on the 'surplus' of population aged 20-29 years and the 'shortage' of the population aged 29-40 years in the cross-section of large cities and surrounding areas.

A quality assessment system is built into procedures of conducting a census, which is a survey that uses various data sources. The implementation of the mechanism for mutual control, research compliance and comparative analyses results in more reliable information. The use of variety of sources promotes their in-depth exploration also in terms of demographic analyses. On the other hand, the use of multiple sources of information makes it possible to obtain inconsistent results being a 'natural' danger. In consequence, divergent estimates force attempts to provide consistent estimates and explain reasons of the differences. The 2011 census was a complex procedure, which for a single individual combined information from two different types of sources: registers and sample surveys. The analysis using integrated data (register and sample survey) requires the development of new theoretical concepts (Zhang, 2011, 2012). The census based on multiple data sources enforced application of modern methodology. In the case of 2011 census in Poland it meant great scientific work related to the development of modern statistical methods such as calibration, statistical data integration, GIS, estimation for small domains, etc. The advantages of the applied methodology are not only difficult to measure and assess, but they should be considered in terms of a precondition for further development of statistics in the most desirable sense: the development of science in response to the needs.

REFERENCES

AL, P. G., BAKKER, B. F. M., (2000). Re-engineering Social Statistics by Micro-integration of Different Sources; An Introduction. "Netherlands Official Statistics". Volume 15. Summer 2000. Special issue: Integrating Administrative Registers and Household Surveys. Statistics Netherlands.

BAFFOUR, B., KING, T., VALENTE, P., (2013). The Modern Census: Evolution, Examples and Evaluation. "International Statistical Review" (2013), 81, 3, pp. 407–425, DOI: 10.1111/insr.12036

BALDRIGE, M., BROWN, C. J., JONES, S., KEANE, J. G., (1985). Evaluating Censuses of Population and Housing. Department of Commerce. United States of America/US Bureau of the Census.

BARWIŃSKI, M., (2014). Ethnic Structure of Polish Population in the Light of the Results of 2011 Census [Struktura narodowościowa Polski w świetle wyników spisu powszechnego z 2011 roku]. „Przegląd Geograficzny", 86 (2), pp. 217–241.

BERKA, C., HUMER, S., LENK, M., MOSER, M., RECHTA, H., SCHWERER, E., (2010). A quality framework for statistics based on administrative data sources using the example of the Austrian census 2011, "Austrian Journal of Statistics", No. 39.

BERKA, C., HUMER, S., LENK, M., MOSER, M., RECHTA, H., SCHWERER, E., (2012). Combination of evidence from multiple administrative data sources: quality assessment of the Austrian register-based census 2011. "Statistica Neerlandica", Vol. 66, Issue 1, 18–33.

BETHLEHEM, J., (2009). The Rise of Survey Sampling, Discussion paper (09015), Statistics Netherlands.

BORKET, M., PENNINX, R., (2011). Policymaking in the Field of Migration and Integration in Europe: An Introduction. [in:] G. Zincone, R. Penninx, M. Borkert (ed.), Migration Policymaking in Europe. The Dynamics of Actors and Contexts in Past and Present, Amsterdam: IMISCOE Research. Amsterdam University Press, pp. 7–21.

BROWN, J., ABBOTT, O., DIAMOND, I., (2006). Dependence in the 2001 one-number census project. "Journal of Royal Statistical Society", A, 169, pp. 883–902.

CBOS, (2013). Was It Worth It to Change the System? Assessment of Social Change After 1989. [Czy warto było zmieniać ustrój? Społeczna ocena przemian po 1989 roku]. Statement from research, BS/73/2013.

CHATFIELD, C., (2002). Confession of a Pragmatic Statistician. "The Statistician", Vol. 51, No. 1. (2002), pp. 1–20.

CSO, (2012). The Report on the Results. The National Census of Population and Housing 2011 [Raport z wyników. Narodowy Spis Powszechny Ludności I Mieszkań 2011], Warsaw.

CSO, (2013). Higher Education Institutions and their Finances in 2012 [Szkoły wyższe i ich finanse w 2012 roku], Warsaw.

FIHEL, A., KACZMARCZYK, P., OKÓLSKI, M., (2006). Labour Mobility in the Enlarged European Union. CMR Working Paper, No. 14/72.

FOSU, G. B., (2001). Evaluation of population census data through demographic analysis. Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects Statistics Division Department of Economic and Social Affairs United Nations Secretariat New York, 7-10 August 2001.

GOŁATA, E., (2009). Economic Activity in Population Census 2011 and Administration Resources [Aktywność ekonomiczna ludności w NSP'2011 a zasoby rejestrów administracyjnych] [in:] Methods and Sources of Information in the public statistics [Metody i Źródła Pozyskiwania Informacji w Statystyce Publicznej], Published by: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan.

GOŁATA, E., (2012). Population Census And Truth (Spis ludności i prawda). „Studia Demograficzne". Polish Academy of Sciences, 161 (1), pp. 23–55.

HOLT, T., (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. Discussion. "The American Statistician", Vol. 61 (1), pp. 1–8.

JANTA, B., (2013). Polish Migrants' Reproductive Behaviour in the United Kingdom. "Studia migracyjne – Przegląd Polonijny", No. 3, 63–96.

JOŃCZY, R., (2010). International Migration From Rural Areas of Opolskie Province After the Polish Accession to the European Union. Selected Economic and Demographic Aspects [Migracje zagraniczne z obszarów wiejskich województwa opolskiego po akcesji Polski do Unii Europejskiej. Wybrane aspekty ekonomiczne i demograficzne]. Published by: Wydawnictwo Instytut Śląski Sp. z o.o. Opole–Wroclaw.

JÓZEFOWSKI, T., RYNARZEWSKA-PIETRZAK, B., (2010). Evaluation of the Possibilities to Use the PESEL Register in the Population Census [Ocena możliwości wykorzystania rejestru PESEL w spisie ludności]. [in:] The Measurement and Information in the Economy [Pomiar i informacja w gospodarce]. Scientific Notebook of the Faculty of Informatics and Electronic Economy. E. Gołata (ed.). Published by: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan.

KICINGER, A., KORYŚ, I., (2011). A new Field to Conquer and Manage: Migration Policymaking in Poland after 1989, [in:] G. Zincone, R. Penninx, M. Borkert (ed.), Migration Policymaking in Europe. The Dynamics of Actors and Contexts in Past and Present, Amsterdam: IMISCOE Research. Amsterdam University Press, pp. 347–376.

KLIMANEK, T., (2012). The Results of the Study of Migratory Behavior of the Inhabitants of Poznan [Wyniki badania zachowań migracyjnych mieszkańców Poznania], [in:] Migration of the Residents of Large Cities [Migracje mieszkańców dużych miast], E. Gołata (ed.). Published by: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan.

KOBUS, P., SMOLKA, M., NOWAKOWSKA, G., (2009). Report on the Works on the Files of Comprehensive Social Insurance System [Raport z prac na zbiorach kompleksowego systemu informatycznego ZUS], CSO: Central Bureau of Census [Centralne Biuro Spisów GUS], Warsaw.

KORDOS, J., (2007). Some Aspects of Post-Enumeration Surveys in Poland. „Statistics in Transition – new series", Vol. 8 (3). pp. 563–576.

KORDOS, J., (2012). The Interplay Between Sample Survey Theory And Practice In Poland [Współzależność pomiędzy rozwojem teorii i praktyki badań reprezentacyjnych w Polsce] „Przegląd Statystyczny" Special Issue 1 – 2012. pp. 61–68. Polish Academy of Sciences.

KUC, M., (2014). Najstarszy człowiek świata żyje w Polsce. W bazie PESEL. [The oldest man in the world lives in Poland. In the PESEL database.], "Gazeta Prawna". 4.06.2014. Warsaw. http://prawo.gazetaprawna.pl/artykuly/801115,najstarszy-czlowiek-swiata-zyje-w-polsce-w-bazie-pesel.html.

LONGVA, THOMSEN, SEVEREIDE, (1998). Reducing Costs of Censuses in Norway Through Use of Administrative Registers. Statistics Norway 1998.

MRYSKALA, P., (2011). Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland. A Handbook of Principles and Practices. Statistics Finland.

NEWELL, R., SMALLWOOD, S., (2010). A cross country review of the validation and/or adjustment of census data. Population Trends, No. 141, Office for National Statistics.

ONS, (2013). Detailed Country of Birth and Nationality Analysis from the 2011 Census of England and Wales.

PARADYSZ, J., (2002). The Non-random Errors in the Fertility Survey within the 1970 Population Census (O błędach nielosowych w badaniu dzietności kobiet w ramach Narodowego Spisu Powszechnego 1970). [in:] Population censuses in the Republic of Poland from 1921 to 2002. Selected Writings by Demographers [Spisy ludności Rzeczypospolitej Polskiej 1921–2002. Wybór pism demografów]. Z. Strzelecki. T. Toczyński (eds.). Polish Demographic Society. Central Statistical Office [Polskie Towarzystwo Demograficzne. Główny Urząd Statystyczny], Warsaw, pp. 479–482.

PARADYSZ, J., (2010). The Necessity for Use of Indirect Estimation in Population Censuses [Konieczność estymacji pośredniej na użytek spisów powszechnych]. [in:] The Measurement and Information in the Economy [Pomiar i informacja w gospodarce]. Scientific Notebook of the Faculty of Informatics and Electronic Economy. E. Gołata (ed.). Published by: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan.

PLEWIS, I., SIMPSON, L., WILLIAMSON, P. (2011). Census 2011: Independent review of coverage assessment, adjustment and quality assurance. Available at: www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-censusproject/independent-assessments/independent-review-of-coverage-assessment-- adjustment-and-quality-assurance/independent-review-final-report.pdf

REGULATION (EU), No 1260/2013 of the European Parliament and of the Council of 20 November 2013 on European demographic statistics Text with EEA relevance. Available at: http://eur-lex.europa.eu/legal content/EN/TXT/?uri=CELEX:32013R1260.

REPORT, (2011). The National Census of Population and Housing 2011 in Assessment of the Ukrainian Minority [Raport. Narodowy Spis Powszechny Ludności i Mieszkań 2011 w ocenie mniejszości ukraińskiej]. 2011. Association of Ukrainians in Poland, Warsaw.

ROSZKA, W., (2013). Statistical Data Integration in Socio-Economic Research [Statystyczna integracja danych w badaniach społeczno-ekonomicznych]. PhD Thesis. Poznan University of Economics. Available at: http://www.wbc.poznan.pl/Content/265243/Roszka_Wojciech_doktorat.pdf.

RUGGLES, S., (2014). Big Microdata for Population Research. "Demography" 51, 287–297.

SAKSON, B., (2002). The impact of "invisible" migration of the eighties on the demographic structure of Polish Population [Wpływ "niewidzialnych" migracji zagranicznych lat osiemdziesiątych na struktury demograficzne Polski]. Monographs and Studies, No. 481, Warsaw School of Economics. Warsaw.

ŚLESZYŃSKI, P., (2004). Regional Disparities Between Population According to the 2002 Population Census and Based on the Current Recording [Regionalne różnice pomiędzy liczbą ludności według narodowego spisu powszechnego w 2002 r. i rejestrowaną na podstawie ewidencji bieżącej], „Studia Demograficzne", 145 (1), pp. 93–103.

ŚLESZYŃSKI, P., (2005). Differences in the Numer of Population Disclosed in the National Census 2002 [Różnice liczby ludności ujawnione w Narodowym Spisie Powszechnym 2002]. „Przegląd Geograficzny", 77 (2), pp.193–212.

Statistics Denmark, (1995). Statistics on Persons in Denmark – a register-based statistical system. Statistics Denmark.

STATISTICS FINLAND, (2004). Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland. Tilastokeskus. Statistikcentralen. Statistics Finland. Helsinki.

STATISTICS NETHERLAND, (2004). The Dutch Virtual Census of 2001. Analysis and Methodology. Ed. Eric Schulte Nordholt, Marijke Hartgers and Rita Gircour. Statistics Netherlands, Voorburg/Heerlen, 2004.

UN, (2006). Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. 2006. New York. Geneva. United Nations Economic Commissions for Europe. Statistical Office of the European Communities. ECE/CES/STAT/NONE/2006/4. UN Publications.

UN, (2007). Register-based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics. United Nations Economic Commission For Europe. United Nations. New York, Geneva, 2007.

UN, (2009). Manual on Census Evaluation. Post Enumeration Surveys. Demographic And Social Statistics Branch, United Nations Statistics Division.

UN, (2010). Post Enumeration Surveys Operational Guidelines. Technical Report. United Nations Statistics Division New York, April 2010.

UN, (2010a). Report on the Results of a Survey on Census Methods used by Countries in the 2010 Census Round. Department of Economic and Social Affairs United Nations. New York, Available at: http://unstats.un.org/unsd/census2010.htm.

UN, (2011). Using Administrative and Secondary Sources for Official Statistics. United Nations Economic Commission For Europe. United Nations. New York, Geneva, 2011.

UN, (2012). Overview of the 2010 round of population and housing censuses in the UNECE region. ECE/CES/ GE.41/2012/20, 16 May 2012.

UN, (2012a). Lessons learned from the Population and Housing Census in Poland. Note by the Central Statistical Office of Poland. Economic Commission for Europe.

UN, (2012b). Quality assessment of register-based census data in Austria, ECE/CES. Working Paper 16, 16 May 2012.

UN, (2013). Overview of National Experiences for Population and Housing Censuses of the 2010 Round. United Nations Statistics Division, 2013, New York June 2013.

UN, (2013a). Field operations. Legislation. Lessons Learned: Key Results of the UNECE Survey on National Census Practices and First Proposals About the CES Recommendations for the 2020 Census Round. Note by the UNECE Steering Group on Population and Housing Censuses. Geneva. 30 September – 3 October 2013. ECE/CES/GE.41/2013/6.

UN, (2013b). Census quality and coverage: Key results of the UNECE Survey on National Census Practices. Note by the UNECE Task Force on census coverage and quality. ECE/CES/ GE.41/2013/8. Geneva. 30 September – 3 October 2013.

WALLER, L., BERRINGTON A., RAYMER J., (2014). New Insights into the Fertility Patterns of Recent Polish Migrants in the United Kingdom. "Journal of Population Research". DOI: 10.1007/s12546-014-9125-5.

WALLGREN, A., WALLGREN, B., (2007, 2014). Register-based Statistics. Statistical Methods for Administrative Data. John Wiley & Sons. Ltd.

WALLGREN, A., WALLGREN, B., (2013). Quality Assessment in Systems with Registers and Sample Surveys. http://www.statistics.gov.hk/wsc/IPS078-P2-S.pdf.

VALENTE, P., (2010). Census taking in Europe: how are populations counted in 2010? "Population & Societies", No. 467.

VAN DER LAAN, P., (2000). Integrating administrative registers and household surveys, "Netherlands Official Statistics", Vol. 15, Summer 2000, Special issue: Integrating administrative registers and household surveys, Statistics Netherlands, Voorburg/Heerlen.

ZASĘPA, R., (1993). Use of Sampling Methods in Population Censuses in Poland, "Statistics in Transition", Vol. 1, Number 1, June 1993.

ZHANG, L.-C., (2011). A Unit-Error Theory for Register-Based Household Statistics. „Journal of Official Statistics", Vol. 27(3), pp. 415–432.

ZHANG, L.-C., (2012). Topics of Statistical Theory for Register-based Statistics and Data Integration. Statistica Neerlandica, Vol. 66, No. 1. pp. 41–63.

ZHANG, L.-C., (2013). Population Size Estimation Based on Multiple Lists. Uncertainty Analysis for Categorical Data Fusion. Open Lecture given at Poznan University of Economics.

ZUMPE, J., DORMON, O., JEFFERIES, J., (2012). Childbearing Among UK Born and Non-UK Born Women Living in the UK, ONS, 25 October 2012.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 659–670

MODELLING ROAD TRAFFIC CRASHES USING SPATIAL AUTOREGRESSIVE MODEL WITH ADDITIONAL ENDOGENOUS VARIABLE

Olusanya Elisa Olubusoye [1], **Grace Oluwatoyin Korter** [2 3], **Afees Adebare Salisu** [4]

ABSTRACT

Road traffic crashes have become a global issue of concern because of the number of deaths and injuries. The model of interest is a linear cross sectional Spatial Autoregressive (SAR) model with additional endogenous variables, exogenous variables and SAR disturbances. The focus is on RTC in Oyo state, Nigeria. The number of RTC in each LGA of the state is the dependent variable. A 33×33 weights matrix; travel density; land area and major road length of each LGA were used as exogenous variables and population was the IV. The objective is to determine the hotspots and examine whether the number of RTC cases in a given LGA is affected by the number of RTC cases of neighbouring LGAs and an instrumental variable. The hotspots include Oluyole, Ido, Akinyele, Egbeda, Atiba, Oyo East, and Ogbomosho South LGAs. The study concludes that the number of RTC in a given LGA is affected by the number of RTC in contiguous LGAs. The policy implication is that road safety and security measures must be administered simultaneously to LGAs with high concentration of RTC and their neighbours to achieve significant remedial effect.

Key words: road traffic crashes, generalized spatial two-stage least squares estimator, instrumental-variable estimation, spillover effects.

1. Introduction

Spatial models that accommodate forms of cross-unit interactions are features of interest in social sciences, biostatistics and geographic sciences. A simple spatial model augments the linear regression model by including an additional

[1] Department of Statistics, University of Ibadan, Oyo State, Nigeria.
E-mail: oe.olubusoye@mail.ui.edu.ng ; busoye2001@ýahoo.com.
[2] Department of Statistics, University of Ibadan, Oyo State, Nigeria. E-mail: kortergrace@gmail.com.
[3] Department Mathematics/Statistics, Federal Polytechnic, Offa, Kwara State, Nigeria.
[4] Department of Economics, University of Ibadan, Oyo State, Nigeria.
E-mail: aa.salisu@mail.ui.edu.ng; adebare1@yahoo.com.

Right Hand Side (RHS) variable known as a spatial lag and considers spill over effects either in the dependent variable or in the disturbance term. When the focus is on the dependent variable, the spillover effect is modelled by including a RHS variable known as a spatial lag. Each observation of the spatial-lag variable is a weighted average of the values of the dependent variable observed for the other cross-sectional units. The matrix containing the weights is known as the spatial-weighting matrix. This model is frequently referred to as a Spatial-Autoregressive (SAR) model.

A generalized version of this model also allows for the disturbances to be generated by a SAR process and for the exogenous RHS variables to be spatial lags of exogenous variables. The combined SAR model with SAR disturbances is referred to as SARAR model. In modelling the outcome for each unit as dependent on a weighted average of the outcomes of other units, SARAR models determine outcomes simultaneously. This simultaneity implies that the ordinary least squares estimator will not be consistent. The RHS variables are a spatial lag of the dependent variable, exogenous variables and spatial lags of the exogenous variables.

The SARAR model allows for additional endogenous RHS variables. Thus, the model of interest is a linear cross sectional SAR model with additional endogenous variables, exogenous variables and SAR disturbances.

The focus is on RTC in Oyo state,Nigeria.The number of RTC in each LGA of the state is the dependent variable. A 33×33 weights matrix; travel density; land area and major road length of each LGA were used as exogenous variables and population was the IV.

Road accidents are a global scourge characteristic of our technological era with a list of victims that grows insidiously longer day by day. Our roads which are supposed to take us places often become sources of sorrow and venues of loss. Most families have found themselves mourning, surrounded by indifference that is all too common as if this was a price or an unavoidable tribute the societies have to pay for the right to travel. The frequencies of deaths, injuries, environmental degradation, material losses and the economic impact of this seemingly unpreventable phenomenon is a global issue of concern.

Worst still, road traffic injuries cost low and middle income countries between 1 and 2 percent of their gross national product which is more than the total development aid received by these countries (WHO, 2004). In Africa, RTC are adjudged to be the second leading cause of deaths between the ages of 15 and 44. In Nigeria, RTC death rate is 162 deaths per 100,000 populations (Ogbodo and Nduoma, 2011). This is against the world average of 22 deaths per 100,000 populations (Sukhai et al., 2011).

Thus, the Nigerian RTC death rate is disproportionately high when compared to the world average by over 636 percent. Young adults' deaths in RTC will affect the workforce of the nation. This will in turn impact negatively on economic activities and indirectly cause a devaluation of the country's gross domestic product.

Consequent upon similar occurrences around the globe, the United Nations General Assembly have designated 2011 to 2020 as a decade of action on road safety for dedicated intervention by governments to bring down the estimated rise in deaths from RTC by 50 percent (Ki-moon, 2012).

This paper, therefore, considers spatial autoregressive dependence in RTC cases, the error term and an additional endogenous RHS variable. In other words, the objective is to determine the hotspots and examine whether the number of RTC cases in a given LGA is affected by the number of RTC cases of neighbouring LGAs and an instrumental variable. The basic underlying assumptions include the existence of spillover effects across the study area and that RTC are caused by certain actions of man, which may not necessarily be aimed at causing accidents.

Section 2 describes past works. Section 3 describes the SARAR model with an additional endogenous variable and the estimation techniques. Section 4 discusses data and results. Section 5 discusses the summary and conclusion.

2. Literature review

Road traffic injuries are a major public health problem. Leveque et al. (2001) chose Years of Potential Life Lost (YPLL) to analyse the trends during the period 1974-1994 and the relative impact of the traffic injuries death on total mortality and on total avoidable mortality in Belgium. The paper analyzed the geographical trends over a 20-year period at the district level. The geographical analysis showed marked differences between districts. Even though a favourable trend was observed for the traffic injuries and deaths in Belgium it was necessary to highlight the important slowing down of this trend during the most recent years. The study concluded it was also necessary to underline the importance of geographical disparities in the distribution of YPLL rates within the entire population.

Cirera et al. (2001) described the characteristics of Motor-Vehicle (MV) injury cases admitted to Emergency Departments (ED), and assessed factors related to injury severity and hospital admission. The subjects were MV injury patients, aged 16 or over and admitted to four EDs in the city of Barcelona (Spain), from July 1995 to June 1996. Severity was assessed with the abbreviated injury scale and the injury severity score. Univariate and bivariate descriptive statistical analyses were performed, as well as multiple logistic regressions. For the 3791 MV-injury cases included in the study period, a larger contribution of cases was noted for males (63.1%), for cases younger than 30 years (55.3%) and for motorcycle or moped occupants (47.1%). After adjusting for age, sex and the presence of multiple injuries, pedestrians, followed by moped and motorcycle occupants were at a higher risk of a more severe injury. Correspondingly, these user groups also showed a higher likelihood of a hospital admission when

attended to in an ED. Injury cases attended to in the ED during night hours were also at a higher risk of a hospital admission.

Geurts et al. (2005) compared characteristics of accidents occurring in black zones to those scattered all over the road. Identifying dangerous accident locations and profiling them in terms of accident-related data and location/environmental characteristics provided new insights into the complexity and causes of road accidents. The case study was a Belgian peri urban region. A technique of frequent item sets (data mining) was applied for automatically identifying accident circumstance that frequently occurred together for accidents located in an outside black zone. Results showed that accidents occurring in black zones were characterized by left-turns at signalized intersections, collisions with pedestrians, losing control of the vehicle (run-off-roadway) and rainy weather conditions. Accidents occurring outside black zones (scattered in space) were characterized by left turns on intersections with traffic signs, head-on collisions and drunken road users. Furthermore, parallel collisions and accidents on highways or roads with separated lanes occurring at night or during the weekend were frequently occurring accident patterns for all accident locations.

Labinjo et al. (2009) explored the epidemiology of Road Traffic Injury (RTI) in Nigeria. The focus was on populations affected and the risk factors for RTI. The RTI rates for rural and urban respondents were not significantly different. Increased risk of injury was associated with male gender among those aged 18-44 years. Simple extrapolations from the survey suggested that over 4 million people were injured and as many as 200,000 potentially killed due to this menace annually.

Aderamo (2012), in a bid to proffer measures to reduce the scourge of road traffic casualties examined the spatial variation of RTC in Nigeria. The models developed relate total number of road accidents, population estimates, length of roads and number of registered vehicles for the country. The regression results revealed that MV deaths had a positive association with population estimate and length of roads. Also, population estimate and length of roads had a significant effect on MV injuries.

In order to enhance prioritization of the burden of RTC, Chandran et al. (2013) calculated Years of Life Lost (YLL) and reduction in life expectancy using population and crash data from Brazil's ministries of health and transport. The potential for reduction in crash mortality was calculated for hypothetical scenarios reducing death rates to those of the best performing region and age category. For males and females at birth, RTC reduced the life expectancy by 0.8 and 0.2 years respectively. The study concluded that many YLL for men and women could be averted if all rates matched those of the lowest-risk region and age category.

To reduce RTC casualties and improve safety and security on roads, usually, highway improvement project selection requires screening thousands of road segments with respect to RTC for further analysis and final selection into improvement projects. Kelle et al. (2013) described a two-step procedure for selecting potential RTC locations for inclusion in highway improvement projects.

The first step of the proposed methodology used odds against observing a given crash count, injury count and run-off road count as measures of risk and a multi-criteria pre-selection technique with the objective of decreasing the number of prospective improvement locations. The second step was based on a composite efficiency measure of estimated cost, benefit and hazard assessment under budget constraint. To demonstrate the two-step methodology, the study analyzed 4 years of accident data at 23000 potential locations, from which final projects were selected.

With a special focus on truck drivers, Rancourt et al. (2013) developed different scheduling algorithms embedded within a tabu search heuristic. This was in attempt to improve safety amongst long haul carriers in vehicle routing and scheduling in the United States. The methods and results confirmed the benefits of using a sophisticated scheduling procedure for long haul transportation.

Metaheuristic algorithms, such as simulated annealing and tabu search are popular solution techniques for vehicle routing problems (VRPs). Harwood et al. (2013) focussed on single VRP similar to travelling salesman problem, and investigated the potential for using estimation methods on simple models with time-invariant costs, mimicking the effects of road congestion. Working with standard VRPs, where the costs of the arcs did not vary with advancing time, changes to the total cost were evaluated following a neighbourhood move simple process. In the cases where a time-varying aspect such as congestion was included in the costs, the calculations became estimations rather than exact values.

3. SARAR model with additional endogenous variables and estimation techniques

$$y = Y\pi + X\beta + \lambda W_1 y + u \tag{1}$$

$$u = \rho W_2 u + \varepsilon \tag{2}$$

where y is an $n \times 1$ vector of observations on the dependent variable; Y is an $n \times p$ matrix of observations on p RHS endogenous variables; and π is the corresponding $p \times 1$ parameter vector; X is an $n \times k$ matrix of observations on k RHS exogenous variables (where some of the variables may be spatial lags of exogenous variables), and β is the corresponding $p \times 1$parameter vector; W_1 and W_2 are $n \times n$ spatial weighting matrices (with 0 diagonal elements); $W_1 y$ and $W_2 u$ are $n \times 1$ vectors typically referred to as spatial lags, and λ and ρ are the corresponding scalar parameters typically referred to as SAR parameters and ε is an $n \times 1$ vector of innovations.

The model in equations 1 and 2 is a SARAR model with exogenous regressors and additional endogenous regressors. Spatial interactions are modelled through

spatial lags, and the model allows for spatial interactions in the dependent variable, the exogenous variables and the disturbances.

The model in equations 1 and 2 is a first-order SAR process with first order SAR disturbances, also referred to as a SARAR(1,1) model, which is a special case of the more general SARAR (*p, q*) model. We refer to a SARAR (1, 1) model as a SARAR model. Setting $\rho = 0$ yields the SAR model $y = Y\pi + X\beta + \lambda W_1 y + \varepsilon$. Setting $\lambda = 0$ yields the model $y = Y\pi + X\beta + u$ with $u = \rho W_2 u + \varepsilon$, which is sometimes referred to as the SAR error model. Setting $\rho = 0$ and $\lambda = 0$ causes the model to reduce to a linear regression model with endogenous variables.

The spatial-weighting matrices W_1 and W_2 are taken to be known and nonstochastic. These matrices are part of the model definition, and in many applications, $W_1 = W_2$. Let $\bar{y} = W_1 y$, let $\bar{y}_i$ and y_i denote the ith element of $\bar{y}$ and y , respectively, and let w_{ij} denote the $(i, j)th$ element of W_1 . Then, $\bar{y}_i = \sum_{j=1}^{n} w_{ij} y_j$.

The weights w_{ij} will typically be modelled as inversely related to some measure of distance between the units. The SAR parameter λ measures the extent of these interactions.

The innovations ε are assumed to be independent and identically distributed or independent but heteroskedastically distributed, Drukker et al. (2013).

The Generalized Spatial Two Stage Least Squares (GS2SLS) estimation technique was used.

4. Analysis and discussion of results

4.1. Data description

In February 1988, the Federal Government created the Federal Road Safety Commission (FRSC) through Decree No. 45 of 1988 as amended by Decree 35 of 1992 referred to in the statute books as the FRSC Act cap 141 Laws of the Federation of Nigeria (LFN). Prior to this time, the Nigerian Police Force established in 1930 was saddled with the responsibility of keeping RTC records. In most instances, especially when the form of RTC was not serious, such events were not usually reported. Thus, it is possible that this data does not contain all the RTC that occurred within the study time and area. Nonetheless, data for this study are reliable and despite possible omissions the findings will not be negatively influenced.

Data on number of RTC and traffic volume was collected from the RS11.3 FRSC Oyo sector command. This research enjoyed the good cooperation of the

FRSC, which is a good potential to obtain quality results. The study focussed on area of land, total length of major roads, travel densities and the residential population for every LGA.

The approximate locations of the road crashes were estimated from the records obtained from the RS 11.3 Oyo FRSC command. The records provided an indicative description of the locality where the RTC occurred and in some cases the site was described using nearest landmarks such as filling stations, roundabouts, stores, markets, garages, institutions or road intersections. Therefore, using the locations and/or nearest landmarks provided in the records together with the Google Earth image, the existing digital road networks and knowledge of the area of study, it was easy to place points on the approximate locations where RTC occurred.

The Google Earth image was particularly helpful because it provided a photographic view of the area of study together with the associated landmarks and road networks. Through the instrumentality of Global Positioning System (GPS) and the information on locations of RTC for each of the unit commands, the coordinates of each RTC location were obtained on the geographic coordinates system of the world.

The coordinate locations generated were subsequently exported into ArcGIS and were plotted as point locations. The point locations, which represent RTC locations, were overlaid on the road network of the LGAs of Oyo State. This made it possible to clip the Oyo state geo-referenced map within the ArcGIS environment to ascertain where each RTC location falls. Therefore, the number of RTC points within each LGA was counted and recorded accordingly.

The Geographic Information System (GIS) was used to create polygon shapefile for the study area. The shapefile was used to create spatial contiguity weight matrix based on the queen criteria. The lengths of the roads and the area encompassing each LGA were calculated using the measuring tool in the software environment.

Population was defined as the population figures for each LGA as reported by the National Bureau of Statistics for the 2006 population census in Nigeria. Travel density was defined as traffic count for each FRSC unit command divided by total major road length in kilometres within each LGA. Major road length was the total length of roads in kilometres from each settlement to another within each LGA. Area was defined as the area per square kilometre encompassing each Local Government Authority. Accident was defined as the total cases of RTC recorded in each LGA whose location was identified by taking the longitudes and latitudes. The logarithms of all the variables were taken for the purpose of this study.

4.2. RESULTS

Figure 1 shows the distribution of RTC across LGAs, with darker colours (blue) representing higher values of the dependent variable. Spatial patterns of RTC are clearly visible. The hotspots for RTC are within Oluyole, Ido, Akinyele,

Egbeda, Atiba, Oyo East, and Ogbomosho South LGAs. This is followed by Ibadan North, Ibadan North West, Ibadan South East, Ibadan South West, Lagelu, Afijio, Oyo West, Ori-Ire and Ogbomosho North LGAs. Each of these categories of LGAs happens to be neighbours.

The content of the spatial-weighting matrix is summarized in Table 1. Some basic information about the normalized contiguity matrix, including the dimensions of the matrix and its storage is displayed. The number of neighbours found is reported as 156, with each LGA having 5 neighbours on average. Each LGA has a minimum of 2 neighbours and a maximum of 8 neighbours. This is an indication that no LGA is isolated, thus spatial dependency is bound to result based on neighbourhood characteristics. The frequency of RTC in a LGA is dependent on the frequency of RTC in the contiguous LGAs. Therefore, to achieve maximum remedial effect when administering road safety remedial measures to a LGA, it is significant to take cognizance of the neighbours of that particular LGA.

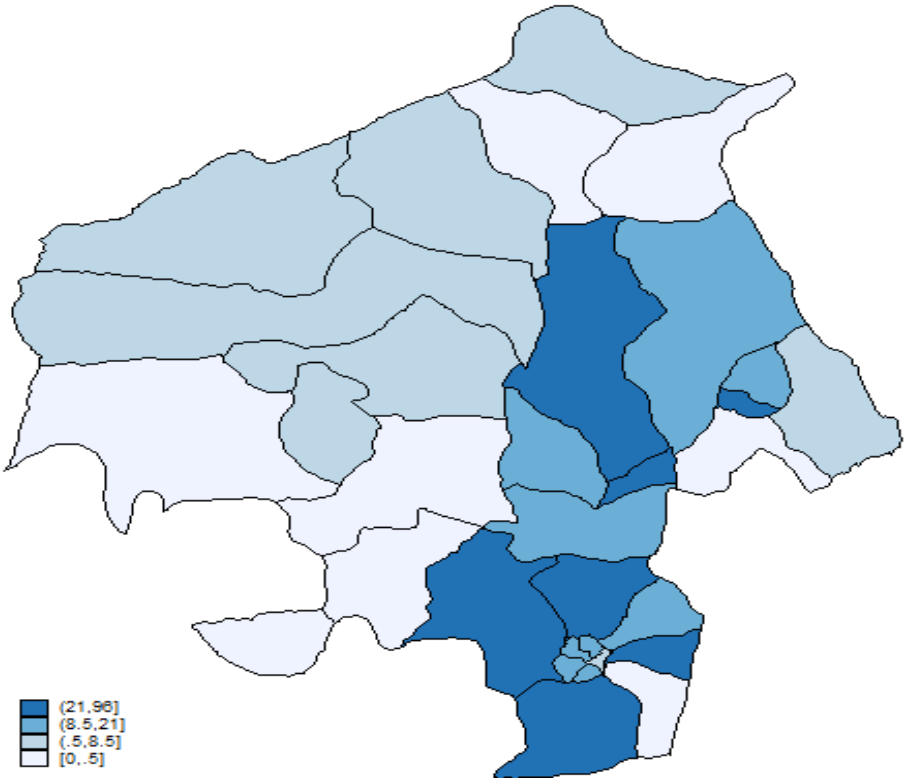


Figure 1. Number of Road Traffic Crashes cases for Local Government Areas in Oyo State, Nigeria

Table 1. Summary of Spatial Contiguity Weighting Matrix

Matrix	*Description*
Dimensions	33 *x* 33
Stored as Links	33 *x* 33
Total	156
Min	2
Mean	4.727273
Max	8

The GS2SLS parameter estimates for the spatial-autoregressive model with spatial-autoregressive disturbances (SARAR) with additional endogenous variable are as shown in Table 2**.**

Table 2. SARAR Model with Additional Endogenous Variable - GS2SLS

Variables	*Coefficient*	*Standard Error*	$P > \|Z\|$
Population	-0.417	0.85	0.624
Major Road Lengths	-0.104	0.54	0.849
Travel Density	0.009	0.18	0.962
Area	-0.175	0.34	0.606
Lambda(λ)	1.205	0.37	0.001
Rho (ρ)	-1.179	0.89	0.185

Given the normalization of the spatial-weighting matrix, the parameter space for λ and ρ is taken to be the interval −1 and 1 (Kelejian and Prucha, 2010). The estimate of λ is positive, large, and significant, indicating strong SAR dependence in RTC cases. In other words, the RTC frequencies for a given LGA is strongly affected by the RTC frequencies in the neighbouring LGAs. One possible explanation for this result may be the existence of freeways linking LGAs. Another may be the result of high travel density on major road networks of particular LGAs. Apparently, there is spillover effect of frequencies of RTC across the neighbourhood.

The estimated ρ is negative and moderate, indicating moderate spatial autocorrelation in the innovations. An exogenous shock to one LGA will cause moderate changes in RTC rate for neighbouring LGAs. For example, an addition of a new freeway linking a LGA will cause moderate changes in RTC rate for neighbouring LGAs. However, maximum road safety remedial effect will only be achieved through new freeways linking all contiguous LGAs simultaneously.

The estimated β vector does not have the same interpretation as in a simple linear model, because including a spatial lag of the dependent variable implies that the outcomes are determined simultaneously. The parameter estimates for population, major road lengths, travel densities and area are not significant. This means that each of the exogenous variables do not contribute significantly to the

incidence of RTC in Oyo state. The signs of the coefficients suggest the following:

Controlling for spatial lag and the LGAs, population is negatively related to the number of RTC occurring within the localities. All other things being equal, LGAs with larger residential populations tend to have fewer RTC. The coefficient indicates the expected number of RTC for every person living in the LGA. One percent increase in population will generate 0.42 percent decrease in the number of RTC cases within each LGA. The high number of people could be the result of high economic activities in the state, which may lead to high traffic density that will slow traffic and thus reduce the frequency of RTC.

Major road lengths characteristics produce a negative coefficient. This means the existence of a freeway link crossing a LGA inversely impact the incidence of RTC cases for the period. One percent increase in major road length will generate 0.10 percent decrease in the number of RTC cases within each LGA. Thus, an increase in major road lengths will generate a decrease in the number of RTC cases within each LGA. Invariably, the existence of more freeways linking LGAs will reduce travel densities, which may in turn reduce RTC. On the other hand, low travel densities could give rise to opportunities for increased speed by drivers which may result in a higher number of RTC.

Travel densities are positively related to the number of accidents. An increase in travel densities will lead to an increase in the number of RTC cases. This suggests that traffic generated tend to be associated with larger crashes. The high intensity of vehicles in urban areas as a result of economic activities and the existence of few major road networks in rural areas could be responsible for this result. One percent increase in travel densities will generate 0.01 percent increase in the number of accident cases.

The sign of the coefficient for area is negative. There are fewer RTC cases in larger LGAs. Every decrease in area per square kilometre encompassing a LGA will lead to an increase in the number of accident cases. This means an increase in the area of administration of LGAs will lead to less RTC cases in each LGA. Every one percent decrease in area per square kilometre encompassing a LGA will generate 0.18 percent increase in the number of RTC cases. This suggests that several factors other than location are responsible for RTC occurrences across the study area.

5. Summary and conclusion

The study concludes that the number of road traffic crashes in a given Local Government Area is affected by the number of road traffic crashes for neighbouring Local Government Areas. Thus, the policy implication of our result is that safety and security measures must be administered within Local Government Areas that have high concentration of road traffic crashes along with

their neighbouring Local Government Areas in order to achieve significant remedial effect.

Although this study examined a few possible exogenous variables for its investigations, improvements could be made by using more precisely measured variables. Also, the significance of more trip generators, such as land use activities, generators of economic activities and different employment activities could be the focus of future works. Studies revealing the causes of accidents such as speed violation, dangerous driving, driving under alcohol influence amongst several other factors will be useful for the implementation of safety and security measures. To enhance succinct investigation into road traffic crashes characteristics, data for more years should generally be used for accident modelling in the theoretical framework of spatial panels.

Additionally, future studies should examine post estimation issues such as how a change in one exogenous variable potentially changes the predicted values for all the observations of the dependent variable in Local Government Areas across the State. Finally, to reduce road traffic accidents, improve safety and security measures and achieve maximum remedial effect policy designs and decision-making on transportation and road networks construction/maintenance should incorporate spillover effects and take cognizance of exogenous variables across each Local Government Area in the state.

REFERENCES

ADERAMO, A. J., (2012). Spatial pattern of road traffic accident casualties in Nigeria. Mediterranean Journal of Social Sciences, 3 (2), pp. 61–72.

CHANDRAN, A., KAHN, G., SOUSA, T., PECHANSKY, F., BISHAI, D. M., GYDER, A. A., (2013). Impact of road traffic deaths on expected years of life lost and reduction in life expectancy in Brazil. Demography, 50(1), pp. 229–36.

CIRERA, E., PLASÈNCIA, A., FERRANDO, J., SEGUÍ-GÓMEZ, M., (2001). Factors associated with severity and hospital admission of motor-vehicle injury cases in a Southern European Urban Area. European Journal of Epidemiology, 17 (3), pp. 201–08.

DRUKKER, D. M., PRUCHA, I. R., RACIBORSKI, R., (2013). A command for estimating spatial-autoregressive models with spatial-autoregressive disturbances and additional endogenous variables. Stata Journal, 13, pp. 287–301.

FEDERAL ROAD SAFETY CORPS (FRSC), Federal Republic of Nigeria. (2012 RTC Report). RS11.3 Oyo Sector Command.

GUERTS, K., THOMAS, I., WETS, G., (2005). Understanding spatial concentrations of road accidents using frequent item sets. Accidents analysis and prevention, 37, pp. 787–99.

HARWOOD, K., MUMFORD, C., EGLESE, R., (2013). Investigating the use of metaheuristics for solving single vehicle routing problems with time varying traversal costs. The Journal of the Operational Research Society, 64 (1), pp. 34–47.

KELEJIAN, H. H., PRUCHA, I. R., (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of Econometrics, 157, pp. 53–67.

KELLE, P., SCHNEIDER, H., RASCHKE, C., SHIRAZI, H., (2013). Highway improvement project selection by joint consideration of cost benefit and risk criteria. The Journal of the Operational Research Society, 64 (3), pp. 313–25.

KI-MOON, B. (2012). Improving global road safety. [pdf] Roadsafe. Available at: <http://www.roadsafe.com/news/article.aspx?article=1636> [Accessed 17 March 2013].

LABINJO, M., JUILLARD, C., KOBUSINGYE, O. C., HYDER, A. A., (2009). The burden of road traffic injuries in Nigeria: results of a population based survey. Injury Prevention, 15 (3), p.157–62.

LEVEQUE, A., HUMBLET, P. C., LAGASSE, R., (2001). Premature avoidable deaths by road traffic injuries in Belgium: Trends and geographical disparities. European Journal of Epidemiology, 17 (9), pp. 841–45.

NATIONAL BUREAU OF STATISTICS, (2007). Social statistics in Nigeria. [pdf] Federal Republic of Nigeria. Available at: <http://www.nigerianstat.gov.ng/ext/latest_release/ssd09.pdf. > [Accessed 17 March 2013].

OGBODO, D., NDUOMA, E., (2011). FRSC: Nigerian roads second worst in the world. [pdf] Thisdaylive. Available at: < http://www.thisdaylive.com/articles/frsc-nigerian-roads-second-worst-in-the-world/103012/> [Accessed 17 March 2013].

RANCOURT, M., CORDEAU, J., LAPORTE, G., (2013). Long-Haul Vehicle routing and scheduling with working hour rules. Transportation Science, 47 (1), pp. 81–107.

SUKHAI, A., JONES, A. P., LOVE, B. S., HAYNES, R., (2011). Temporal variations in road traffic fatalities in South Africa. Accident Analysis and Prevention, 43, pp. 421–28.

WORLD HEALTH ORGANIZATION, (2004). World report on road traffic injury prevention. Summary World Health Organization. [pdf] World Health Organization. Available at: <http://www.who.int/violence_injury.../road.../world.../summary_en_rev.pdf > [Accessed 17 March 2013].

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 671–690

BAYESIAN ACCELERATED FAILURE TIME AND ITS APPLICATION IN CHEMOTHERAPY DRUG TREATMENT TRIAL

Kumar Prabhash[1], **Vijay M Patil**[2], **Vanita Noronha**[3], **Amit Joshi**[4], **Atanu Bhattacharjee**[5]

ABSTRACT

The Cox proportional hazards model (CPH) is normally applied in clinical trial data analysis, but it can generate severe problems with breaking the proportion hazard assumption. An accelerated failure time (AFT) is considered as an alternative to the proportional hazard model. The model can be used through consideration of different covariates of interest and random effects in each section. The model is simple to fit by using OpenBugs software and is revealed to be a good fit to the Chemotherapy data.

Key words: Survival Analysis, Faliure Time, Metronomic, Cisplatin.

1. Introduction

Accelerated Failure Time (AFT) models for time to event data give the scope to work with a parametric form of the hazard function. It is possible to accumulate random effects as frailty part, and they can be easily fitted with statistical software. However, standard statistical methods for survival analysis are dependent on asymptotic statistical inference. Bayesian methods can be an alternative choice for survival data analysis. This work was influenced by the analysis of data for drug treatment effect, comparison among chemotherapeutic patient on a duration of survival. This primary application is then applied to illustrate the methodology in this paper, but the illustrated approach is also appropriate to other types of study. The paper is organized as follows. A discussion of accelerated failure time models is given in section 3. The models are then illustrated in the Chemotherapy data and their fitness evaluated in different sections. The Data methodology is explained

[1]Department of Medical Oncology, Tata Memorial Hospital, India
[2]Department of Medical Oncology, Tata Memorial Hospital, India
[3]Department of Medical Oncology, Tata Memorial Hospital, India
[4]Department of Medical Oncology, Tata Memorial Hospital, India
[5]Department of Biometrics, Chiltern Clinical Research Ltd, Bangalore, India

in section 4. Section 5 provides the data modelling problem. Availability of different methods and their extension are detailed in section 6. Data analysis and results are provided in section 7. Section 8 gives the discussion and concluding remarks.

2. Accelerated Failure Time (AFT) Model

An Accelerated Failure Time (AFT) is a parametric model to give the alternative of the proportional hazard model. In the case of the proportional hazard model the effect of covariates is measured in terms of multiplication by a constant. In contrast to the proportional hazards model, the AFT model measured the effect of covariates with positive or negative terms by some constant discussed by David (2003). Let the $S_1(t)$ and $S_2(t)$ be the survival function in treatment Group-I and Group-II respectively. The AFT defines that $\phi > 0$ and

$$S_1(t) = S_2(\phi t) \tag{1}$$

The interpretation as follows : the percentage of individuals in the treatment Group-I that lives longer than time-(t) equal the percentage of individuals in the Group-II that lives longer than ϕt. Now the survival time t can be replaced by M_{1t} and M_{2t} as median survival time for treatment Group-I and Group-II. Now we have the hypothesis

$$S_1(M_{1t}) = S_2(M_{2t}) = 0.5 \tag{2}$$

and

$$\phi M_{1t} = M_{2t} \tag{3}$$

The AFT is an extension over limited application of the Proportional Hazard (PH) model. The parameters of interest in regression analysis become more robust in AFT than PH models, especially for ignored covariates (Hougaard (1994). In this paper, we propose an AFT regression model by adding a random effect to explore the influencing factor for survival duration among chemotherapy-treated patients. The aim of this work is to explore the treatment effect of Metronomic chemotherapy in comparison with Cisplatin chemotherapy. Three methods are applied to compare the treatment effectiveness,i.e (I) AFT regression with right-censored observation (II) AFT through consideration of Correlation Structure and (III) Bayesian extension of AFT models. The data are analyzed using OpenBugs code. The DIC value is used to select the best fit parametric model (Spiegelhalte 2002).

3. Data Motivation

Conventional MTD based chemotherapy dosing approach has lead to unsatisfactory efficacy results with excess of toxicity in comparison with new modality of drug administration termed as metronomic chemotherapy. It is an alternative to the traditional chemotherapeutic treatments (Hanahan 2000). The experimental work of metronomic chemotherapy was done by Folkman and his colleagues (Folkman 1971 and Hanahan 1996). The efficacy of metronomic chemotherapy on tumor with lack of toxicity of metronomic chemotherapy in mouse model was encouraging (Klement 2002). Multiple clinical studies in different tumors at different body sites have confirmed the efficacy of metronomic chemotherapy. In head and neck cancers too; metronomic chemotherapy in a palliative setting has shown a promise. The standard Cisplatin based palliative chemotherapy in head and neck cancers leads to a marginal improvement in overall survival. This improvement comes at a cost of severe side effects; Cisplatin in high dose causes emesis, nephrotoxicity, electrolyte disturbances and neurotoxicity. Hence an alternative treatment strategy was warranted in this situation. The proposed study was conducted in the department of Medical Oncology. Tata Memorial Hospital (TMH), Mumbai between 2011 to 2013, India. Patients attending the outpatient department of Medical Oncology (TMH) were selected for the present study subject to fulfillment of the selection criteria. Patients warranting palliative chemotherapy in head and neck cancers were randomized into 2 arms. One arm received 3 weekly Cisplatin for 6 cycles and another arm received oral metronomic chemotherapy untill progression. These patients were followed untill death. In this study, we were interested in the survival disparities between Metronomic and Cisplatin. The duration of survival was of two types: Overall survival and Progression Free Survival. The individual-specific information for a patient that is used in this study is age (age of the patient at diagnosis in complete years), Overall survival time, Progression-free survival time, Previous treatment and type of treatment (Metronomic or Cisplatin). We have 110 patients of chemotherapeutic effect, 57 patients from Metronomic and 53 patients from Cisplatin therapeutic groups. Table 1 provides a summary of the characteristics of the chemotherapeutic patients included in this study. In Figure 1, we plot the median duration estimates for both Metronomic, and Cisplatin groups of overall survival and progression-free survival.

4. Modeling Problems

The PH and AFT models are two attractive choices for survival analysis. The non-parametric Kaplan-Meier (K-M) curve can be used as the pointer for selection of suitable models. Figure 1 reveals the comparison of treatment groups in terms of survival duration. The graphical exploration of K-M curve and cumulative survival function are given in Figure 2. As basic assumption about proportional hazards, it is expected that the difference between the two functions will be constant. But Figure 2 does not give the strong evidence about the pattern of such expectation. We find that the survival rate does alter noticeably with the duration and it emerges that the metronomic group tends to have higher survival rates than the Cisplatin therapeutic group. So, considering random effects in survival models should develop the estimates of the contributing factors. Instead, the both functions are nearly same for the initial few days of time, almost up to the 50 days and then differentiate. It gives an idea about the possible violation of the proportional hazard model's assumptions. It is reasonable that the hazard functions are nontrivial but identical at any time t=0 through gradual difference with the increment of t. However, it breaks the assumptions about constant hazard ratio assumption for the proportional hazard model. In the presence of non-proportionality occurrences, the accelerated failure time model is applicable into two sample frameworks. It assumes the equality about scale change in the hazard function over the period.

5. Methods

5.1. Regression with Right-Censored Observations

Let T be the follow-up times, C is the censoring indicator and X is the baseline covariates. The actual survival time is defined as $Z \sim Weibull(\gamma, \lambda)$. The density and hazard fnction of Z is denoted as

$$f_0(z) = \lambda \gamma z^{\gamma-1} exp(-\lambda z^{\gamma}), h_0 = \lambda \gamma z^{\gamma-1} \tag{4}$$

Further, the hazard function is defined with

$$h(z|X) = exp(\beta^T X) h_0(z) = exp(\beta^T X) \lambda \gamma z^{\gamma-1} \tag{5}$$

It is possible to formulate the AFT from Weibull distribution through con-

sideration of $-\mu/\sigma$, as Intercept, α as a regression parameter by

$$\gamma = \sigma^{-1}, \lambda = exp(-\mu/\sigma), \beta = -\alpha/\sigma \tag{6}$$

5.1.1 Maximum Likelihood Estimation of Parameters

The aim is to estimate the parameter vector (γ, λ, β) based on this data through maximum likelihood. The corresponding likelihood function is defined as ,

$$L_1(\gamma, \lambda, \beta) = \prod_{i=1}^{n} h(T_i|X)^{c_i} S(T_i|X) \tag{7}$$

In order to account for X_{1i} potentially being censored, make additional assumptions and modify the likelihood function as follows: first, we explicitly specify a distribution for X_{1i} by specifying the density f_θ where $\theta \in R^d$ indicates the parameterization of f, for $d \geq 1$. Note that they assume the same distribution for all observations. Often it seems sensible to assume $f_\theta = f_{\mu,\sigma^2}$ as Normal, maybe after taking the logarithm of X_{1i}. Define for each observation the binary random variable $R_i = 1\{X_{1i} \geq c_i\}$ that indicates the status of the observation, i.e. whether it is observed or left-censored. The probability mass function p_{R_i} of R_i is a simple Bernoulli distribution with success probability $\pi_i = P(X_{1i} > c_i) = \int_{c_i}^{\infty} f_\theta(x)dx$ that X_{1i} is observed.

5.2. AFT through consideration of Correlation Structure

Let T_i, C_i and X_i be the failure time, censoring time and $p \times 1$ covariate for the i^{th} subject. Further, T_i is conditionally not dependent on $C_i|Xi$. The semi-parametric AFT model is defined with

$$T_i = X_i^T \beta + \varepsilon_i, i = 1, ..n \tag{8}$$

The term β is a regression parameter, ε_i(error) is identically distributed with random variables. It is assumed that ε_i are free from X_i. The i^{th} individuals observed data is defined as $Y_i = min(C_i, T_i)$. The matrix representation of X_i is very crucial for estimation of regression coefficients. There are several features that may occur about the representation of X_i like margin-specific regression coefficient, identical regression coefficient or mixture of margin-specific regression and identical regression. The presence of correlations between measurements is a natural problem for estimation procedures. In this work the ε_i are assumed to have I.I.D and correlation structure without spec-

ification and correlation with "Exchangeable" are considered to perform this analysis. The general extension of Generalized Estimating Equation (GEE) is applied in this work to carry the algorithm. (Chiou 2014a, Chiou 2014b).

5.3. Bayesian Modeling

The AFT model is defined as

$$log(t_{ij}) = \alpha + x_{ij}\beta + \Omega_{ij} + \sigma\varepsilon_{ij} \tag{9}$$

The term $\alpha + x_{ij}\beta$ is the linear predictor of a subset of intercept and linear dependence of regression predictor and parameters. The term Ω_{ij} effects model and ε_i is the error term. Now, we can write

$$f(t_0/\lambda_{it}) = \frac{1}{\sigma t_0} f_0(\frac{\log(t_0) - \lambda_0}{\sigma}) \tag{10}$$

$$S(t_0|\lambda_0) = S_0(\frac{log(t_{ij}) - \lambda_{ij}}{\sigma}) \tag{11}$$

$$h(t_0|\lambda_{ij}) = \frac{1}{\sigma t_{ij}} h_0(\frac{log(t_{ij}) - \lambda_{ij}}{\sigma}) \tag{12}$$

The terms $f_0(.)$ and $s_0(.)$ are the base failure distribution and corresponding survival function. In case of logistic model, it can be defined with

$$V_{ij} = \alpha + x_{ij}\beta + \Omega_{ij} \tag{13}$$

$$S_0(\varepsilon) = \frac{1}{1 + exp(\varepsilon)} \tag{14}$$

and

$$S(t_0/V_{ij}) = S_0(\frac{log(t_{ij}) - V_{ij}}{\sigma}) \tag{15}$$

$$S(t_0/V_{ij}) = \frac{1}{1 + [t_{ij} exp(-V_{ij})]^{1/\sigma}} \tag{16}$$

$$f(t_{ij}|V_{ij}) = S_0(\frac{log(t_{ij}) - V_{ij}}{\sigma})^2 exp(\frac{log(t_{ij}) - V_{ij}}{\sigma}) \tag{17}$$

It is assumed that $S_0(.)$ is following logistic distribution. Further, the Likelihood is defined as

$$L_2 = \prod_{i=1}^{n} \prod_{j=1}^{n_i} [\frac{1}{\sigma t_{ij}} f_0(\frac{log(t_{ij}) - V_{ij}}{\sigma})]^{V_i} S_0(\frac{log(t_{ij}) - V_{ij}}{\sigma})^{1-c_{ij}} \tag{18}$$

Let $p(a)$ gives the prior distribution for the parameter a and $p(b)$ for the parameter b. The posterior distribution can be stated as

$$p(b,V,a/t) \propto L(t|b,V)p(W/a)p(b)p(a) \tag{19}$$

In Model1, it is assumed that $t_i \sim weibull(\rho,\lambda_i)$. The rate is defined as

$$\log(\lambda_i) = \beta_0 + \mu_{j(i\in j)}, (j = 1,2) \tag{20}$$

The terms β_0 , μ_j and $\log(\rho)$ are assumed to have prior distributions with $N(0,\sigma_\beta^2)$,$N(0,\sigma_\mu^2)$ and $N(0,\sigma_\rho^2)$ respectively.

The Model2, is defined as

$$\log(\lambda_i) = \beta_0 + \mu_{j(i\in j)} + v_i, (j = 1,2) \tag{21}$$

and v_i is enclosed to take care about random effect. The prior distribution of v_i is obtained through $v_i \sim N(0,\sigma_v^2), \tau_v = \sigma_v^2, \sigma_v \sim U(0,3)$.

In Model3, the shape parameter is attached for individual specific observation through $t_i \sim Weibull(\rho_i,\lambda_i)$. The prior distribution of ρ_i is assumed through $log(\rho_i) \sim N(0,\sigma_\rho^2)$ The individual specific random effect has not been considered in this model.

The model performance is observed through p_D,$\bar{D}$ and DIC respectively. The model with the smallest DIC value is considered as suitably fitted in this scenario. The posterior mean estimates observed from three models are given in Table 2 and the corresponding p_D,$\bar{D}$ and DIC values of the model parameters are detailed in Table 4.

6. Data Analysis and Results

Based on the two retrospective data and related studies the predicted PFS for Arm B and Arm A were assumed with 5 months and 2.5 months respectively. Power of 80% was adopted and a total of 33% increment of PFS by the metronomic arm over cisplatin was expected. A type one error of 0.05 was taken; with a 2 tailed p value of 0.05% considered as significant. A total sample size was calculated as 110. The intention to treat the patients was adopted to conduct the primary endpoint analysis. The PFS between both the arms were compared by Kaplan-Meier curve, with the unstratified log-rank test. The multivariate Cox proportional hazard was carried on PFS and OS. The covariates were selected through forward LR method. The hazard

Table 1: Demographic, baseline characteristics and important prognostic details according to the arm

Parameters	Cisplatin=53	Metronomic=57
Median	45(29-70)	48(31-69)
M:F ratio	44.9	49.8
Median monthly Income (USD)	30.7 (7.7-461.5)	38.4 (7.7-461.5)
Localization		
Local	11 (20.7%)	13(22.8%)
Regional	22(41.5%)	21(36.8%)
National	20(37.8%)	23(40.4%)
Tobacco chewer	39(73.6%)	36 (63.2%)
Cigarette smoker	17 (32.1%)	16 (28.1%)
Median pack years	20 (1-30)	20 (2-80)
Subsite of tumour		
Oral cavity	41(77.4%)	43 (75.4%)
Oropharynx	8(15.1%)	10 (17.5%)
Larynx	3(05.7%)	1 (01.8%)
Hypopharynx	1(01.8%)	3 (05.3%)
Locally advanced disease/relapse	50(94.3%)	54 (94.7%)
Metastatic disease	03(05.7%)	03 (05.3%)

ratio with 95% confidence interval was documented.

The median duration of PFS was higher in Arm B (i.e. median 101 days, 95% CI: 58.2-143.7 days) in comparison with Arm A(i.e. median 66 days, 95% CI; 55.8-76.1 days). The log-rank test shows there is a significant difference between both the arms ($p = 0.014$). The factors [age(0.07) and arm(0.015)] influencing the better PFS were obtained through Cox PH model. The hazard ratio in Arm A was 1.58 (95% CI,1.09,2.38). The median OS was significantly higher in Arm B (i.e. median 249 days , 95% CI: 222.48-275.52 days)) in comparison to Arm A(i.e. median OS 152 days (134.19-247.81 days). The log-rank test also confirms a significant difference between both the arms ($p = 0.02$). The only factor arm was found influencing for better OS and it was obtained through Cox PH model. The hazard ratio not in favor of Arm A was 1.63 (95% CI, 1.05, 2.50). The goal of this work is to explore whether Metronomic therapy provides more survival duration of cancer patients. Table 3 gives the estimates of λ and γ for PFS and OS separately. The Hazard Ratio(HR) and Event Time Ratio(ETR) estimates

Table 2: Posterior Mean estimates of parameters based on 20,000 MCMC runs

Parameters	Model1	
	Mean(SD)	(2.5%, 97.5%)
σ_0	1.63(0.32)	(0.78,1.98)
σ_1	1.02(0.60)	(0.07,1.95)
μ_1	0.78 (1.01)	(0.06,0.56)
μ_2	1.05 (1.03)	(0.06,0.86)
β_0	3.32(1.04)	(0.61,4.64)

Parameters	Model2	
	Mean(SD)	(2.5%, 97.5%)
σ_0	2.27(1.25)	(0.49,4.74)
σ_1	3.22(0.96)	(1.55,4.92)
μ_1	3.94 (0.09)	(0.01,3.92)
μ_2	0.53 (0.03)	(0.00,0.55)
β_0	0.94(0.05)	0.80,0.996)

Parameters	Model3	
	Mean(SD)	(2.5%, 97.5%)
σ_0	1.67(0.28)	(0.97,1.98)
σ_1	0.91(0.58)	(0.06,1.95)
μ_1	0.55 (0.91)	(0.05,0.29)
μ_2	0.82 (0.92)	(0.05,0.56)
β_0	3.56(0.91)	3.83,4.68)

Table 3: Posterior Mean estimates of parameters based on 20,000 MCMC runs

Parameters	PFS	OS
	Estimate(SE)	Estimate(SE)
λ	0.00(0.00)	0.00(0.00)
γ	1.16(0.09)	0.00(0.11)
Arm	0.63(0.21)	0.41(0.22)
Age	-0.02(0.01)	-0.01(0.01)

Table 4: Bayesian Model goodness of fit with Model 1, Model2 and Model3

Model	$\bar{D}$	*DIC*	p_D
1	2183	2179	2.98
2	2185	2183	20.16
3	2178	2186	3.96

Table 5: HR estimates obtained through Cox PH and AFT Model

Parameters	Cox PH	
	PFS	OS
Age	0.61(0.35, 1.11)	0.63(.335,1.105)
Arm	1.58(1.09,2.38)	1.63(1.05,2.50)
Parameters	AFT Model	
	PFS	OS
Age	0.97(0.95,0.99)	0.48(0.96,1.00)
Arm	1.88(1.22,2.83)	1.51(0.98,2.34

Table 6: Estimates obtained through regression models

Parameters	PFS		OS	
	Estimate(SE)	p-value	Estimate(SE)	p-value
Intercept	4.23(0.47)	0.00	5.18(0.43)	0.00
Arm	-0.54 (0.19)	0.00	-0.30(0.16)	0.00
Age	-0.02(0.00)	0.00	0.01(0.00)	0.00

Table 7: Estimates obtained through aftgee regression modeling

Parameters	Model1		Model2	
	Estimate(SE)	p-value	Estimate(SE)	p-value
Intercept	3.84(0.57)	0.00	3.84(0.48)	0.00
Age	0.01(0.01)	0.14	0.01(0.01)	0.09
Sex	-0.05(0.26)	0.82	-0.05(0.28)	0.83

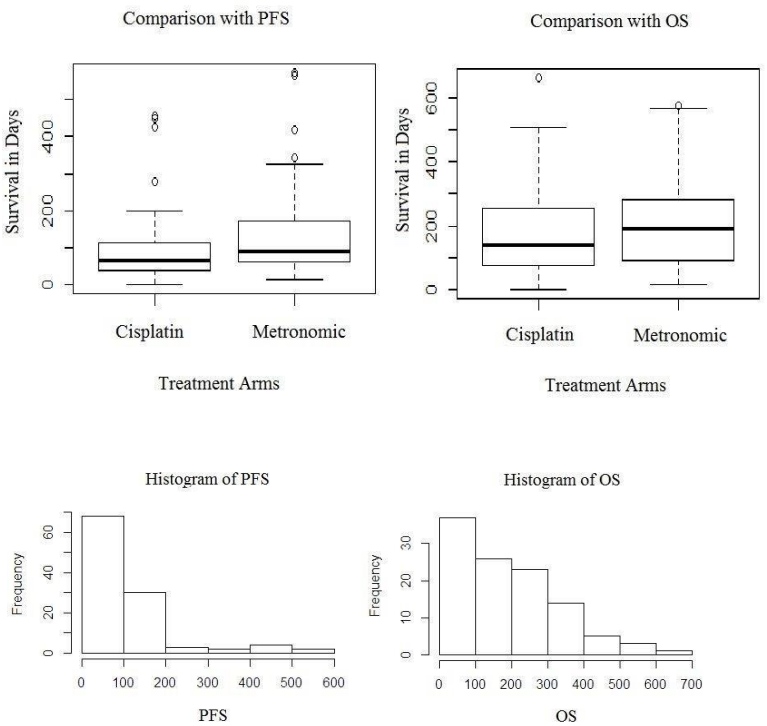


Figure 1: Comparison of Treatment Groups in Terms of Survival Duration

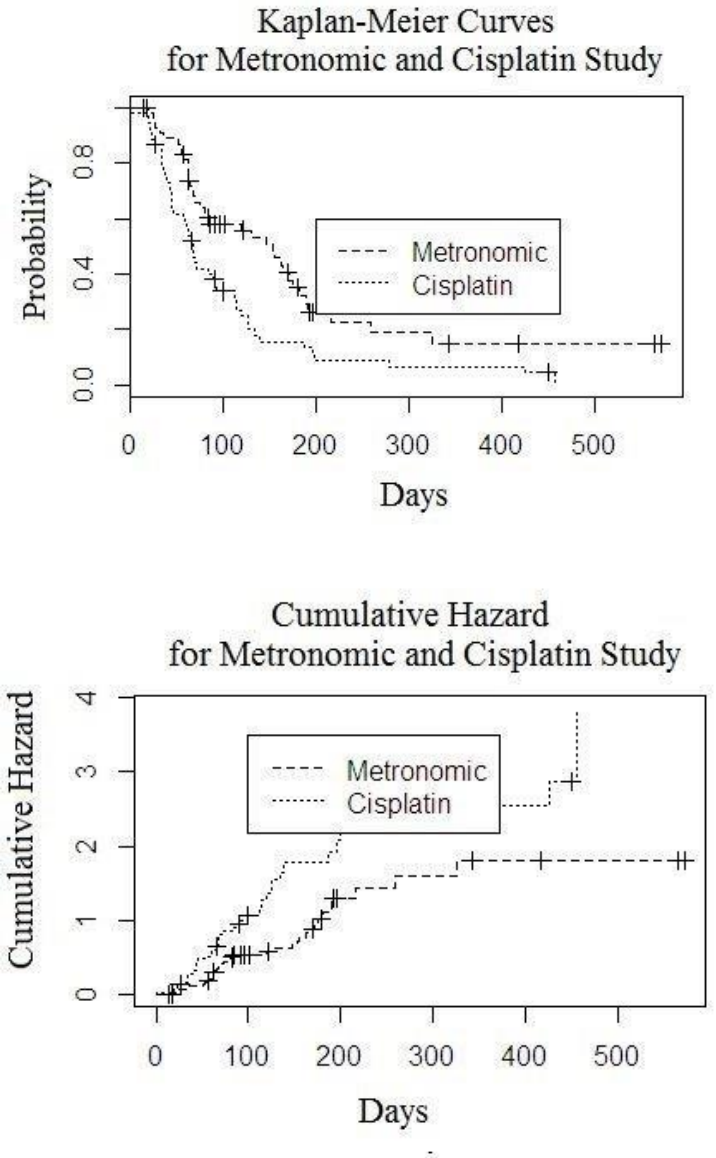


Figure 2: Treatment effect comparison

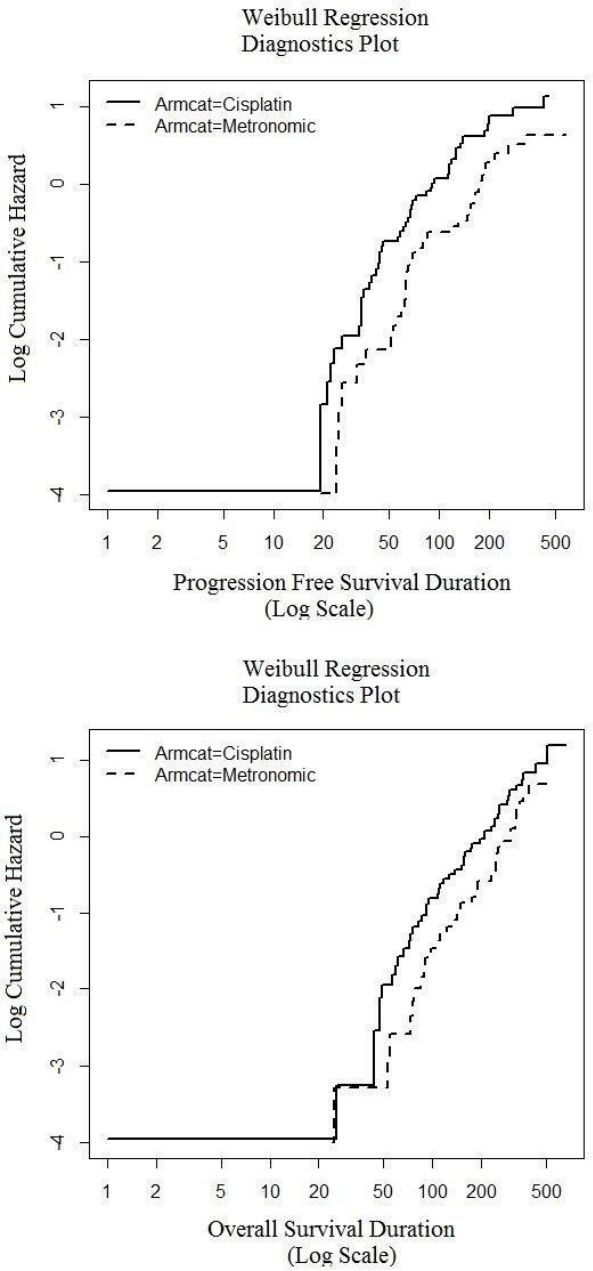


Figure 3: Diagnostic Plot for PFS and OS

obtained through regression models are detailed in Table 5. The mean and median survival duration in days is compared. In both the tables, duration of survival in Metronomic therapy is found better in comparison with Cisplatin. Further, the work contributes to the exploration of the influence of factors for the different survival between the groups. The work is carried out through the model to try to explain the effect of therapy on the time of survival. The coefficient of Arm is found to be significantly related to both models. It provides that patients treated with Cisplatin are likely to have more hazard in comparison with Metronomic arm(Table 5). The estimates of the regression coefficient of Arm and Age are observed with -.54(0.19) and -.02(0.00) respectively (Table 5). The coefficient of Sex is found to be non-significantly related to both models. It proposes that male and female patients are likely to have equal recurrence durations (Table 7). The data analysis is carried out through Bayesian approach. The prior distributional assumption is points for each parameter of interest. We desired to dominate the data information for posterior distribution value formulation. The non-informative prior is used to dominate the data value for posterior distribution for all parameters. The posterior sampling-based procedures Markov Chain Monte Carlo (MCMC) is applied for model generation. Model 1 is observed with minimum DIC value, i.e. 2179. It can be concluded that Model1 is more appropriate in this study data.

7. Discussion and Conclusion

This study provides that PFS is better in a metronomic arm in comparison to cisplatin arm. The result of cisplatin is similar with recently published studies with single-agent platinum (morton 1985,wittes 1976,clavel 1994). The CPH models are widely explored for censored regression modeling for covariates of interest. Accelerated failure time (AFT) models are another and rarely explored approaches for regression modeling for censored data model. The CPH modeling (Cox 1972) and AFT modeling (Kalbeisch 2002) are available methods for right-censored observations for survival data analysis. The frailty effects are widely applied into PH model through consideration of parametric model or with arbitrary distribution (Klein 1999,Horowitz 1999,Anderson 1995,Walker 1997,Sargent 1998,Pickles 1995).Recently, the parametric and non-parametric frailty distribution is applied through AFT (Sargent 1998). The gamma frailty has also been explored into AFT models (Pan 2001). The PH model is widely applicable tool for survival analysis. But exposure to address different types of distribution is limited. The widely applied

distributional assumption, i.e. Weilbull and Gompertz are only suited for monotonically increase or decrease distribution. It is very difficult to avail different types of distributional assumption through PH model. In contrast, AFT is open to carry different distributional assumption into the model assumptions (David 2003). Recently, Bayesian approach for AFT models has been explored. A Bayesian Semi-parametric approach is elaborated to an AFT model (Walker 1999). It also applied for an AFT model with interval-censored and structured correlated data (Komarek 2007). The random effect into AFT model is applied for multivariate doubly censored data for cluster observations (Komarek 2008). Dirichlet process prior is used for mixing distribution to deal with Semi parametric regression model for censored data (Ghosh 2006). A fully Bayesian approach for the median regression by Polya tree prior is proposed (Walker 1999). An application of normal mixture prior distribution has also been illustrated for AFT model (Komarek 2007,Komarek 2005). The AFT can be directly linked with expected death time to the covariates of interest through linear regression modeling. The semi parametric extension of AFT is more appropriate for undetermined error distributions. The rank-based approach (Prentice 1978) and least squares (Buckley 1979, Jin 2007) are two methods to handle the semi parametric extension of AFT. The specified distribution for error is called as parametric AFT model(Tsiatis 1990,Therneau 2014). Failure to specify the proper distribution for the error may generate the bias estimation about censored data. The Buckley-James (BJ) estimator (Buckley 1979) is the tool to work with parametric AFT.

In this paper, we examine the chemotherapeutic data by the Bayesian parametric AFT model. The OS and PFS were observed separately for better information about survival duration on chemotherapeutic effect. The OpenBUGS is utilized for data analysis through parametric AFT models. The DIC criterion is used to know the best fitted models. The K-M curve is plotted for non-parametric statistical inference. It is found that the Bayesian AFT model with random effect is suitable for analysis of chemotherapy data. The results observed in this study shows that metronomic arm is having significant influence on duration of survival on chemotherapeutic patients. The model can be more robust by allowing other parametric assumption. However, it can be difficult to carry the computational work (Anderson 1995,Sargent 1998,Folkman 1971). The DIC is applied as diagnostic criteria to test the model. The DIC is adopted to check the model fitting suitability. It is avaliable in OpenBugs. However, other model comparision tools can be

used to compare the models in different computational platforms. The parametric AFT model is applied to chemotherapeutic data. The computational difficulties may be greater if we shift from parametric approach (Anderson 1995,Sargent 1998,Folkman 1971). Although, the semiparametric approach can also be applied and it has been found successful in survival data analysis. However, semi-parametric approach can also be applied to AFT model as an extension of this work (Anderson 1995, Sargent 1998,Komarek 2007,Komarek 2005,Christensen 1988).

8. Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

9. Acknowledgements

The authors thank the two anonymous referees for their cautious reading and constructive suggestions which have led to improvement on earlier versions of the manuscript.

REFERENCES

ANDERSON, J. E., LOUIS, T. A., (1995). Survival analysis using a scale change random effects model. Journal of the American Statistical Association, 90, 669–679.

AGGARWAL, S. K. (1998). Calcium modulation of toxicities due to Cisplatin. Met Based Drugs, 5, 77–81.

BUCKLEY ,J., JAMES, I., (1979). Linear Regression with Censored Data. Biometrika, 66 (3), 429–436.

CHIOU, S. H., KANG, S., KIM, J., YAN, J., (2014a). Marginal Semiparametric Multivariate Accelerated Failure Time Model with Generalized Estimating Equations. Lifetime Data Analysis, 24 (4), 599–618.

CHIOU, S. H., KANG, S., KIM, J., YAN, J., (2014b). Fast Accelerated Failure Time Modeling for Case-Cohort Data. Statistics and Computing, 24 (4), 559–568.

CELMINS, A., (1987). Least squares model fitting to fuzzy vector data. Fuzzy sets and systems, 22 (3), 245–269.

CHRISTENSEN, C., JOHNSON, J. W., (1988). Modelling accelerated failure time with a Dirichlet process. Biometrika, 75, 693–704.

CLAVEL, M., VERMORKEN, J. B., COGNETTI, F., et al., (1994). Randomized comparison of cisplatin, methotrexate, bleomycin and vincristine (CABO) versus cisplatin and 5-fluorouracil (CF) versus cisplatin (C) in recurrent or metastatic squamous cell carcinoma of the head and neck A phase III study of the EORTC Head and Neck Cancer Cooperative Group. Annals of Oncology, 5 (6), 521–526.

COX, D. R., (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society, 34 (2), 187–220.

DAVID, C., (2003). Modelling Survival Data in Medical Research. Chapman Hall/CRC Texts in Statistical Science.

FOLKMAN, J., (1971). Tumor angiogenesis: therapeutic implications N Engl J Med, 285, 1182–1186.

GHOSH, S. K., GHOSAL, S., (2006). Semiparametric Accelerated Failure Time Models for Censored Data. Bayesian Statistics and Its Applications. Anamaya Publishers.

HANSON, T., JOHNSON, W. O., (2004). A Bayesian semiparametric aft model for interval-censored data. J. Comput. Graph. Statist, 13, 341–361.

HANAHAN, D., FOLKMAN, J., (1996). Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. Cell, 86, 353–364.

HANAHAN, D., BERGERS, G., BERGSLAND, E., (2000). Less is more, regularly: metronomic dosing of cytotoxic drugs can target tumor angioňgenesis in mice., J Clin Invest, 105, 1045–1047.

HOROWITZ, J. L., (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. Econometrica, 67, 1001–1028.

HOUGAARD, P., MYGLEGAARD, P., JOHNSEN, K. B., (1994). Heterogeneity models of disease susceptibility with application to diabetic nephropathy. Biometrics, 50, 1178–1188.

JIN, Z., HUANG, L., (2007). lss: An S-PLUS/R Program for the Accelerated Failure Time Model to Right Censored Data Based on Least-Squares Principle. Computer Methods and Programs in Biomedicine, 86 (1), 45–50.

KALBEISCH, J. D., PRENTICE, R. L., (2002). The Statistical Analysis of Failure Time Data. John Wiley Sons.

KOMAREK, A., E. LESAFFRE, E., (2008). Bayesian Accelerated Failure Time Model With Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions. Journal of the American Statistical Association, 103 (482), 523–533.

KOMAREK, A., LESAFFRE, E., (2007). Bayesian Accelerated Failure Time Model for Correlated Interval-Censored Data with a Normal Mixture as Error Distribution. Statistica Sinica, 7, 549–569.

KOMAREK, A., HILTON, J. F., (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. J. Comput. Graph. Statis, 14, 726–745.

KLEMENT, G., HUANG, P., MAYER, B., GREEN, S. K., MAN, S., BOHLEN, P., HICKLIN, D., KERBEL, R. S., (2002). Differences in therapeutic indexes of combination metronomic chemotherapy and an anti-vegfr-2 antibody in multidrug-resistant human breast cancer xenografts. Clin Cancer Res, 8 (1), 221–232.

KLEIN, J. P., PELZ, C., ZHANG, M., (1999). Modeling random effects for censored data by a multivariate normal regression model. Biometrics, 54, 497–506.

MORTON, R. P., RUGMAN, F., DORMAN, E. B., STONEY, P. J., WILSON, J. A., MCCORMICK, M., VEEVERS, A., STELL, P. M., (1985). Cisplatinum and bleomycin for advanced or recurrent squamous cell carcinoma of the head and neck: a randomised factorial phase III controlled trial. Cancer chemotherapy and pharmacology, 15 (3), 283–289.

MUNOZ , R., MAN, S., SHAKED, Y., LEE, C. R., WONG, J., G., FRANCIA AND KERBEL, R. S. (2006). Highly efficacious nontoxic preclinical treatment for advanced metastatic breast cancer using combination oral uft-cyclophosphamide metronomic chemotherapy. Cancer Res, 66, 3386–3391.

NGUYEN, HUNG, T., WU, (2006). Fundamentals of statistics with fuzzy data. Berlin Springer.

POLVERINI, P. J., NOVAK, R. F., (1986). Inhibition of angiogenesis by the antineoplastic agents mitoxantrone and bisantrene. Biochem Biophys Res Comunication, 140, 901–907.

PRENTICS, R. L., (1978). Linear Rank Tests with Right Censored Data. Biometrika, 65 (1), 167–180.

PAN, W., (2001). Using frailties in the accelerated failure time model. Lifetime Data Analysis, 7, 55–64.

PICKLES, A., CROUCHLEY, R., (1995). A comparison of frailty models for multivariate survival data. Statistics in Medicine, 14, 1447–1461.

SARGENT, D. J., (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. Biometrics, 54, 1486–1497.

WALKER, S., MALLICK, B. K., (1999). A Bayesian Semiparametric Accelerated Failure Time Model. Biometrics, 55 (2), 477–483.

WALKER, S. G., MALLICK, B. K., (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. Journal of the Royal Statistical, Society B , 59, 845–860.

WALKER, S., MALLICK, B. K., (1999). A Bayesian Semiparametric Accelerated Failure Time Model. Biometrics, 55 (2), 447–483.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., LINDEVAN, D. A., (2002). Bayesian measures of model complexity and fit. J. Roy. Statist. Soc. Ser. B., 64, 583–616.

SURENDIRAN, A., BALAMURUGAN, N., GUNASEELAN, K., AKHTAR, S., REDDY, K. S., ADITHAN, C., (2010). Adverse drug reaction profile of cisplatin-based chemotherapy regimen in a tertiary care hospital in India: An evaluative study. Indian J Pharmacol, 42 (1), 40–43.

THERNEAU, T., (2014). "survival: A Package for Survival Analysis in S. R package version". R package version 2, 37–7.

TSIATIS, A. A., (1990). Estimating Regression Parameters Using Linear Rank Tests for Censored Data. The Annals of Statistics, 18 (1), 354–372.

WITTES, R. E., CVITKOVIC, E., SHAH, J., GEROLD, F. P., STRONG, E. W., (1976). CIS-Dichlorodiammineplatinum (II) in the treatment of epidermoid carcinoma of the head and neck. Cancer treatment reports, 61 (3), 359–366.

YABUUCHI, Y., WATADA, JUNZO, NAKAMORI, Y., (1997). Fuzzy principal component analysis for fuzzy data. Fuzzy Systems, 1997., Proceedings of the Sixth IEEE International Conference on, 2, 1127–1132.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 691–722

CALENDAR AND SEASONAL EFFECTS ON THE SIZE OF WITHDRAWALS FROM ATMS MANAGED BY EURONET

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

ABSTRACT

This study analyses the calendar effects on withdrawals from Automated Teller Machines (ATMs) (daily data) managed by the Euronet network for the period from January 2008 to March 2012. Our study focuses on the identification of specific calendar and seasonal effects in the ATM cash withdrawal series of the company in the Polish provinces of Małopolska and Podkarpackie. The results of the analysis show that withdrawals depend strongly on the day of the week. On Fridays more cash is withdrawn than on other days, and Saturdays and Sundays are the days of the week with the lowest level of withdrawals. In a month, it can be seen that cash withdrawals take place more often in the second and in the last weeks of the month. This observation suggests that withdrawals from ATMs can be related to the profile of wage withdrawals. In Poland, in the public sector wages are paid at the beginning of the month, and in the private sector at the end of the month. The time series of withdrawals also reflect seasonality. The largest amounts are withdrawn in July, August and December. Reason for the increased demand for cash are the summer holidays and the Christmas season. The results reflect consumer habits which show pronounced calendar and seasonal effects.

Key words: calendar effects, seasonal effects, replenishment management.

1. Introduction

The most important virtual banking services are phone banking, mobile banking, Internet banking and ATM banking. Banks collect large amounts of data on their customers. They try to use these data in order to make better decisions. Trends and seasonal patterns of behaviour that are reflected in the data may be used to improve a bank's customer value management strategies. This is important with respect to risk management, profit maximization, fraud detection

[1] AGH University of Science and Technology, Faculty of Management, POLAND, E-mail: henryk.gurgul@gmail.com.
[2] AGH University of Science and Technology, Faculty of Management, POLAND, E-mail: msuder@agh.edu.pl

and marketing (Bretnall et al. 2008). On the basis of trends and seasonal patterns econometric models may be suitable for short-term forecasting. These forecasts may help to adjust the replenishment to the need of customers and at the same time to avoid a large, unnecessary (and unproductive) cash flow waiting in ATMs for their customers instead of being profitably invested.

In the last two decades of the 20th century *Automated Teller Machines*, which formed ATM networks, were an important addition to the banking system. The first ATM was installed in New York in Chemical Bank at the end of the seventies. This dynamic rise in the number of ATMs in recent years has reached saturation level (this is the upper bound of the number of ATMs necessary to fulfil the needs of the inhabitants of a country) in developed countries. As per the data for 2009, the relative availability of ATMs is found to be highest in Canada (1,800), followed by the USA (1,382) and Australia (1,230). In some countries the number of ATMs installed has shown a tendency to fall (Gerdes et al., 2005, Schmitz and Wood, 2006). The reason for this observation is the policy of banks. From the point of view of bank managers the distribution of cash by ATMs generates significant costs.

In addition, using ATMs reduces deposits and, therefore, the profits of banks have a tendency to decrease. However, the elimination of cash flow through ATMs would not be desirable for banks because many bank clients prefer cash services. Therefore, the elimination of ATMs could mean the loss of many clients (Takala and Viren, 2007). Many banks regard ATMs as important tools that can reduce banking transaction costs, excessive personal costs, branching costs, etc.

This is true especially in the case of villages and small towns in developed countries and in developing countries, where cash is still the main medium of transactions.

As Kumar and Kumar (2014) stress, the use of Automated Teller Machines (ATM) in India is so popular that many people understand ATM as Any Time Money. In metropolitan cities in India, if you walk half a kilometre in any direction, then you will get at least one ATM. On the basis of the information we can imagine the importance of ATM in the day-to-day life of many people across the world.

One of the benefits of ATMs in recent years has been that a person holding a card of any bank can withdraw money from an ATM of any other bank, although the number of transactions without charge is limited in different countries (in India to 5 per month).

The first ATM in Poland was installed in the late eighties in the division III of PeKaO in Warsaw. However, the ATM started to play a significant role in Poland at the end of the nineties.

The increase in the number of ATMs implies a rise in operations, the share of ATMs in cash operations and in the turnover in ATMs. The number of ATM operators has also increased. Empirical studies have proved a significant statistical dependence between the number of ATMs and the number of ATM networks operating in a city, region or country.

The Euronet network was created in 1994 in Budapest. In the same year it installed the first ATM in the Hotel Marriott in Warsaw. Meanwhile, the number of ATMs operated by Euronet expanded rapidly. Currently, the total number of ATMs in the Euronet network amounts to approximately 4.5 thousand. Its share in the total number of ATMs installed in Poland is approximately 25%. The total number of withdrawals from Euronet ATMs in the first quarter of 2012 was approximately 2.5 billion PLN.

The management of this large network has become an important issue. The most important keywords of the management process are ATM localization, the size of replenishment, the time of replenishment and the convoy logistics. These factors incur most of the costs of operating the ATMs which are covered by Euronet. When it comes to reducing the cost of operating ATMs the forecasting of cash withdrawals from ATMs may have an essential role. Correct forecasts can contribute to a reduction in ATM management costs; there should not be too much inert money, which does not generate any profit in ATMs. However, an ATM may not be empty, which is very important from the point of view of customer satisfaction. This topic also seems to be important from other points of view. The revenue of the network operator is proportional to the number of transactions. The size of withdrawals has no impact on revenue. The number of withdrawals is a stable variable, so that the number of withdrawals may be one of the exogenous variables which explain the size of withdrawals. Moreover, the size of withdrawals can be forecasted by the number of withdrawals - we can use the average withdrawal as a proxy.

In addition, forecasting the number of withdrawals is important with respect to the throughput of an ATM. The lack of queues is a very important issue from the point of view of customer satisfaction.

The prediction of cash demand from a given ATM makes it possible to determine the risk level and replenishment strategy. Forecasts must take into account not only the number and nature of people in the neighbourhood of an ATM but also the behaviour and habits of potential customers of the ATM. Since withdrawals have the structure of time series in modelling, there should be taken into account not only systematic factors but also stochastic determinants. The time series of withdrawals show “calendar effects” in particular. This means that the size of withdrawals depends usually on the season, month of the year, day of the week, time of the day, etc. The identification of calendar effects is of great importance with respect to the choice of strategy and principles for the replenishment of ATMs.

This paper mainly deals with calendar and seasonal effects, therefore the second section of this paper provides an overview of the literature concerning the calendar effect and questions which arise from ATM management. Section 3 describes the dataset and gives the most important descriptive statistics with respect to different localizations of ATMs. In the following sections the results are presented and calendar effects are analyzed. A summary of results is given in the concluding section.

2. Literature overview

The research on statistical properties of time series of the number of withdrawals from ATMs is important with respect to building relevant forecasting models for the ex ante prediction of withdrawals. The owners of ATM networks can use the forecasts in two ways. First of all. the network operator supplies services to debit card issuers. This is a source of income (charges – the so-called interchange) for every transaction conducted.

Proper forecasts make it possible to determine the total charges for using ATMs. In addition, the forecasting of withdrawals is an important factor for managing replenishment in ATMs. Proper withdrawal forecasts can reduce the costs of managing ATMs.

ATM cards became a common part of life in the last 30 years. However, the statistical properties of withdrawals from ATMs have not yet been investigated thoroughly, with respect to several aspects, from both a theoretical and practical point of view.

A better knowledge of consumer habits is very important with respect to the replenishment strategy of ATMs (Galbraith and Tkacz, 2007). An understanding of these habits may be helpful in the prediction of individual consumption (Esteves, 2009, and Duarte et al., 2016) or in forecasting the retail sales statistics (Carlsen and Storgaard, 2010).

Holden and El-Bannany (2004) demonstrate that Automated Teller Machines (ATMs) play an important role in increasing Return On Asset (ROA) in their analysis of banks in the United Kingdom. Kondo (2010) investigates whether this conclusion also applies to Japanese banks. He finds that ATMs do not have any influence on ROA of Japanese banks. However, he documented that ATMs had positive effects on fees and commissions (income) from 2000 to 2003, and positive effects of ATMs on interest income have also been seen recently. Kondo (2010) concludes that in Japan, ATMs do not influence ROA, which includes the overall profits of bank transactions. However, they contribute to particular businesses in that they can make the most of their abilities.

A study by Snellman and Viren (2009) reported results on the dependence between the structure of banking systems, the number of ATM networks and the location of ATMs.

Calendar effects are frequently observed in economic time series, including financial time series. This follows from the observation that most economic time series depend directly or indirectly on hours, days, months, quartals and other time intervals. The phenomenon of the time dependence of the investor, customer, and consumer activity on time were discovered in the seventies by Cleveland and Devlin (1980) and Liu (1980). They stressed the importance of the number of working days in the week and their link with seasonal effects. These data determine the dynamics of the time series being studied to a large extent. The importance of seasonal effects is reflected in updating these effects in statistics

(national accounts). National accounts update not only seasonal effects but also effects of working days in the period being studied.

Beside the number of working days, the day of the week, the week of the month, the month of the year and other calendar effects such as national or church holidays also have an impact on the dynamics of respective time series. Findley and Monsell (2009) stressed that there is a necessity to take into account the days of the week. Monthly activity can depend on the number of days in a particular month and on what days of the week are being studied.

A major effect can be caused by the Easter season. This season is not at invariant time. Every year the Easter season may vary, usually beginning in March, but sometimes in April. Thus, it is difficult to approach the Easter season by quarters. Sometimes it is in the first quarter, sometimes in the second. The Easter season may have an impact on quarter indices a year apart, e.g. on the ratio of the values of time series in the first quarter of the following year and in the first quarter of the previous year.

The time series of ATM withdrawals reflect such calendar effects as days, when money is due, holidays, seasonality of demand, cycles and trends. These calendar effects are analyzed in papers, e.g. Simutis et al., 2008, in the framework of the logistics of ATM services.

In practice calendar effects often coincide with seasonal effects. Calendar effects can be detected after different volatility sources have been analysed. The most important procedure is filtering, which removes seasonal effects. The methods for extracting seasonal variations are not proper tools for extracting calendar effects, which are not periodic (Cleveland and Grupe, 1983). As demonstrated by Findley et al. (1998), properly accounting for calendar effects broken down by day is of great importance. The reason is that in this way the statistical properties of the time series of withdrawals become better and one can apply more adequate models.

In the scientific literature there are several methods for the detection, estimation and correction of time series which show calendar effects. On the basis of a large sample, Cleveland and Devlin (1980) established the empirical distribution for the monthly time series of withdrawals. They detected the main frequencies for these time series. Findley and Soukup (1999, 2000, 2001) examined the usefulness of empirical distribution for the detection of the day-effect. McElroy and Holland (2005) used a nonparametric test in order to assess maxima in samples.

In order to assess the importance of the day of the week, dummy days of the week are usually used. However, these dummies are strongly correlated from one side. They also show seasonal effects, so in order to establish stable estimations of the day-effect, proper computer programmes are necessary, as demonstrated by Young (1965), Bell and Hillmer (1983), Cleveland and Grupe (1983). The best known of them, such as X-12-ARIMA and Tramo-Seats, allow the user to combine calendar effects, national holidays and the calendar effects of other holidays. This problem is complex because of movable feasts (holidays like

Easter, Ramadan or Chinese New Year). Findley and Soukup (2000) assumed that the cost of ignoring calendar effects may be considerable. Hansen et al. (2005) take for granted that all the potential calendar effects do not follow from an economic theory. Therefore, it is necessary to take into account all the potential calendar effects.

Kufel (2010) cites two methods for the detection of cyclicity. The first one uses a panel of dummies and the second one harmonic components. He detected cyclicality within the periods: year, month, week and day, in the time series of withdrawals from one ATM in Toruń.

Calendar effects are most important in the case of daily withdrawals from ATMs because this topic is not sufficiently handled in the scientific literature (Rodrigues and Esteves (2010)).

Rodrigues and Esteves analyzed the following calendar effects in their paper on ATMs installed in Portugal:

- *day-of-the-week effect*: in this case it is assumed that withdrawals are dependent on the day of the week.
- *week-of-the-month* effect: in Portugal wages and salaries are paid monthly, thus the intra month effects of withdrawals cannot be ignored.
- *month-of-the-year effect*: the analysis comprises *12 month-of-the year effects*. This analysis can take into account typical seasonality of consumption.
- *Holidays:* Christmas, New Year, two movable feasts (Carnival and Easter) and other holidays with permanent dates. In their analysis, Rodrigues and Esteves (2010) also used the time series of withdrawals directly before and after holidays, based on Sullivan et al. (2001). Preholiday periods comprise days before the closing of banks and post-holiday periods refer to periods just after the banks close.

Empirical research by Rodrigues and Esteves (2010) have proven the presence of significant calendar effects on withdrawals from ATMs installed in Portugal.

To some extent, our research refers to the investigations presented by Rodrigues and Esteves (2010). In the next section we will present the dataset used in the empirical part of our paper.

3. Dataset and its properties

The research on calendar effects and seasonal effects occurring in the time series of the number of withdrawals from ATMs was based on data on withdrawals from ATMs operated by the Euronet company and located in two areas of Lesser Poland (Małopolskie) and Subcarpathia (Podkarpackie). The analysis was based on data on 222 ATMs covering the period January 2010-December 2012.

Information on the location of the ATMs may be found in Table 1.

Table 1. The number of ATMs in different types of locations in the two voivodeships

Province	Małopolskie		Podkarpackie		Total	
Type of localization	N	%	N	%	N	%
Bank Branch	47	27.49	25	49.02	72	32.43
Hipermarket	28	16.37	8	15.69	36	16.22
Shop	25	14.62	6	11.76	31	13.96
Shopping Center	24	14.04	6	11.76	30	13.51
Petrol Station	22	12.87	2	3.92	24	10.81
Transport	4	2.34	0	0.00	4	1.80
Entertainment	1	0.58	0	0.00	1	0.45
Hotel	1	0.58	0	0.00	1	0.45
Other	19	11.11	4	7.84	23	10.36
Total	171	77.03	51	22.97	222	100.00

Due to the large number of time series examined, detailed empirical results will be presented for the sum of the number of withdrawals from all ATMs and for six selected ATMs which operate in different locations.

In Figure 1 we present a graph of the time series of the overall number of withdrawals from all ATMs, while in Table 2 we provide descriptive statistics for this time series.

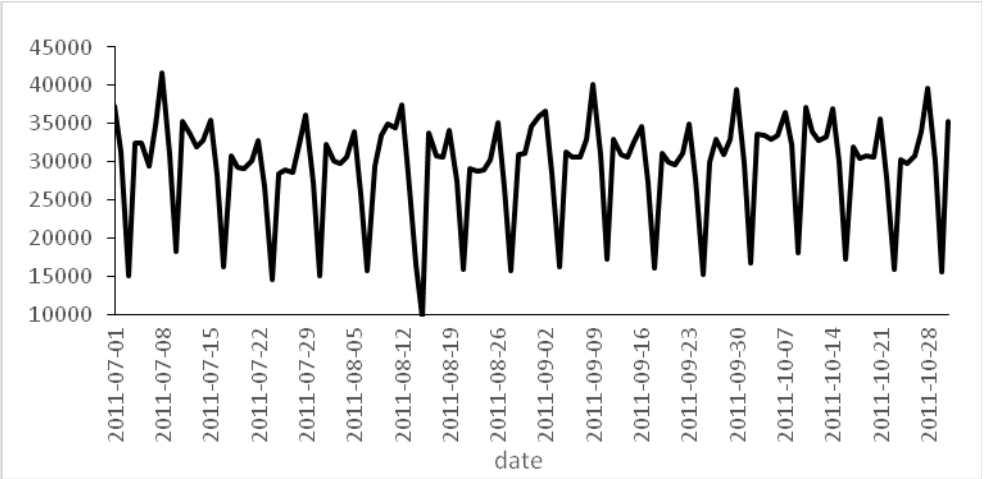


Figure 1. Overall number of daily withdrawals from all the 222 ATMs in the time period July 2011-October 2011

Table 2. Descriptive statistics of the time series of the overall number of daily withdrawals

Average	Median	Standard deviation	Coeff. of variation	Minimum	Maximum	Skewness	Kurtosis
29 139.2	30 810.5	7835.02	26.89%	0	48 178	-1.28	1.47

The analysis of the time series shown in Figure 1 proves that there was not a clear trend in the overall number of withdrawals from ATMs over the period being studied. The preliminary analysis confirms, however, that in this series one can observe a number of seasonal as well as calendar effects (a notable example is the increase in the number of withdrawals in May, June, November and December). In addition, the data presented in Table 2 allows the claim that approximately 30 000 withdrawals a day take place in the selected ATMs. Therefore, there are approximately 130 daily cash withdrawals in each of the machines on average.

More information on the structure of the number of withdrawals from the ATMs follows from the analysis of plots of the time series of the number of withdrawals from the selected ATMs (Figure 2) and their descriptive statistics (Table 3).

Table 3. Descriptive statistics of the time series of withdrawals from selected ATMs.

	ATM 1	ATM 2	ATM 3	ATM 4	ATM 5	ATM 6
Province	Małopolskie	Małopolskie	Podkarpackie	Podkarpackie	Małopolskie	Małopolskie
Town/city	Wieliczka	Zakopane	Stalowa Wola	Rzeszów	Kraków	Kraków
Location type	Shop	Petrol Station	Bank Branch	Shopping Center	Hipermarket	Other (pharmacy)
Average	121.44	157.78	112.69	274.29	200.11	125.74
Std. deviation	47.237	68.05	41.92	94.68	66.49	47.61
Coefficient of variation	38.89%	43.13%	37.20%	34.51%	33.22%	37.87%
Minimum	0	0	0	0	0	0
Maximum	271	462	218	523	412	226
Skewness	-0.54	1.004	-0.695	-0.733	-1.19	-0.547
Kurtosis	0.112	1.30	0.113	0.034	2.279	0.537

It can be seen that there are differences in the structure of the number of withdrawals from particular ATMs. For example, the time series of the number of withdrawals from ATM No. 2 shows trend and seasonality. For the remaining time series, a visual inspection does not establish the presence or absence of seasonal effects or calendar effects. The analysis of the information presented in Table 3 proves that the values of basic statistical measures of the number of withdrawals from ATMs are quite varied. For example, by comparing the average

number of daily withdrawals for the six selected ATMs one may notice that these values are quite different. For example, for ATMs installed in shopping centres there are around 120 withdrawals on average, while more than twice that number took place in a shopping centre in Rzeszów. The number of withdrawals is also affected by such factors as the type of location and availability of ATMs in particular locations. In the case of the remaining variables one can also notice significant differences among the selected ATMs.

Figure 1 for ATM 2 provides evidence about the cycle of one year (Figure 3 encompasses full three years). In the case of other ATMs we do not find a cycle of one year. However, one can detect 1 week seasonality. Thus, the plots of time series are for a time frame of 4 months.

A preliminary analysis proves that the number of withdrawals varies significantly among the ATMs analysed. It seems interesting to test whether such variability is also true in the case of seasonal and calendar effects.

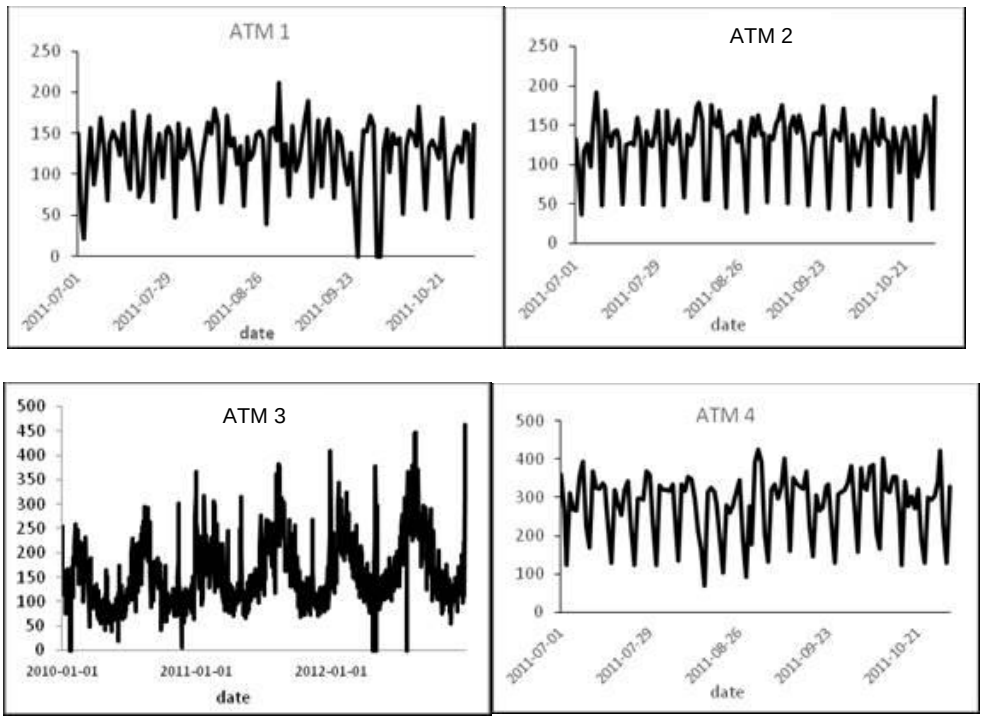


Figure 2. Plots of the time series of the numbers of withdrawals in selected ATMs

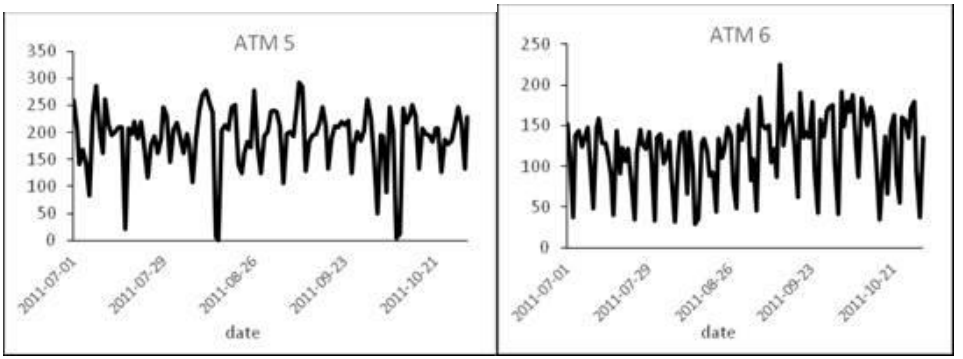


Figure 2. Plots of the time series of the numbers of withdrawals in selected ATMs (cont.)

4. Methodology and empirical results

The deterministic structure of most time series can be described using two basic classes of components: trend and seasonality. In some cases, there may be additional components such as calendar effects. The trend represents a general non-linear (or linear) component, which allows an assessment of the overall direction of the dynamics of the phenomenon examined during the period under consideration (e.g. a linear trend in a stable period, an exponential trend during a development phase). Formally, a seasonal component represents a regularly repeated pattern in the behaviour of the phenomenon analysed.

In this part of the article we present the methods for testing the presence of such components of the time series as seasonality and calendar effects. We also discuss the results of the empirical analysis. We stress that the analysis in our paper is conducted within framework of deterministic seasonality. We do not aim in this study to establish the impact of stochastic factors on seasonality pattern.

4.1. Seasonal effects

In the case of many time series (e.g. electricity or gas consumption) the presence of seasonal effects can be relatively easy to verify solely on the basis of a visual inspection. In the case of the time series of the number of withdrawals from ATMs, the verification of the presence of seasonality solely on the basis of a visual inspection of the respective plots was possible only in the case of some ATMs.

Based on Figure 2, which presents the plots of the numbers of withdrawals, one can determine the presence of seasonality only in the case of the ATM located in Zakopane (ATM, No. 2). In the case of the remaining ATMs, the identification of the presence of this deterministic component turned out to be impossible based on a visual inspection of the respective plots.

In order to identify seasonality in the time series examined in this paper we use a method based on spectral density estimators, also known as periodograms.

For the time series $x_1, x_2, ..., x_n$ the periodogram is defined as:

$$I_n\left(\omega_j\right) = \frac{1}{n}\left|\sum_{k=1}^{n} x_k \exp\left\{-2\pi i\left(k-1\right)\omega_j\right\}\right|^2 \tag{1}$$

where $\omega_j = j/n$ j-th frequency for $j = 1, 2, ..., \left[\frac{n}{2}\right]$, and $\left[x\right]$ stands for the floor of number x.

The results of a spectral analysis of the time series of the overall number of withdrawals are presented in Figure 3.

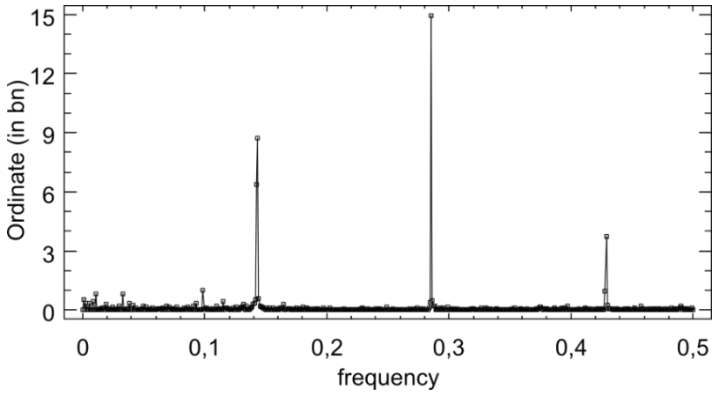


Figure 3. Periodogram of the time series of the overall number of withdrawals

A spectral analysis of the time series of the total number of withdrawals from ATMs confirms the existence of a weekly cycle. The peak which appears for frequencies near to 0.14 indicates the presence of weekly seasonal effects. In the above periodogram one can also see changes in spectral density for frequency 0.285 and 0.428, which are referred to as harmonic fluctuations (Franses, 1996). This indicates the existence of length cycles of 3.5 and 2.33 days respectively, that is cycles whose multiple is the weekly cycle. For the examined time series of the overall number of withdrawals from ATMs, the spectral analysis showed no presence of monthly or annual seasonality.

The results of spectral analysis of the time series of the ATMs selected are presented in Figure 4.

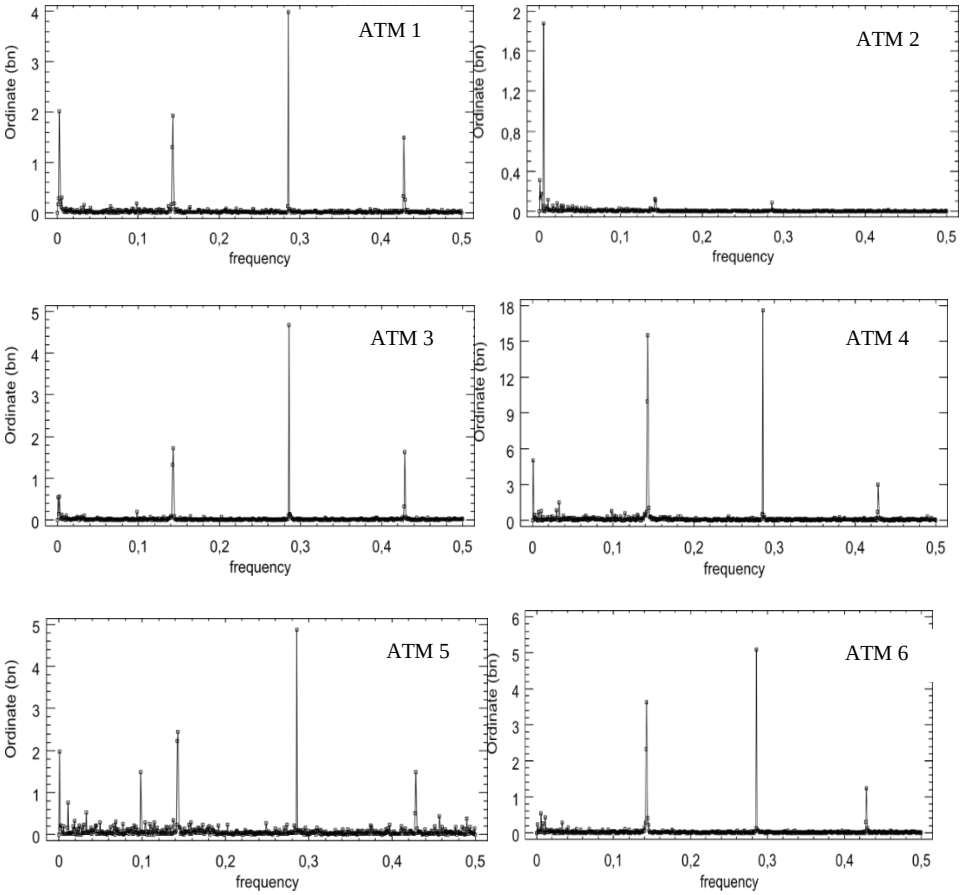


Figure 4. Periodogram of the time series of the number of withdrawals in selected ATMs

A spectral analysis of six selected ATMs shows differences among these machines. A weekly cycle was found for all ATMs, except for ATM No. 2. On the other hand, the noticeable peaks for a frequency close to 0 that occur for ATMs No. 1, 2, 4 and 5 indicate the presence of annual seasonality. A slight increase in the value of the spectral density for frequencies near 0.033, which occurred for ATMs No. 4 and 5, corresponds to a cycle around 30 days in length, i.e. a monthly cycle. One should take this into account and draw attention to the fact that due to the different lengths of individual months of the year, the spectral analysis may not prove the presence of monthly seasonality even when the effect of the day of the month has a significant impact on the dynamics of the series of withdrawals.

The last stage of research aimed at tracing cycles of different lengths was the spectral analysis conducted for the remaining 216 ATMs. Table 4 shows the results of the analysis, which focused on testing whether there are cycles of

different lengths for all ATMs with respect to the type of location (nominal and percentage values).

Table 4. Nominal and percentage results of spectral analysis.

Location type	Weekly cycle		Monthly cycle		Annual cycle	
	N	[%]	N	[%]	N	[%]
Bank Branch	72	100.0	16	22.2	35	48.6
Hipermarket	36	100.0	8	22.2	16	44.4
Shop	30	96.8	6	19.4	12	38.7
Shopping Center	30	100.0	7	23.3	15	50.0
Petrol Station	22	91.7	4	16.7	11	45.8
Transport	4	100.0	2	50.0	2	50.0
Entertainment	1	100.0	0	0.0	1	100.0
Hotel	1	100.0	0	0.0	0	0.0
Other	23	100.0	5	21.7	11	47.8
Total	219	98.6	48	21.6	103	46.4

Thus, the analysis of spectral density confirmed the presence of weekly cycles for almost 99% of the time series analysed. In addition, in the case of almost 50% of the ATMs examined we found annual seasonality, and for slightly more than 20% of the ATMs analysed we confirmed the existence of monthly cycles.

The results obtained indicate that seasonality and the cycles present in the time series of withdrawals are important components, which should be taken into account when selecting the models used for prediction purposes.

4.2. Calendar and seasonal effects

Calendar and seasonal effects are a common phenomenon observed in economic time series, including financial time series. The latter follows from the fact that most of economic time series directly or indirectly refer to particular days, months, quarters and other units of time. One should also pay attention to the fact that the presence of weekly cycles may be considered in the context of seasonal effects. The same applies to the analysis of the properties of monthly withdrawals (annual cycle). However, in the case of analysing the impact of the day of withdrawals (the monthly cycle) one should conduct the analysis in the context of calendar effects. This follows from the fact that not all months are of the same length.

Some indication of the possibility of the existence of calendar effects in the time series examined follows from a visual inspection of the plots of the series being studied.

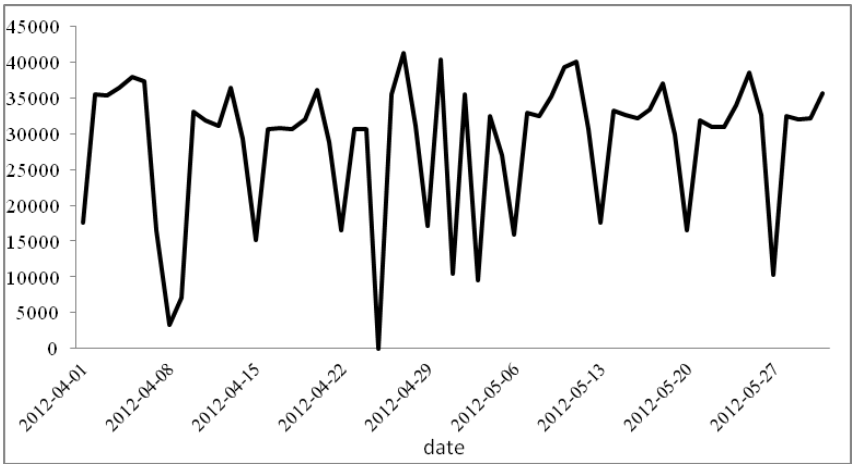


Figure 5. Time series of the overall number of withdrawals in the period April-May 2012

Analysing Figure 5, which presents the time series of the overall number of withdrawals for the period April-May 2012, one may notice that the number of withdrawals on Sundays is much smaller than on other days of the week. In addition, if Sundays are also non-trading days (e.g. Easter, or the Feast of Pentecost) the number of withdrawals is even smaller. Moreover, quite visible calendar effects include the increase in the number of withdrawals prior to upcoming holidays and longer periods when people do not work ("long weekends") and an increase in the number of withdrawals just before the 10^{th} and 30^{th} day of each month.

The analysis of the plots of time series is only suggestive of the occurrence of the calendar effect. In order to formally verify whether seasonal and calendar effects are indeed present in the series being studied we performed an analysis based on an examination of the variability of the average withdrawals in an individual year, month or week, and then calculated and compared other basic descriptive statistics (quantiles, minimum, maximum, standard deviation, coefficient of variation, skewness and kurtosis). The verification of individual calendar effects was based on comparing the respective withdrawals at individual moments. We proceeded with mean tests, such as the non-parametric analysis of variance (Kruskal-Wallis ANOVA rang test) and the Dunn test. In all the tests the significance level was set equal to 5%.

As in the case of testing for seasonality, the examination of calendar effects was conducted in two stages. First, we tested the existence of seasonal and calendar effects using the sum of the daily withdrawals from all 222 ATMs. This approach was used to determine the general trends, which are present in time series of withdrawals from ATMs during days of the week and particular periods of a month or on special days of the year.

The second stage was a similar analysis carried out separately for each of the ATMs. Detailed results of this analysis will be presented for the six ATMs selected. The knowledge of the occurrence of individual calendar effects for each ATM may turn out to be very useful when choosing methods of predicting the number of withdrawals.

In the analysis the following types of calendar effects were taken into account:

- Day of week. This effect allows us to capture differences in the numbers of withdrawals in ATMs over different days of the week.
- Day of month. This effect allows us to select the days of the month on which the number of withdrawals significantly rises/drops.
- Month. In this analysis it is assumed that in some months of the year the number of withdrawals may significantly rise.
- A Special event in the year. This effect allows us to analyse the number of withdrawals from ATMs on special days like holidays or long weekends. Five special events which may influence the number of withdrawals were taken into account:
- E1-non-trading days, e.g. New Year's Day, Easter, Christmas; E2- trading days during a long weekend; E3-trading days prior to a long weekend or holidays; E4-trading days after a long weekend or holidays; E5-holidays, for example Grandfather's Day, Grandmother's Day, Valentine's Day, Women's Day.

The study of the effect of the day of the week was carried out using original data. Since some effects may overlap, although the effect of the day of the week affects the evolution of the number of withdrawals most strongly, the remaining effects were examined based on weekly-seasonally-adjusted data.

4.3. The effect of the day of the week

In Figure 6 boxplots for the time series of the overall number of withdrawals on different days of the week in the period 2010-2012 are presented. Information on measures of position, measures of dispersion and asymmetry for withdrawals on different days is presented in Table 5.

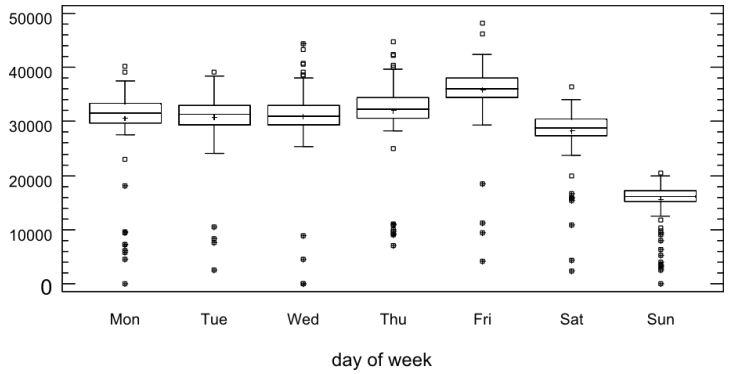


Figure 6. Boxplots of the overall number of withdrawals in the period April-May 2012

The results of the analysis indicate a variation in the number of withdrawals for individual days of the week. It can be seen from Figure 6 that, on average, the highest number of withdrawals from ATMs takes place on Fridays, and the lowest - on Saturdays and Sundays. From Monday to Thursday the average number of withdrawals is relatively stable. This test indicates the statistical significance of the differences between the average numbers of withdrawals on individual days of the week (p-value ≈ 0). These results were also confirmed by the Dunn test.

The analysis of the results presented in Table 5 shows that the variation in the number of withdrawals for the individual days of the week is on a similar level. The coefficient of variation is quite low for all days of the week (14-22%).

In addition, withdrawals on all days are characterized by relatively high left asymmetry. In other words, the predominant number of withdrawals is greater than the daily average. High kurtosis, on the other hand, indicates a high level of flattening of the distribution, and thus the existence of a large number of values is similar to the mean value for all days of the week.

Table 5. Descriptive statistics of the time series of the overall numbers of withdrawals on different days of the week

Day of the week	Average	Median	Standard deviation	Coeff. of variation [%]	Skewness	Kurtosis
Monday	30 538.5	31 569	6308.62	20.66	-3.00	9.97
Tuesday	30 848.1	31 266	4704.21	15.25	-3.27	16.06
Wednesday	31 066.8	30 947.5	5616.18	18.08	-2.96	15.24
Thursday	31 986.7	32 300.5	5743.08	17.95	-2.44	8.66
Friday	35 820.9	36 025	5016.39	14.00	-3.36	18.04
Saturday	28 235.7	28 850	4383.95	15.53	-3.25	14.59
Sunday	15 519.3	16 149	3384.46	21.81	-2.57	7.33

A similar analysis was performed for the six ATMs selected. The results are presented in Figure 7.

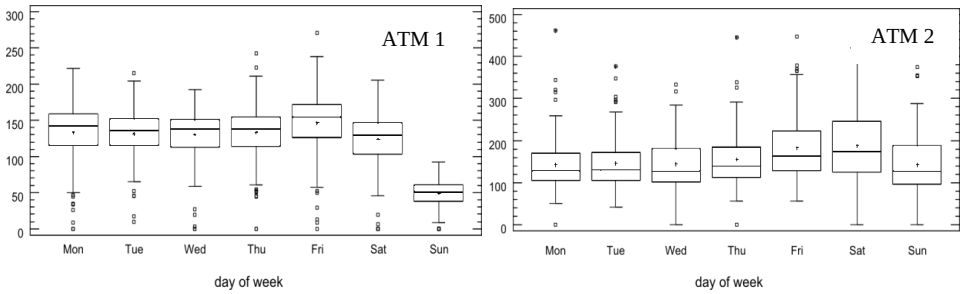


Figure 7. Boxplots of the overall number of withdrawals in six selected ATMs on different days of the week

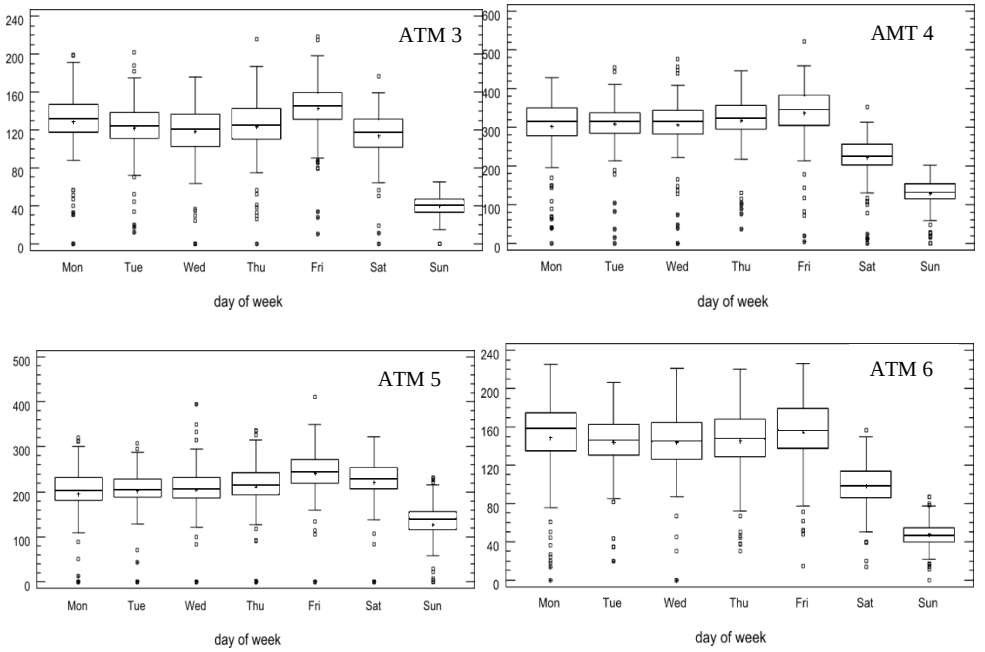


Figure 7. Boxplots of the overall number of withdrawals in six selected ATMs on different days of the week (cont.)

The results of the analysis obtained for the six ATMs selected confirm the earlier findings obtained for the sum of withdrawals on individual days of the week for all the ATMs. It is worth noting that for the cash machine located in Zakopane (ATM machine No. 2) the average number of withdrawals on Sunday did not differ significantly from the corresponding values on other days of the week. For ATMs No. 1 and No. 5, located in a store and a hypermarket, the average number of withdrawals on Saturday does not deviate significantly from the number of withdrawals on weekdays.

4.4. The effect of the day of the month

In many economic publications the effect of the 10th day of the month is examined. This phenomenon is related to the fact that a significant part of society receives salaries (by bank transfer to bank account or in the form of cash) on this particular day of the month. In order to verify whether the day of the month causes significant differences between the number of withdrawals from ATMs, the boxplots of the numbers of withdrawals on different days of the month are presented in Figure 8. Table 6 presents the basic descriptive statistics of the series.

An analysis of the boxplots presented in Figure 8 shows that the number of withdrawals from ATMs varies among different days of the month. It is noteworthy that the average number of withdrawals rises between the 6th and the 10th day of the month. On the 10th day of the month the average reaches its

maximum. This phenomenon can be explained by the fact that immediately prior to that date and right after that day many people pay cash to make various payments (e.g. to pay the bills) since the withdrawal date very often falls on this day. In contrast, the number of withdrawals on the 11th day of the month is significantly smaller than the day before.

Similar results were obtained by Kufel (2010). However, he noticed the highest number of withdrawals not only on the 10th but also on the 11th days of the months. He claimed that the latter follows from the fact that many people receive salaries in their bank accounts before the 10th day of the month.

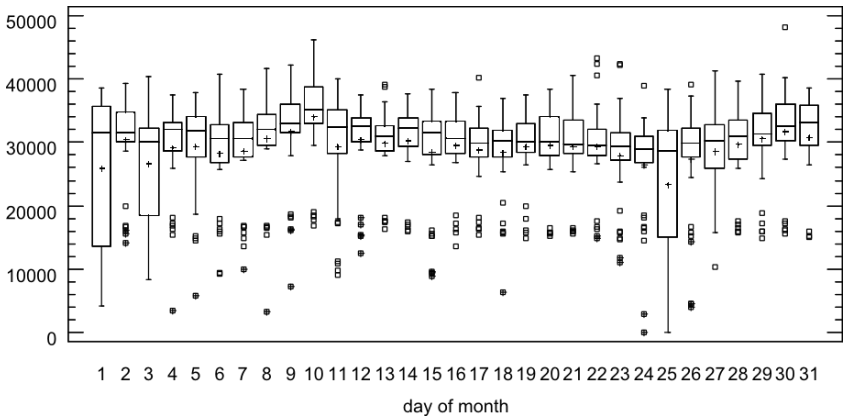


Figure 8. Boxplots of the overall number of withdrawals in the six selected ATMs on different days of the month.

A similar structure and variability in the number of withdrawals is observed between the 25th and 30th day of the month and on the 31st day of the month. Between the 25th and 30th day of the month the average number of withdrawals rises and reaches the maximum on the 30th day of the month and then significantly drops on the 31th day of the month. The source of this phenomenon seems similar to the previously described one.

When analysing the descriptive statistics calculated for the number of withdrawals on individual days of the month (Table 6) one may note quite a high coefficient of variation for the 1st day of the month compared to other days of the month. This is due to the fact that on the first day of the month there are three holidays a year (in January, May, November), and withdrawals on these days are relatively small. A similar situation can be observed on the 3rd, 11th and 25th day of the month. Holidays also fall on these days (Constitution Day, Independence Day, Christmas), although in the case of these dates the differences between the numbers of withdrawals compared to other days of the month are less pronounced. The results of the means and the median tests in each group indicate that the day of the month in fact leads to differences between the numbers of withdrawals from ATMs.

Figure 8 shows the results of a similar analysis for the six ATMs selected. The results obtained confirm the general rule concerning the number of withdrawals on particular days of the month. There are, however, some differences between the results obtained for each of the ATMs, namely the effect of the 10th day of the month is notable for five out of the six ATMs analysed. For ATMs No. 1 and No. 5 this phenomenon is especially visible because the average number of withdrawals on this day exceeds by far the number of withdrawals on the remaining days of the month.

As expected, the effect of the 10th day of the month is not evident for the ATM machine located in Zakopane. This follows from the fact that this machine is used mainly by tourists who normally plan their trip in advance and do not expect that they will receive their cash or transfer exactly on the 10th day of the month.

Table 6. Descriptive statistics of the time series of the overall numbers of withdrawals on different days of the month

Day of the month	Average	Median	Standard deviation	Coeff. of variation [%]	Skewness	Kurtosis
1	25832.3	31453	12182.5	47.16	-0.75	-1.14
2	30384.4	31548.5	6857.15	22.57	-1.25	0.75
3	26654.3	30141	8957.86	33.61	-0.92	-0.43
4	29242.5	32046	7429.04	25.41	-1.80	3.25
5	29414.3	31902	7357.85	25.01	-1.66	2.32
6	28275.1	30603.5	7677.42	27.15	-1.12	0.66
7	28686.4	30573.5	7442.66	25.94	-1.18	0.36
8	30617.1	32158	7965.74	26.02	-1.59	3.27
9	31630.1	33019	8114.45	25.65	-1.36	1.54
10	34033.5	35230.5	7501.26	22.04	-1.10	0.89
11	29401.3	32400	8892.48	30.25	-1.25	0.33
12	30422.6	32679	6386.73	20.99	-1.79	2.25
13	29924.3	31026.5	5713.61	19.09	-1.22	1.20
14	30239.1	32301.5	6079.01	20.10	-1.48	1.28
15	28528	31569	8352.45	29.28	-1.35	0.65
16	29577.5	30714.5	6095.58	20.61	-1.33	1.29
17	28795.7	29810.5	5713.33	19.84	-1.03	0.99
18	28382.9	30174.5	6787.38	23.91	-1.52	2.35
19	29335.9	30077	5754.76	19.62	-1.29	1.26
20	29467	30116	6517.18	22.12	-1.02	0.59
21	29392.6	29770.5	6488.82	22.08	-0.85	0.59
22	29330.2	29534	6730.48	22.95	-0.44	0.86
23	27928.6	29404.5	7650.05	27.39	-0.65	0.31
24	26514.5	28907.5	8400.37	31.68	-1.71	2.93
25	23369	28703	12362.9	52.90	-0.86	-0.65
26	27449.1	29926	9216.56	33.58	-1.47	1.43
27	28671.5	30178.5	6766.04	23.60	-0.82	0.93
28	29713.8	30916	6376.17	21.46	-0.95	0.49
29	30553.1	31315	6921.78	22.65	-0.95	0.58
30	31693.4	32540	7681.09	24.24	-0.78	0.60
31	30867.6	33090	7267.9	23.55	-1.38	0.98

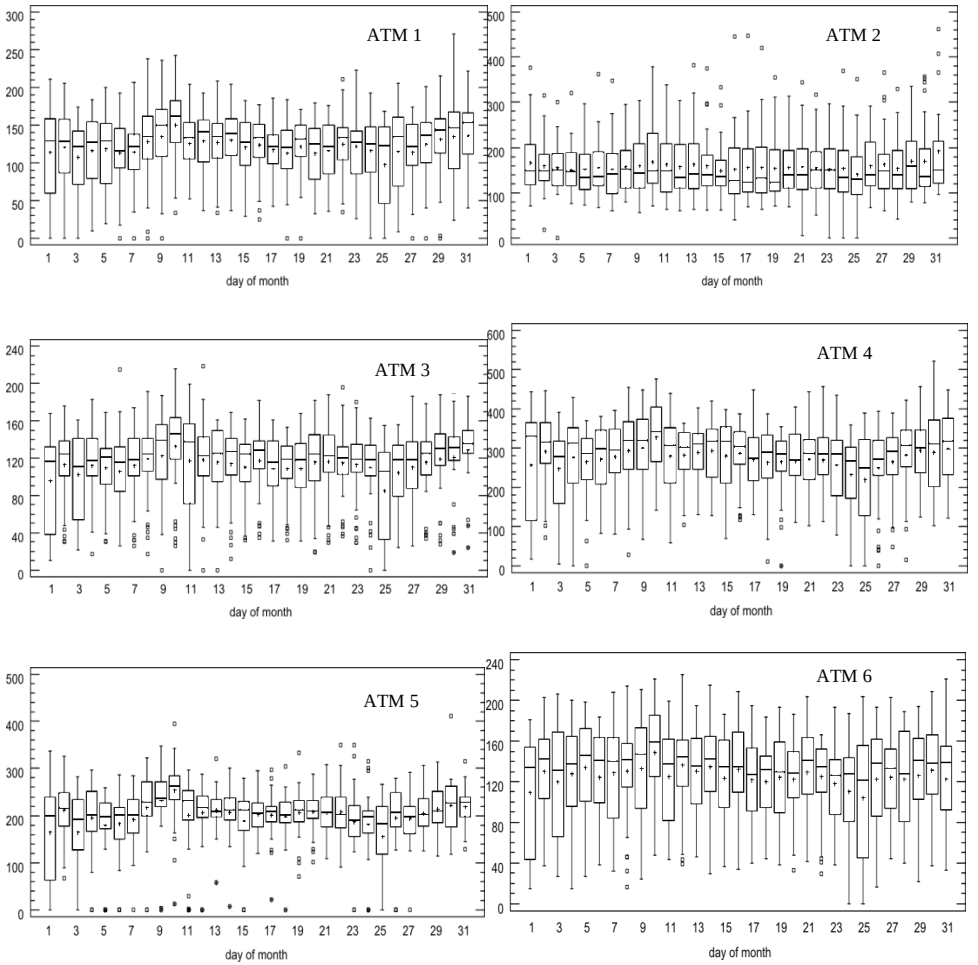


Figure 9. Boxplots of the overall number of withdrawals in the six selected ATMs on different days of the month

4.5. The effect of the month of the year

In this subsection we present the results of examining the differences between the numbers of withdrawals from all ATMs in particular months of the year. The results are presented in Table 10 and Figure 11.

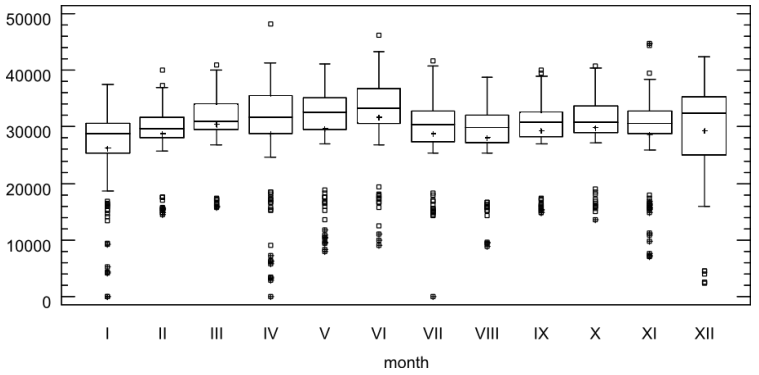


Figure 10. Boxplots of the overall number of withdrawals in all ATMs in different months of the year

The results presented in Figure 10 indicate that in the early months of the year the average number of withdrawals from ATMs achieves lower values compared with other months of the year. The latter may be a consequence of the usual increase in household expenditure during the holiday period (St. Nicholas', Christmas, New Year's Eve), which implies a greater demand for cash (see the boxplots for December) and leads to a reduction in the financial capacity of the public at the beginning of the year.

The slightly increased number of withdrawals in June also seems worth underlining. This may be related to the beginning of preparations for the holidays. The results of the analysis of variance for medium-sized withdrawals in individual months indicate statistically significant differences. For both tests p-value ≈ 0.

For most months of the year the value of the coefficient of variation (see Table 7) calculated for the number of withdrawals reached low levels (approximately 20%). Slightly higher values of this measure were reported for January, April, May, November and December.

Table 7. Descriptive statistics of the time series of the overall number of withdrawals in different months of the year

Month	Average	Median	Standard deviation	Coeff. of variation [%]	Skewness	Kurtosis
I	26292.8	28865	8604.66	32.73	-1.36	1.35
II	28738.4	29693	5894.62	20.51	-1.16	0.92
III	30377.1	31036	6009.74	19.78	-1.28	1.18
IV	28749.5	31677	10307.1	35.85	-1.24	0.80
V	29741.9	32605	8734.87	29.37	-1.24	0.38
VI	31683.5	33335.5	8144.14	25.70	-1.20	0.76
VII	28767.6	30412	7090.69	24.65	-1.39	2.36
VIII	28118.4	29826	6623.56	23.56	-1.38	1.20
IX	29285.9	30723.5	6116.2	20.88	-1.23	0.95
X	29927.8	30854	6384.43	21.33	-1.16	0.74
XI	28615	30608.5	8047.09	28.12	-1.15	0.90
XII	29395.7	32428	9637.02	32.78	-1.31	1.27

Figure 11 shows the results of the analysis of the effect of the month of the year for the six ATMs selected. In general, the results of this analysis confirm the general trends observed in the time series of overall withdrawals from all ATMs in different months. For all six ATMs the test results showed significant differences between the numbers of withdrawals in individual months. Special attention must be paid to the results obtained for ATM No. 2, which is located in Zakopane. The average number of withdrawals clearly indicates the months in which the largest number of tourists visit Zakopane. These are the months during the summer holiday season and during the winter holidays. A relatively high average number of withdrawals occurs also in December and January, i.e. during Christmas and the New Year holiday period.

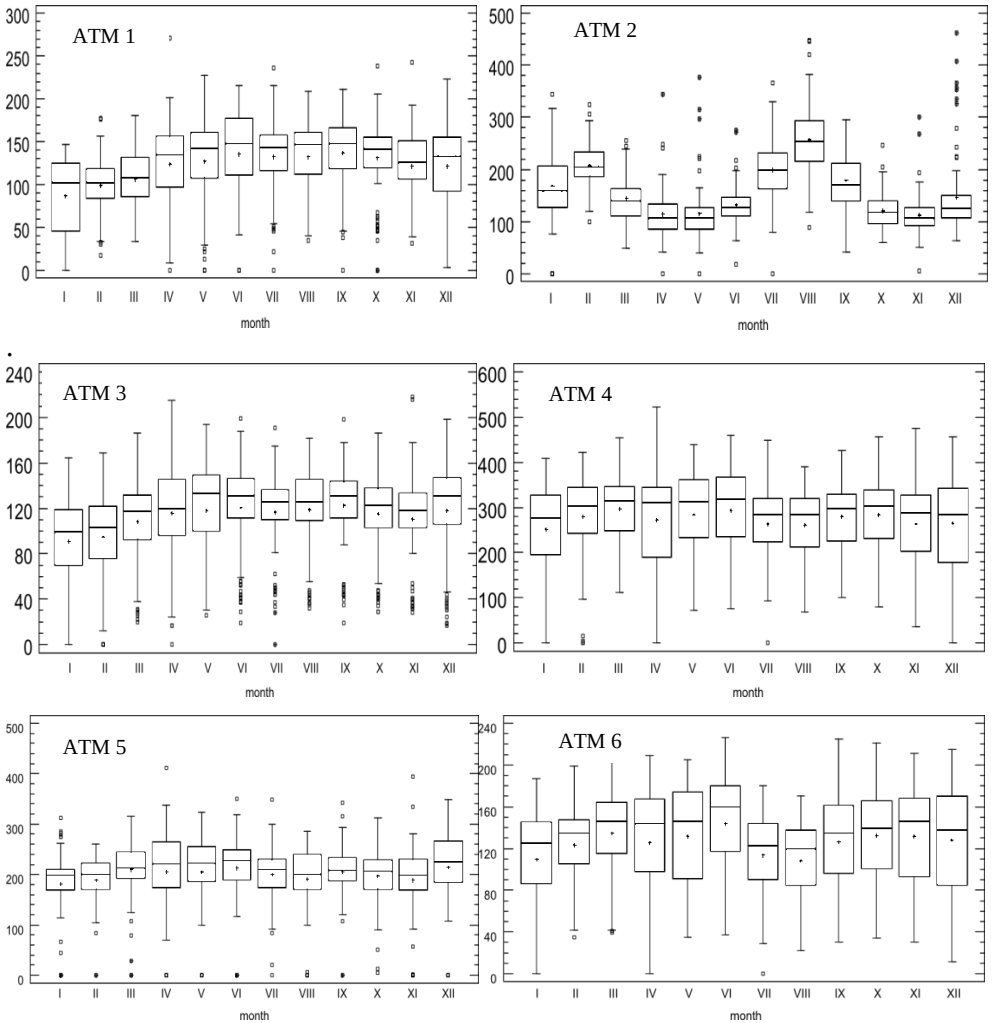


Figure 11. Boxplots of the overall number of withdrawals in the six selected ATMs in different months of the year

4.6. The effect of special days

In the last stage of the research we examined whether the number of withdrawals on the five types of the special day defined in previous sections is significantly different from the corresponding values reported for usual calendar days. The results of this comparison are presented in Table 8.

The comparison of means and medians for the days when the events E1, E2, E3, E4, E5 took place and for the regular calendar day shows that the number of withdrawals is indeed significantly different (p-value in nonparametric ANOVA close to zero).

In order to judge which events during the year cause differences between the numbers of withdrawals the Dunn test was additionally performed. This test shows the significance of the difference between the means in two selected subgroups. The results indicate that only the E4 event (the day after a long weekend) does not significantly affect the number of withdrawals from ATMs in comparison with other days. The remaining types of events statistically significantly affect the number of withdrawals from ATMs. The analysis of the boxplots presented in Figure 12 also confirms the results of the test.

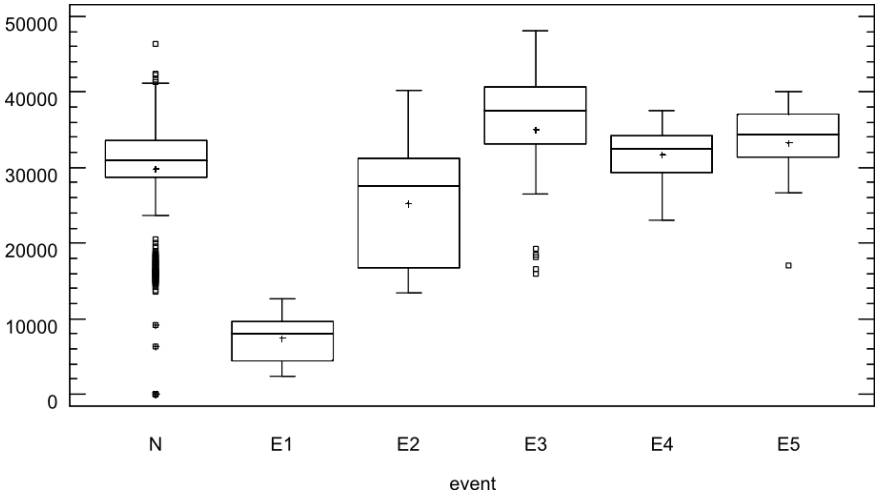


Figure 12. Boxplots of the overall number of withdrawals from all ATMs on five types of the special day and remaining (usual) days.

Table 8. Descriptive statistics of the time series of the overall numbers of withdrawals on five types of the special day and remaining (usual) days.

Type of special day	Number of days	Average	Median	Standard deviation	Coeff. of variation [%]	Skewness	Kurtosis
N	945	29843.6	30927	6532.5	21.89	-1.33	2.06
E1	38	7473.03	8090	3040.64	40.69	-0.29	-1.23
E2	42	25253	27657	8426.37	33.37	0.08	-1.47
E3	33	35065.8	37465	8612.17	24.56	-1.10	0.40
E4	22	31773.8	32564	3653.17	11.50	-0.62	0.41
E5	16	33350	34367	5726.79	17.17	-1.64	3.48

On E1-type days, i.e. on non-trading days, one may notice a reduction in the average number of withdrawals in comparison with other days. This is due to the fact that during the holidays, most shops are closed and potential customers usually leave or stay at home. As a result, the availability of ATMs is reduced and the demand for cash falls, and hence also the number of withdrawals.

The same is true on business days that occur within long weekends. On such days the number of cash withdrawals also decreases. This is most likely a consequence of the fact that people take trips away from home. However, on these days the number of withdrawals may rise in tourist destinations like Zakopane.

An increase in the number of withdrawals from ATMs on days before holidays or long weekends (E3) may result from the fact that before this type of event many people usually withdraw more cash in order to prepare for celebrations or trips. It seems natural that after such events, i.e. after spending cash previously withdrawn, there is a need to once again take cash from an ATM. The results of the research show that the average number of withdrawals after a long weekend (E4 event) is slightly higher than on normal days, but this is not a statistically significant difference. As already mentioned, the analysis showed a significant increase in the number of withdrawals from ATMs on occasional holidays. One can, therefore, presume that before this type of holiday the demand for cash increases. The following Figure (Figure 13) shows the results of similar studies carried out for the six ATMs selected.

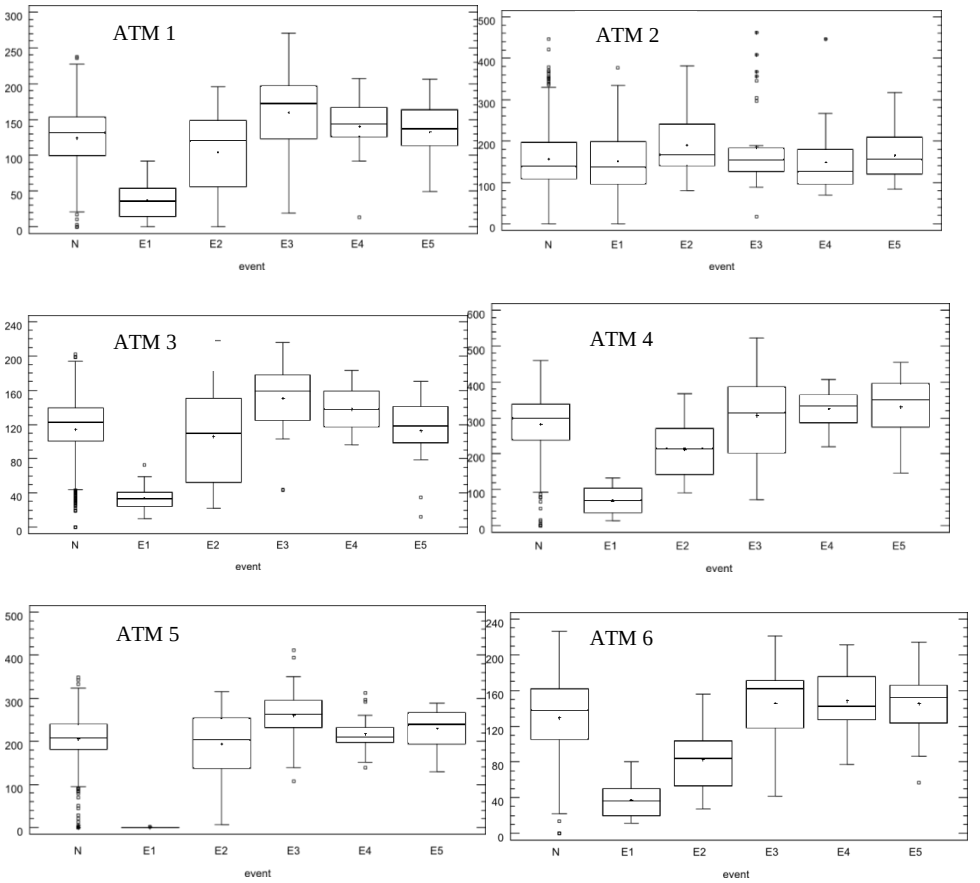


Figure 13. Boxplots of the overall number of withdrawals from the six selected ATMs on five types of the special day and remaining (usual) day.

The analysis of the number of withdrawals from the six ATMs selected on five types of the special day gives mixed results (see Figure 13). The results of the analysis of variance for all ATMs showed a significant difference between the average numbers of withdrawals on different types of the special day. However, when focusing on the six ATMs selected the difference between the average numbers of withdrawals on special days and on normal ones was not significant in all cases.

For all ATMs except ATM No. 2 the average number of withdrawals on holidays was different from the average number of withdrawals on normal days. The effect of working days during the long weekend causes a difference in the number of withdrawals for four out of the six ATMs (except ATMs No. 3 and No. 5).

The effect of pre-holiday days significantly affects the average number of withdrawals from five out of the six ATMs (except ATM No. 4), while the effect of after-holiday days (E4) significantly affects the number of withdrawals in

ATMs No. 3, 4, 6. The effect of specific days of E5 type turned out to be important only for ATM No. 3.

To sum up this part of the analysis, it can be concluded that all the effects examined have a significant impact on the number of withdrawals from ATMs operated by the Euronet company for at least half of the six ATMs selected. Only in the case of the number of withdrawals on the E5-type days (occasional holidays like Grandmother Day, Grandfather Day, Valentine's day, Women's Day) were no statistically significant differences found for most of the ATMs. One can note, however, that certain properties are individual to each ATM. For this reason, it seems that in order to effectively use calendar effects in forecasting models the respective testing should be carried out separately for each ATM.

Tables 9 and 10 contain information on the number of ATMs (nominal and percentage) for which the individual calendar effects turned out to be a statistically significant (at the level of 5%) factor influencing the average number of withdrawals. What was also examined was whether the type of location can significantly affect the structure of withdrawals on the five types of the special day.

On the basis of Table 9 one may claim that the day of the week substantially affects the number of withdrawals in 221 out of the 222 ATMs analysed. For more than 70% of the ATMs, it also turned out that the day of the month is an important factor affecting the volume of withdrawals. A statistically significant effect of the day of the month on the number of withdrawals was reported for all devices in the category "other". Also in the case of the group of ATMs linked directly with trade locations (supermarkets, shopping centres) the share of machines for which significant differences between the average number of withdrawals on different days of the month occurred was high and oscillated at around 80%. For the remaining types of location this share was lower. For a very high share of ATMs (93.9%) the number of withdrawals turned out to depend on the day of the month. In this category the smallest impact was reported in the case of ATMs located in bank branches (85.7%).

Table 9. The results of testing calendar and seasonal effects (the effects of days, weeks, months)

Type of localization	Day of the week		Day of the month		Month of the year	
	N	[%]	N	[%]	Number	[%]
Bank Branch	72	100	43	59.5	62	85.7
Hipermarket	36	100	30	82.5	34	95
Shop	31	100	20	64	30	96
Shopping Center	30	100	26	85.7	29	95.9
Petrol Station	23	95.8	16	64.9	23	97.3
Transport	4	100	3	78.6	4	100
Entertainment	1	100	1	66.7	1	100
Hotel	1	100	1	50	1	100
Other	23	100	23	100	23	100
Total	221	99.5	158	71.3	208	93.9

Table 10. The results of the analysis of calendar effects (the effect of special types of day)

Type of localization	E1		E2		E3		E4		E5	
	N	[%]	N	[%]	N	[%]	N	[%]	N	[%]
Bank Branch	69	96.4	42	58.3	29	40.5	14	19.1	2	2.4
Hipermarket	36	100	22	60	7	20	4	12	0	0
Shop	30	95.9	15	46.9	9	28.6	4	12.2	0	0
Shopping Center	26	87.5	11	37.5	17	55	6	20	1	2.5
Petrol Station	24	100	14	56.8	5	21.6	3	10.8	1	5.4
Transport	3	83.3	1	16.7	1	16.7	1	16.7	0	0
Entertainment	1	100	0	16.7	0	0	0	16.7	0	0
Hotel	1	90.5	0	28.6	0	38.1	0	19.1	0	7.1
Other	17	75	12	50	6	25	12	50	6	25
Total	210	94.5	105	47.4	77	34.5	37	16.7	7	3.1

In Table 10 we present the similar results for the time series of withdrawals on special days. The analysis shows that for 69 ATMs the average number of withdrawals on E1-type days significantly differs from the average on normal days (Normal (usual) days are understood as the days which are not E2-E5 special days). In this group of ATMs the lowest sensitivity to changes in the number of withdrawals on specific days is shown by these ATMs which are located in places that belong to the "other" category, and the highest sensitivity for the ATMs located in hotels, petrol stations and shops.

In the case of other types of the special day changes in the number of withdrawals turned out to be considerably smaller. The average number of withdrawals on trading days during holidays or long weekends (E2-type days) differs significantly from the average volume of withdrawals on normal days for almost half of the ATMs. However, this share is distributed very differently for each location: in the case of ATMs located in stores and bank branches this is around 60%, while for ATMs in entertainment centres the share is only about 16.7%.

On E3-type days, that is days before holidays and long weekends, the number of withdrawals from ATMs changes significantly in the case of 34.5% of all ATMs. Usually these changes affect ATMs operating in hypermarkets.

Of all the types of the special day the smallest changes in the number of withdrawals were observed for E5 type days. Only for nine ATMs (3.1% of the total number of machines analysed) the nonparametric ANOVA shows a statistically significant difference between the average number of withdrawals on these days and normal days.

Differences between the results of research on the importance of some of the effects in the first and second part of the analysis (i.e. the results reported in the section dealing with the overall number of withdrawals and withdrawals from the six individual ATMs) arise from the fact that the impact of the withdrawals from specific ATMs on the time series of the total number of withdrawals is significantly different. This means that the statistical significance of the calendar

effect is mostly caused by withdrawals from those ATMs which are generally characterized by the largest withdrawals (i.e. ATMs located in shopping malls and hypermarkets in large cities).

The results of the analysis of calendar and seasonal effects, which were carried out in the framework of consumer behaviour, which shows the overall number of withdrawals, clearly demonstrate the importance of the phenomena discussed. The results of the study show that trends associated with selected moments within a week, month or year in which money is withdrawn in larger or smaller amounts, reflect certain regularities and habits which characterize customers.

The conclusions of the second part of the study, in which calendar and seasonal effects were examined at the level of individual ATMs, suggest the necessity of ATM-specific adjustment of prognostic models. Taking into account only those calendar effects which turned out to be significant in describing withdrawals from a particular ATM allows us to create an optimal individual prediction model. In other words, the results of the study suggest that for understanding and predicting the number of withdrawals from ATMs it is necessary to take into account such effects as the effects of the day of the week, the effects of the week of the month, the effects of the month of the year, as well as the effects of holiday days.

5. Conclusions

In the eighties the ATM became an important part of the international financial system with respect to financial services. The role of ATMs in Poland is still very important and it seems to us it is going to be even more important. The sound management of an ATMs network can reduce the operating costs of ATMs. This is important not only for the operator but also for the customer. A lower cost of average ATM maintenance can convince the owner to install further ATMs in order to save the customers' time, especially of those who are residents in suburbia and on the outskirts.

The main part of the total cost is cash, which is in an ATM and does not bring any profit. Instead of generating profit it works for the customer. In ATM logistics it is very important that the proper amount of cash is in an ATM. Cash should be available to the customer for 24 hours a day but should be not in excess, because cash in ATMs does not generate any profit. Thus, an important problem is how to stock on ATM with cash. From this point of view it is very important to forecast correctly customer demand for cash (withdrawals) in the area where the ATM is installed.

In our paper we assess the main factors determining forecasts of demand for cash from ATMs. On the basis of the time series of withdrawals from ATMs from Małopolska and Podkarpackie provinces delivered by Euronet we prove the existence of calendar effects.

The results of our study indicate that the size and the number of withdrawals are different on particular days. The dependence of the size of withdrawals on the day of the week, the week of the month or the month of the year is well established in the case of customers from the regions being studied.

Since many factors determine the number and size of withdrawals, and for particular ATMs the characteristics may be quite different, forecasts should be calculated for each ATM separately. It is necessary to take into account only those calendar effects which are significant.

The most important properties are reflected in seasonality and calendar effects. The presence of these effects has been proved in this contribution. The authors did not check other properties of ATMs time series like stationarity or long memory. However, the authors are going to conduct further analyses with respect to these properties.

As mentioned in the introduction, the pattern of withdrawals (seasonality, calendar effects) and their statistical properties may be helpful in choosing the proper methods and models for forecasting withdrawals from ATMs.

Our results confirm the importance of calendar effects on withdrawals from the ATMs which we studied.

REFERENCES

BELL, W. R., HILLMER, S. C., (1983). Modeling Time Series with Calendar Variation. Journal of the American Statistical Association 78, pp. 526–534.

CARLSEN, M., STORGAARD, P. E., (2010). Dankort withdrawals as a timely indicator of retail sales in Denmark, Danmarks Nationalbank Working Papers, No. 66.

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, 371, 75, pp. 487–496.

CLEVELAND, W. P., GRUPE, M. R., (1983). Modeling time series when calendar effects are present. Applied Time Series Analysis of Economic Data, Zellner, A. (editor), U.S. Department of Commerce, U.S. Bureau of the Census, Washington D. C., pp. 57–67.

DURATE, C., RODRIGUES, P., RUA, A., (2016). A Mixed Frequency Approach to Forecast Private Consumption with ATM/POS Data, International Journal of Forecasting (forthcoming)

ESTEVES, P. S., (2009). Are ATM/POS Data Relevant When Nowcasting Private Consumption?, Banco de Portugal Working Paper, 25.

FINDLEY, D. F., MONSELL, B. C., (2009). Modeling Stock Trading Day Effects Under Flow Day-of-Week Effect Constraints. Journal of Official Statistics, Vol. 25 (3), pp. 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 & Economic Statistics, Vol. 16 (2), pp. 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, pp. 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, pp. 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, pp. 743–53.

GALBRAITH, J. W., TKACZ, G., (2007). Electronic Transactions as High-Frequency Indicators of Economic Activity. Bank of Canada Working Paper, 2007–58.

GERDES, G. R., WALTON, J. K., LIU, M. X, PARKE, D. W., (2005). Trends in the Use of Withdrawal Instruments in the United States. Federal Reserve Bulletin 91 (Spring), pp. 180–201.

HANSEN, P. R., LUNDE, A., NASON, J.M., (2005). Testing the Significance of Calendar Effects. Working Paper 2005-2, Federal Reserve Bank of Atlanta.

HOLDEN, K., EL-BANNANY, M., (2004). Investment in information technology systems and other determinants of bank profitability in the UK, Applied Financial Economics, 14, pp. 361–5.

KONDO, K., (2011). Do ATMs influence bank profitability in Japan? Applied Economics Letters, 2010, 17, pp. 297–303

KUFEL, T., (2010). Ekonometryczna analiza cykliczności procesów gospodarczych o wysokiej częstotliwości obserwowania, **[**Econometric analysis of the cyclical nature of economic processes with high frequency data**]**, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, Toruń.

KUMAR, S., KUMAR, D.K.S., (2014). ATM Usage Price in India, IUP Journal of Operations Management, Vol. 13, Issue 3, pp. 65–72.

LIU, L. M., (1980). Analysis of Time Series with Calendar Effects. Management Science 26, pp. 106–112.

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.

RODRIGUES, P., ESTEVES, P., (2010). Calendar effects in daily ATM withdrawals. Economics Bulletin, vol. 30, No. 4, pp. 2587–2597.

SCHMITZ, S., WOOD, G., (2006). Institutional Change in the Withdrawals System and Monetary Policy. Routledge London.

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, pp. 416–421.

SNELLMAN, H., VIREN, M., (2009). ATM networks and cash usage. Applied Financial Economics, 19 (10), pp. 841–851.

SULLIVAN, R., TIMMERMANN, A., WHITE, H., (2001). Dangers of Data Mining: The case of calendar effects in Stock Returns. Journal of Econometrics, 105, pp. 249–286.

TAKALA, K., VIREN, M., (2007). Impact of ATMs on the Use of Cash. Communications and Strategies, No. pp. 66, 47–61.

YOUNG, A.H., (1965). Estimating Trading-day Variation in Monthly Economic Time Series. Technical Paper 12, Bureau of the Census.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 723–736

THE IDENTIFICATION OF TRAINING NEEDS FOR HUMAN CAPITAL QUALITY IMPROVEMENT IN POLAND – A STATISTICAL APPROACH

Marta Dziechciarz–Duda[1], **Józef Dziechciarz**[2]

ABSTRACT

The Ministry of Science and Higher Education has launched the Competency Development Programme in the form of additional financial means for activities to equip students with unique, the so-called soft skills necessary in scientific careers and on the labour market. Courses developing skills such as team work ability, leadership, creativity, independent thinking and innovative approach to problem solving will be financed.
For that purpose, a thorough analysis of needs is necessary. Existing databases describing the quality of human capital in Poland should be analysed in order to identify those competencies that graduates of universities are missing.
The arsenal of possible statistical tools applicable for that purpose covers a wide range of techniques, from the simplest methods of descriptive statistics to advanced multivariate statistical analysis.
The study will attempt to identify the missing soft competences based on existing statistical data, e.g. Polish human capital database.

Key words: soft competencies, human capital, statistical methods.

1. Introduction

The European and National Qualification Frameworks define the output of education in terms of knowledge, skills and personal attitudes. Although there is no direct indication on weights attributed to those three groups of teaching and learning results, one may expect some kind of equilibrium. The reality in higher education (and similarly in other segments of education and training system) is such that students are assessed mainly for knowledge than for skills, and in a minimal degree for acquired personal attitudes (here referred to as soft skills).

On the other hand, employers declare a completely opposite hierarchy of expected employee's characteristics. The most important are soft skills, than skills,

[1] Wroclaw University of Economics. E-mail: Marta.Dziechciarz@ue.wroc.pl.
[2] Wroclaw University of Economics. E-mail: Jozef.Dziechciarz@ue.wroc.pl.

and knowledge, although desirable, is considered as additional, supplementary criterion. In a 2015 survey, 77 percent of employers surveyed by CareerBuilder said they were seeking candidates with soft skills, and 16 percent of the respondents considered such qualities more crucial than hard skills (see for example The 10 Unique Soft Skills … (2015), Why Attitude is more Important … (2015). Similar results for Polish employers have been found in Dziechciarz et. al (2006), Kurkliński and Maszybrocki (2008), Maszybrocki (2010). An extensive literature review on the topic may be found in Getting Youth … (2013).

In response to this challenge, the Ministry of Science and Higher Education launched in mid-2015 the Competency Development Programme in the form of additional financial means for activities to equip students with unique, the so-called soft skills necessary in scientific careers and on the labour market. Courses developing skills such as teamwork ability, leadership, creativity, independent thinking and innovative approach to problem solving will be financed.

To guarantee the Competency Development Programme is a success, a thorough analysis of needs is necessary, including the assessment of compliance (convergence) of the declared level (self–assessment) of possessed soft competences with the declared employer's needs (demand – supply analysis). Existing databases describing the quality of human capital in Poland should be analysed in order to identify those competencies that graduates of universities are missing.

The possible approach to assess the level of soft skills is to measure it in tests proving that a person is able to use an individual skill, (on a given level. Such objective measurement is extremely expensive and time-consuming, and requires frequent updates of results. This is the reason that the most widespread mode of the assessment of the level of soft skills is the subjective approach. The respondent declares to which extend she/he is able to use an individual soft skill.

The natural arsenal of tools for the purpose of looking into large database of self-assessment statements is the multivariate statistical analysis framework, starting with basic descriptive statistics, along with correlation and dependence measures, factor and correspondence analysis, to classification techniques.

2. The objective of the analysis

The objectives of the analysis of compliance (convergence) of the declared level (self-assessment) of possessed soft competences with the declared employer needs (demand – supply analysis) are the following.

– Identifying the desired (by employers) competence profile and its confrontation with the declared (by potential employees) possession level of soft skills.
– Assessment of compliance (convergence) of the declared needs for soft competences (demand, employers' declarations) with the declared possession of soft competences (supply, potential employee declarations).
– Credibility assessment of respondents' declarations.

And additionally:

- Priority setting, whether demand or supply determines the directions of trainings.
- Decomposition (identification) of convergent and divergent indications.
- Testing of the applicability of selected multivariate statistical analysis tools and techniques for the purpose of soft skills analysis.

The fifth edition of the Study of Human Capital in Poland was selected as the main source of statistical data. The details of the study and the source data are available on the web page of the study (http://en.bkl.parp.gov.pl). The study will be referred to as BKL in further text. The database used contains information coming from a large number of respondents (http://bkl.parp.gov.pl/bazy–danych14). It covers over 64,000 employers and 70,890 of potential employees[3]. The database used for analysis in this article was created during the implementation of the 2014 edition of the Project *Study of Human Capital in Poland*. The database will be referred to as BKL2014 in further text.

2. Soft competencies

The classification of soft competencies, developed for the purposes of the Study of Human Capital in Poland, consists of twelve groups, both for employees and for employers. The list includes the following soft competences:

Table 1. The classification of soft competencies developed for the purposes of the Study of Human Capital in Poland

Symbol		Soft competency description
C01	Z01	Seeking and analysis of information, and drawing conclusions
C02	Z02	Technical imagination, handling and repairing technical devices
C03	Z03	Performing calculations
C04	Z04	Working with computers and using the Internet
C05	Z05	Artistic and creative skills
C06	Z06	Physical fitness
C07	Z07	Self-organisation of work and showing initiative
C08	Z08	Contacts with other people
C09	Z09	Organisation and conducting office work
C10	Z10	Managerial skills and organisation of work
C11	Z11	Availability
C12	Z12	Fluent use of Polish language (linguistic correctness, wide vocabulary, ease of speaking)

Note: the symbols with description C refer to Employees, the symbols with description Z refer to Employers. The soft competencies for Employees are further classified into subcategories. The list of subcategories is given in Appendix.

Source. BKL http://bkl.parp.gov.pl.

[3] Unfortunately, the database contains descriptions of variables only in Polish language.

2.1. Decomposition of competences in accordance with respondents' characteristics

In order to see the basic characteristics of the declared level of possessed soft competences, the simple plots have been used. The respondents with higher education diploma (least numerous group) manifest a relatively high self-assessment with low variability. Respondents with basic education show a much lower level of soft competences, accompanied with a much higher variability (figure 1).

As it was expected, the respondents with secondary education show a more optimistic picture of their soft competences than people with basic education and less favourable in comparison with those with higher education.

Such statements give moderate, cautious ground to assess that respondents carefully formulate their self-assessment, at least in accordance with the level of their education. It gives some reasoning for the use of subjective statements as the basis for analysis.

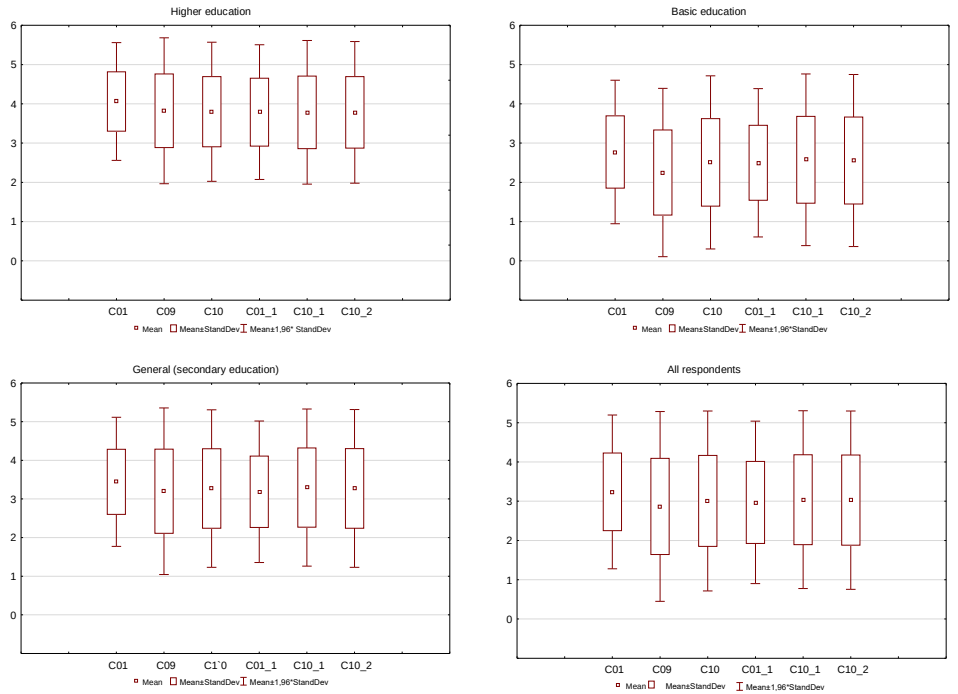


Figure 2. Box and whiskers plots, the level of possessed soft competences, declaration of potential employees broken down by educational groups

Source: Own calculation on data from BKL2014.

In order to identify soft competences, which have a tendency to strongly correspond with the level of education, the PROFIT analysis (*PROperty FITting*) has been used. The results of multidimensional scaling approach are illustrated in Figure 2. The obtained preference map is the basis for the assessment of the rela-

tions and interdependence between the objects and the respondents. PROFIT analysis creates the vector preference map, combining the perceptual map obtained by multidimensional scaling with the data on the preferences towards the surveyed objects from the point of view of their characteristics. This method puts together the results of multidimensional scaling and multiple regression analysis (see, e.g. Zaborski, (2013); Walesiak and Gatnar, (2004)).

A disquieting issue coming from the results is an observation on respondents with higher education. It shows that the level of education has a prevailing role in the self-assessment of the soft skill level. It would be more credible, provided young academics judge their skill in more moderation.

The labour market situation of a respondent (employed, unemployed) does not differentiate declarations of soft skills self-assessment.

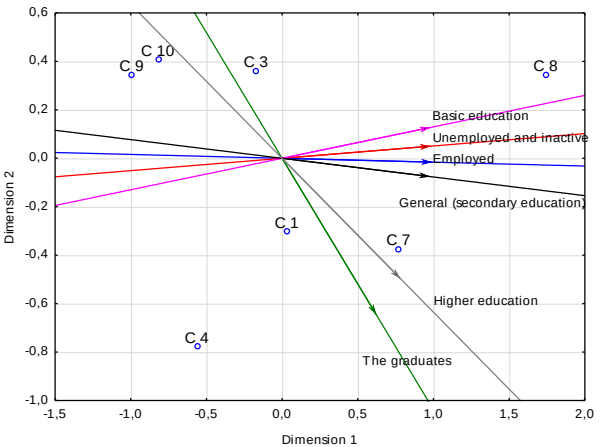


Figure 1. Evaluation of the level of competencies by employment and education

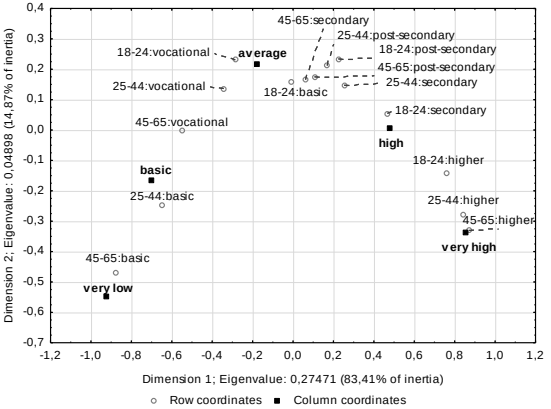
Source: Own calculation on data from BKL2014.

In order to identify soft competences, which have a tendency to strongly correspond with the age groups and level of education, the correspondence analysis has been performed. The upper part of Figure 2 illustrates the coincidence of the respondents' education level (4 categories) combined with the age of respondents (3 categories), along with a subjective evaluation of her/his soft competences.

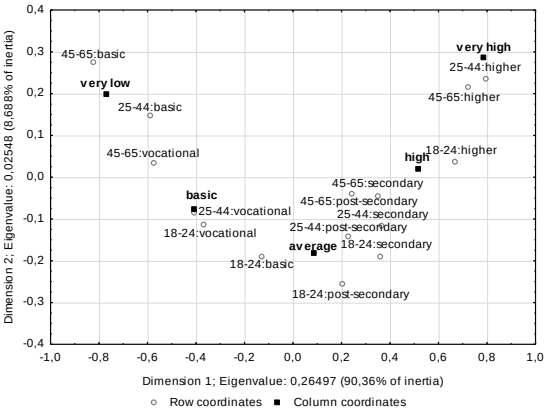
In the case of all three analysed competencies, the achieved percent of total inertia described in the two first dimensions reached a value of around 99%, individual inertia vary from 83.41% to 90.36% for the first dimension; and 8.69% to 14.87% for the second dimension.

In the study, a variant of correspondence analysis for many nominal variables was used, i.e. with multi-dimensional contingency matrix. A comprehensive description of the algorithm of the correspondence analysis, computational details, and its applications can be found in the classic text by Greenacre, [1984] or other descriptions, e.g. Stanimir [2005]. The data that was used for calculations describes the situation in Poland in 2014; BKL2014 database, analysed data subset: study of working-age population (70890 respondents).

C01: *seeking and analysis of information, and drawing conclusions*



C09: *organisation and conducting office work*



C10: *managerial skills and organisation of work*

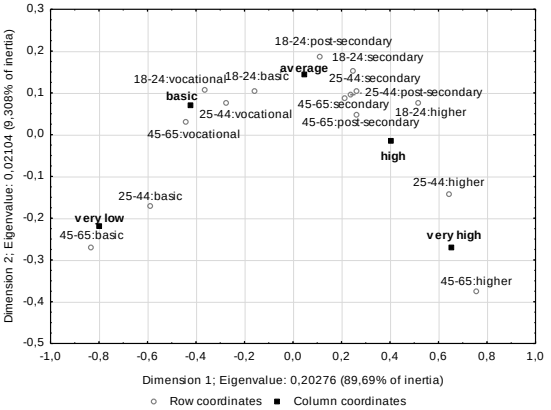


Figure 3. Correspondence analysis of self-assessment of soft skills combined with education and age groups

Source: Own calculations on data from BKL2014.

The visualized output of correspondence analysis with multi-dimensional contingency matrix gives more ground for approving the respondents' ability to a fair assessment of their soft competences. On may observe that the level of education and age groups strongly correspond with the declared level of soft competences. This correspondence is true for all three analysed competences.

2.2. Analysis for selected job type. Accountant

General inference for the whole sample of respondents may be used for general purposes only. The specific job type requires specific skills in general, and specific soft skills in particular. In Figure 4, a schematic match of the demand side with the supply side of competencies is shown. It this particular situation, the demand and supply of competencies for accountancy job.

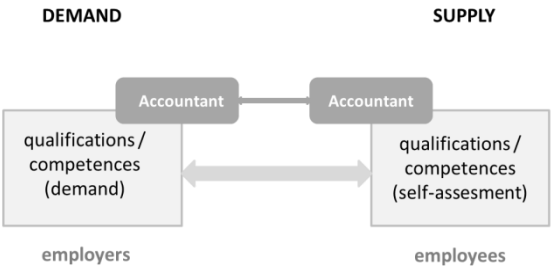


Figure 4. Schematic match of demand side with supply side of competencies

Source: Own elaboration based on BKL.

An attempt to compare expectations on the employers' side with the employee' judgement gives interesting insight into differences.

Evaluation of the level of competencies by Employers (demand)	Mean (1-4)	Evaluation of the level of competencies by Employees (self-assessment)	Mean (1-5)
Performing calculations	3,34	Performing calculations	4,07
Working with computers and using the Internet	3,16	Contacts with other people	4,28
		Team work	4,18
		Can easy make contacts with colleagues	4,13
		Being communicative	4,13
Self-organisation of work and showing initiative (organisation and conducting office works on time)	3,12	Conducting office works on time	4,13
		Making decisions	4,08
		Self-organisation of work and showing initiative	3,88
Contacts with other people, either with colleagues or customers	2,88	Logical thinking, analysis of facts	4,07
Seeking and analysis of information, and drawing conclusions	2,88	Fluent use of Polish language	4,05
Organisation and conducting office works	2,71	Organisation and conducting office works	4,03
Fluent use of Polish language (linguistic correctness, wide vocabulary, ease of speaking)	2,71	Continuous learning new things	3,95
Availability	2,38	Seeking and analysis of information, and drawing conclusions	3,88
Managerial skills and organisation of work	2,13	Resistance to stress	3,62
Physical fitness	2,07	Availability	3,72
Technical imagination, handling and repairing technical devices	1,75	Working with computers and using the Internet	3,70
		Basic knowledge of MS Office	3,53
Artistic and creative skills	1,67	Solving conflicts between people	3,70
		Managerial skills and organisation of work	3,62

Figure 5. Importance of soft skills. Assessment by employers of accounting department and self-assessment by employees

Source: Own calculations on data from BKL2014.

For the employer, the ability of effective work goes first. The employee looks for a good working atmosphere. The most substantial difference one may see in a high position of *working with computers and using the Internet* competence, ranked very high in employers' hierarchy, and low in employee's hierarchy. The difference in other direction is manifested by competence: *contacts with other people, either with colleagues or customers.* As shown in Figure 5, it is very high in employees' hierarchy, and considered not so important by employers (the discussed competence is divided into sub-competences on the side of the employee, details in Appendix).

2.3. Comparison of the assessment of competencies

Evaluation of self-assessment of competencies for various positions gives insight into the way employees consider their ability to take responsibility working on particular positions. In order to identify soft competences, which have a tendency to strongly correspond with the job type illustrated in Figure 6 and 7, the PROFIT analysis has been used once more. In Figure 6 the supply (employee) side is shown. One may easily see groups of jobs (e.g. customer advisor, HR specialist) with set of soft competencies (subcategories, the list in Appendix).

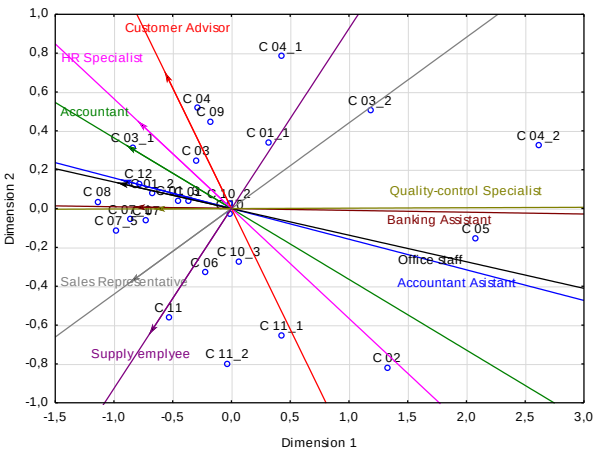


Figure 6. Evaluation of competencies for various positions (self-assessment, supply side; (subcategories, list in Appendix)

Source: Own calculations on data from BKL2014.

Evaluation of competencies for various positions required by employers gives insight into the way employers consider the set of skills needed to take responsibility working on particular positions. One may easily see groups of jobs with a set of soft competencies. To large extent, the sets shown in Figure 6 and 7 are the same (similar), but some differences might be seen in the assessment of job requirements with

respect to soft skills.

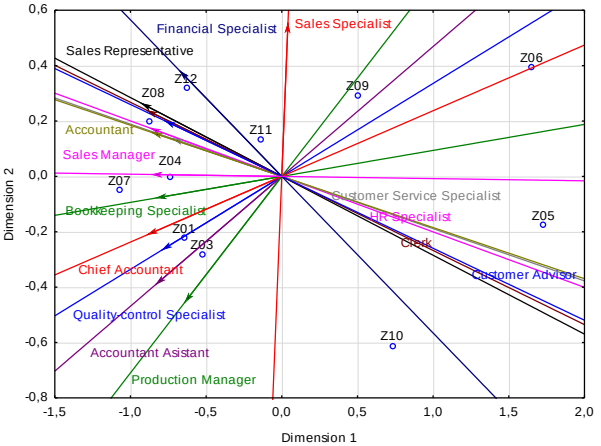


Figure 7. Evaluation of competencies for various positions (employer; demand side)

Source: Own calculations on data from BKL2014.

2.4. Classification of soft competencies

For more formalised insight into the similarities and differences of demand and supply side assessment, the classification study has been performed. The competences that are most needed to perform the position of an accountant were used as an exercise example. For classification, the hierarchical agglomerative method was used. This method proved to be most efficient in the sense of the carried out simulation experiment, Walesiak, Dudek, (2009). The technique provides a large homogeneity of identified classes. Due to the variable nature, the Generalized Distance Measure (GDM) technique for non-metric data was applied. `Dist.GDM` procedure of the platform `R` was used from `clusterSim` package. `Dist.GDM` procedure calculates Generalized Distance Measure for variables measured on ordinal scale or metric scale (ratio & interval). GDM2 method was used for variables measured on ordinal scale. `Average` method was employed for the classification. This hierarchical agglomeration technique is available in the `hclust` (`stats` package) of program R. As a result, a dendrogram was obtained.

Two experiments has been done. The results of classification on the demand side, employer, are displayed in Figure 8. Nine variables describing the assessment of the importance of competences formulated by employers were used for the classification.

The symbols in Figures 8 and 9 stand for: 1 – *Seeking and analysis of information, and drawing conclusions*; 2 – *Performing calculations*; 3 – *Working with*

computers and using the Internet; 4 – *Self-organisation of work and showing initiative (organisation and conducting office work on time)*; 5 – *Contacts with other people, either with colleagues or customers*; 6 – *Organisation and conducting office work*; 7 – *Managerial skills and organisation of work*; 8 – *Availability*; 9 – *Fluent use of Polish language (linguistic correctness, wide vocabulary, ease of speaking).*

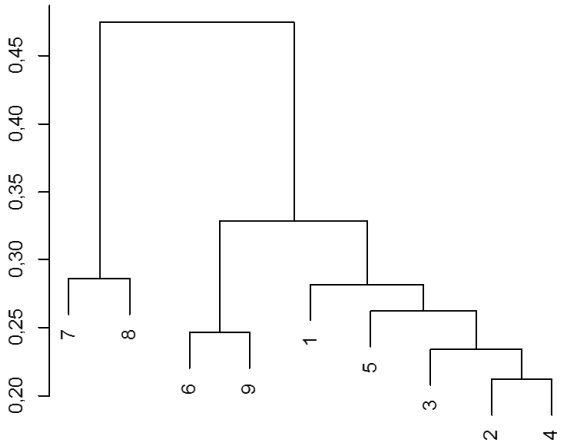


Figure 8. Dendrogram build on variables depicting employers' evaluation of the level of importance for soft competencies (demand side)

Source: Own calculations.

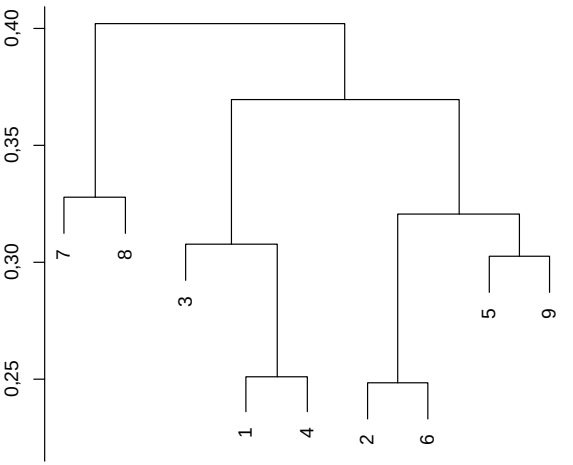


Figure 9. Dendrogram built on variables depicting employees' self-assessment of the level of possession of soft competencies (supply side)

Source: Own calculations.

From the results, it might be seen that despite differences the classifications are relatively similar. It has consequences for training design, and for recruitment policy.

3. Conclusions

In conclusion, it might be stated that, on the general level (for all respondents), there exists substantial compliance (convergence) of the declared level (self-assessment) of possessed soft competences with the declared employer's needs (demand side). This statement is also true for specific job types.

It means that in practice the competence profile desired by employers finds fulfilment in the declared (by potential employees) possession level of soft skills.

The statements on compliance are true conditionally, depending on the level of credibility assessment of respondents' declarations. Inferences draw from statistical analysis give ground for a careful acknowledgement that the statements contained in self-assessment may be considered true.

An important question on priority setting, whether demand or supply determines the directions of trainings, may be answered by saying that due to substantial similarity between statements on demand and supply side this question proved to be not crucial.

The identification of divergent indications should be done on a disaggregated level. Individual job types may produce such divergent indications, especially in situations in which the employee stresses the importance of work atmosphere, and the employer seeks for work efficiency and effectiveness.

For solving the listed research problems, selected multivariate statistical analysis tools and techniques proved their applicability, especially for the purpose of soft skills demand and supply analysis.

Acknowledgement

This study was carried out as a part of the project *Non-metric multivariate data analysis as a tool for study of adults situation in the context of demographic changes* financed by Narodowe Centrum Nauki (National Science Centre) in Poland. Project number: 2012/05/B/HS4/02499.

REFERENCES

BERNAIS, J., JĘDRALSKA, K., (eds.), (2015). Uniwersytet w perspektywie kształcenia przez całe życie [University in the Lifelong Learning Perspective]. UE Katowice.

BILANS KAPITAŁU LUDZKIEGO [HUMAN CAPITAL BALANCE], (2014). PARP, Warszawa; http://bkl.parp.gov.pl.

BORG I., GROENEN P., (2005). Modern multidimensional scaling. Theory and applications. Springer, New York.

DZIECHCIARZ, J., (2012). System jakości kształcenia [The Quality of Education], in: Dziechciarz, J., Błaczkowska A., Grześkowiak, A., Król, A., Stanimir, A., Analiza wybranych aspektów wyników egzaminu gimnazjalnego [Analysis of selected aspects of lower secondary school exam results], UE Wrocław.

DZIECHCIARZ, J., (2015). Pomiar i wycena wiedzy, umiejętności i kompetencji nabytych w formalnych i nieformalnych formach kształcenia [Measurement and Valuation of Knowledge, Skills and Competence Acquired in the Formal and Informal Education Forms], in: Wdowiński, P., (ed.), Nauczyciel akademicki wobec nowych wyzwań edukacyjnych [Academic Teacher's new Educational Challenges], Uniwersytet Łódzki. pp. 25–43.

DZIECHCIARZ, J., (2015). Measurement of Rate of Return in Education. Research Directions, in: Velencei, J., (ed.), Business, Management and Economics, Obuda University, Budapest, pp. 39–56.

DZIECHCIARZ, J., (2015). O pojęciu jakości w pomiarze efektów pracy uniwersytetu [The Concept of Quality in the Measurement of the University Performance], Ekonometria, No. 4, pp. 79–92.

DZIECHCIARZ, J., (2015). O pomiarze efektywności nakładów na edukację i szkolenia w kontekście kształcenia przez całe życie [On the Measurement of the Efficiency of Investment in Education and Training in the Context of Lifelong Learning], in: Bernais, J., Jędralska, K., (eds.), Uniwersytet w perspektywie kształcenia przez całe życie [University in the Lifelong Learning Perspective], UE Katowice, pp. 42–52.

DZIECHCIARZ, J., BŁACZKOWSKA, A., GRZEŚKOWIAK, A., (2009). Econometric Evaluation of Education Systems; in: Rinderu, P., (ed.), Creating an Observatory on Europe-wide Transparency of Academic Qualifications, Editura Universitaria Craiova; Craiova, pp. 112–131.

DZIECHCIARZ, J., DZIECHCIARZ–DUDA, M., KRÓL, A., TARGASZEWSKA, M., (2015). Various Approaches to Measuring Effectiveness of Tertiary Education, Archives of Datascience, 2015/1, pp. 1–25.

DZIECHCIARZ, J., et al. (eds.), (2006). Rynek pracy aglomeracji wrocławskiej. Stan i perspektywy [The Labour Market of the Wroclaw Agglomeration. Status and Prospects], AE Wrocław.

DZIECHCIARZ–DUDA, M., DZIECHCIARZ, J., (2016). Multivariate Statistical Analysis in Missing Skills Identification, in: Michelberger, P., (ed.), Management, Enterprise and Benchmarking in the 21st Century, Óbuda University, Budapest, pp. 109–122.

DZIECHCIARZ–DUDA, M., KRÓL, A., (2012). Próba zastosowania modelu Mincera do oceny wpływu wyższego wykształcenia na poziom wynagrodzeń [An Application of Mincer Model in the Analysis of Higher Education Influence on the Wages' Level], Ekonometria, No. 3 (37), pp. 56–69.

DZIECHCIARZ–DUDA, M., KRÓL, A., (2013). On the non-monetary benefits of tertiary education, Econometrics, No. 3 (41), pp. 78–94.

DZIECHCIARZ–DUDA, M., PRZYBYSZ, K., (2011). Rynek usług edukacyjnych dla osób starszych. Analiza cech studentów uniwersytetów trzeciego wieku [Educational Market for Elderly People – Analysis of University of the Third Age Students' Characteristics], in: Garczarczyk, J., (ed.) Metody pomiaru i analizy rynku usług. Pomiar jakościowy. Zastosowania i efektywność [The Methods of Measurement and Analysis of Services Market. Qualitative Measurement. Application and Effectiveness], UE Poznań, pp. 44–55.

DZIECHCIARZ–DUDA, M., PRZYBYSZ, K., (2014). Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni [Education and Labour Market Needs. Classification of University Graduates], in: Jajuga, K., Walesiak, M., (eds.), Taksonomia, No. 22/327, pp. 303–312.

Getting Youth in the Door: Defining Soft Skills Requirements for Entry-level Service Sector Jobs (2013). International Youth Foundation, Baltimore.

GÓRNIAK, J., (ed.) (2015). (Nie)wykorzystany potencjał. Szanse i bariery na polskim rynku pracy [Opportunities and Barriers on the Polish Labour Market. (Not)Exploited Potential], PARP, Warszawa.

GREENACRE, M., (1984). Theory and Applications of Correspondence Analysis, Academic Press, London.

KURKLIŃSKI, L., MASZYBROCKI, M., (2008). Ocena kształcenia absolwentów studiów o kierunku ekonomia oraz finanse i rachunkowość a potrzeby rekrutacyjne instytucji finansowych [Economics, Finance and Accounting Graduates Qualification Evaluation versus Financial Institutions Recruitment Needs], Związek Banków Polskich, Warszawa.

MASZYBROCKI, M., (2010). Praktyki w bankach i zakładach ubezpieczeń dla studentów kierunków ekonomicznych (wyniki badań ankietowych) [Internships in Banks and Insurance Companies for Business Students (Survey Results)], Wiadomości Ubezpieczeniowe 1/2010, pp. 139–148.

ROUSSEEUW, P., STRUYF, A., HUBERT, M., (2015). Package 'Cluster', https://cran.r–project.org/web/packages/cluster.

STANIMIR, A., (2005). Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych [Correspondence Analysis as a Tool for the Economic Phenomena Investigation], AE Wrocław.

The Ten Unique Soft Skills Employers Desire in new Hires, (2015). http://www.entrepreneur.com/article/234864.

WALESIAK, M., DUDEK, A., (2009). Ocena wybranych procedur analizy skupień dla danych porządkowych [Finding Groups in Ordinal Data. An Examination of some Clustering Procedures], in: Jajuga, K., Walesiak, M. (eds.), Taksonomia, No. 16, pp. 41–49.

WALESIAK, M., DUDEK, A., (2015). Package 'ClusterSim', https://cran.r–project.org/web/packages/clusterSim.

WALESIAK, M., GATNAR, E., (eds.), (2009). Statystyczna analiza danych z wykorzystaniem programu R [Statistical Data Analysis using the Program R], PWN, Warszawa.

WDOWIŃSKI, P., (ed.) (2015). Nauczyciel akademicki wobec nowych wyzwań edukacyjnych [Academic Teacher's new Educational Challenges]; Uniwersytet Łódzki.

WHY ATTITUDE IS MORE IMPORTANT THAN INTELLIGENCE, (2015). http://www.entrepreneur.com/article/253095

ZABORSKI, A., PEŁKA, M., (2013). Geometrical presentation of preferences by using profit analysis and R program, Folia Oeconomica, 285, pp. 191–197.

APPENDIX

The classification of some of employees' soft competencies is further divided into subcategories.

Table 2. The subcategories of employees' soft competencies

Symbol	Description of soft competency subcategories
C01_1	quick summarizing large amounts of text
C01_2	logical thinking, analysis of facts
C01_3	continuous learning of new things
C03_1	perform simple calculations
C03_2	perform advanced mathematical calculations
C04_1	basic knowledge of MS Office package type
C04_2	knowledge of specialized programs, writing programs, Web pages
C04_3	use of the Internet
C07_1	independent decision making
C07_2	entrepreneurship, initiative
C07_3	Creativity
C07_4	resistance to stress
C07_5	timely implementation of planned activities
C08_1	cooperation within the group
C08_2	easy networking with colleagues
C08_3	being communicative
C08_4	resolving conflicts between people
C10_1	assigning tasks to other employees
C10_2	coordinating the work of other employees
C10_3	disciplining other employees
C11_1	frequent trips
C11_2	flexible working hours

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 737–748

INTERVAL ESTIMATION OF HIGHER ORDER QUANTILES. ANALYSIS OF ACCURACY OF SELECTED PROCEDURES

Dorota Pekasiewicz[1]

ABSTRACT

In the paper selected nonparametric and semiparametric estimation methods of higher orders quantiles are considered. The construction of nonparametric confidence intervals is based on order statistics of appropriate ranks from random samples or from generated bootstrap samples. Semiparametric bootstrap methods are characterized by double bootstrap simulations. The values of bootstrap sample below the prearranged threshold are generated by the empirical distribution and the values above this threshold are generated by the distribution based on the asymptotic properties of the tail of the random variable distribution. The results of the study allow one to draw conclusions about the effectiveness of the considered procedures and to compare these methods.

Key words: accuracy of estimation, order statistic, percentile bootstrap method, quantile, semiparametric bootstrap method, Value at Risk.

1. Introduction

Quantiles of a random variable distribution are used in different kinds of economic and financial research. They are applied in defining, for example, measures of poverty and wealth in the analysis of population income and Value at Risk measure in the studies of market risk. Value at Risk is defined as *p*-quantile of random variable being the value of losses from investments.

Nonparametric and semiparametric quantile estimation methods are the subject of interest when the quantile order is greater than 0.9. In the group of nonparametric procedures, bootstrap and non-bootstrap methods are considered. One of them is the percentile bootstrap method (Efron, Tibshirani 1993), and the other is the best exact nonparametric method (Zieliński, Zieliński 2005). Bootstrap semiparametric methods are based on information about the tail distribution of the random variable (Pandey et al. 2003). The accuracy of the

[1] Department of Statistical Methods, University of Łódź, Poland. E-mail:pekasiewicz@uni.lodz.pl.

estimation, defined as the length of the confidence interval, is analysed. The accuracy of quantile estimation for selected distributions, for the nonparametric non-bootstrap procedure of quantile estimation, known as the best exact method, is determined analytically. But for bootstrap methods the simulation study is used. In the paper Pareto and Student t-distribution are considered. The selection of distributions is associated with the possibility of choosing these parameters for which the distribution is characterized by a thin or fat tail. Simulation methods allow one to estimate the probability that the confidence interval includes the real value of the quantile, and additionally they allow to investigate whether this probability is approximately equal to the prearranged confidence coefficient. The application of the semiparametric methods require estimating the generalized Pareto distribution parameters. This distribution is used in approximation of the tail of the random variable distribution. In the paper, two methods of estimating the generalized Pareto distribution parameters are considered. One of them is the probability weighted moments method based on the classical empirical distribution and the other is the probability weighted moments method based on the level crossing empirical distribution (Huang, Brill 1999).

2. The best exact nonparametric estimation of quantile

Let us assume that we investigate a population with regard to random variable *X* with unknown continuous distribution *F*. Let $X_1, X_2, ..., X_n$ be a simple random sample drawn from this population and $1-\alpha$ be the fixed confidence coefficient.

Nonparametric interval estimation of quantile Q_p of order $p \in (0, 1)$ is associated with a random variable *K*, which denotes the number of observations in the sample smaller than the quantile Q_p. Random variable *K* has binomial distribution with the probability function:

$$P(K=k)=\binom{n}{k}(p)^k(1-p)^{n-k} \quad \text{for} \quad k = 0, 1,2, \ldots, n. \tag{1}$$

Let $X_{(r)}^{(n)}$ and $X_{(s)}^{(n)}$ denote order statistics of rank *r* and *s*, $(1 \le r \le s \le n, r, s \in N)$ respectively. The probability that the value of the quantile Q_p is in the interval $\left(X_{(r)}^{(n)}, X_{(s)}^{(n)}\right)$ is calculated by the formula:

$$P\left(X_{(r)}^{(n)} \le Q_p \le X_{(s)}^{(n)}\right)=\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}. \tag{2}$$

The values of order statistics are determined so that the right side of this formula is equal to the fixed confidence coefficient, i.e. $\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}=1-\alpha.$ Sometimes it results in an unequivocal confidence interval, especially for small sample sizes, when choosing ranks of order statistics to obtain the confidence interval for the *p*th quantile at the confidence level $1-\alpha$ is impossible. In these cases it is necessary to randomize (Zieliński 2008):

$$\lambda P\left(X_{(r)}^{(n)} \le Q_p \le X_{(s)}^{(n)}\right)+(1-\lambda)P\left(X_{(r')}^{(n)} \le Q_p \le X_{(s')}^{(n)}\right)=1-\alpha. \tag{3}$$

We calculate λ and we take $\left(X_{(r)}^{(n)}, X_{(s)}^{(n)}\right)$ as the confidence interval for Q_p with probability λ or $\left(X_{(r')}^{(n)}, X_{(s')}^{(n)}\right)$ with probability $1-\lambda$. The obtained interval is not always symmetric under the value of the quantile estimator.

The accuracy of the interval estimation is given by the formula:

$$d=\lambda\left(E\left(X_{(s)}^{(n)}\right)-EX_{(r)}^{(n)}\right)+(1-\lambda)\left(E\left(X_{(s')}^{(n)}\right)-EX_{(r')}^{(n)}\right), \tag{4}$$

where

$$E\left(X_{(k)}^{(n)}\right)=\int_{-\infty}^{\infty} x g_{k;n}(x)dx=\frac{n!}{(k-1)!(n-k)!}\int_{-\infty}^{\infty} x\left[F(x)\right]^{k-1}\left[1-F(x)\right]^{n-k} f(x)dx\ ,$$

$$k=r,s,r',s'. \tag{5}$$

Then, for the random variable with known distribution *F*, it is possible to calculate the precision of quantile estimation.

In Table 1 the minimum sample sizes for selected quantiles and different confidence coefficients are presented. In Table 2 are shown the ranks of order statistics allow one to obtain confidence interval for quantile on the confidence level approximately equals 0.95 for selected *p*-quantiles and selected sample. sizes.

Table 1. The minimum sample sizes for nonparametric estimation of *p*-quantiles

p	$1-\alpha$				
	0.9	0.925	0.95	0.975	0.99
0.80	11	12	14	15	21
0.90	22	25	29	23	44
0.95	45	51	59	72	90
0.99	230	258	299	368	459

Source: Own calculations.

Table 2. The ranks of order statistics in estimation of Q_p

p	Sample sizes					
	300		600		1000	
0.80	227, 254 (0.9491)	228, 256 (0.9515)	459, 498 (0.9495)	460, 499 (0.9527)	773, 823 (0.9479)	774, 824 (0.9506)
0.90	261, 281 (0.9451)	261, 282 (0.9524)	528, 560 (0.9499)	528, 561 (0.9509)	884, 923 (0.9494)	884, 924 (0.9514)
0.95	277, 292 (0.9491)	278, 293 (0.9548)	561, 582 (0.9467)	561, 583 (0.9512)	938, 965 (0.9474)	937, 964 (0.9504)
0.99	291, 300 (0.9499)	290, 300 (0.9507)	590, 599 (0.9412)	590, 600 (0.9558)	985, 998 (0.9495)	985, 999 (0.9517)

Source: Own calculations.

3. The percentile bootstrap method of quantile estimation

The next analysed procedure of quantile estimation is the percentile bootstrap method (Domański, Pruska 2000).

Based on the simple random sample $X_1, X_2, ..., X_n$ we generate N bootstrap samples $X_1^*, X_2^*, ..., X_n^*$, from the bootstrap distribution:

$$P\left(X^* = x_i\right) = \frac{1}{n}, \quad \text{for } i=1,2,\ldots,n, \tag{6}$$

where $x_1, x_2, ..., x_n$ are elements of the sample $X_1, X_2, ..., X_n$.

Next, for each bootstrap sample we compute the quantile $X_{p,k}^*$, where k =1, 2, …, N. Therefore, after N replications we get the sequence of ordered quantiles (sorted from least to greatest) $X_{p,(1)}^*, ..., X_{p,(N)}^*$, which allow one to approximate the distribution of quantile Q_p. Using this sequence we determine the percentiles of ranks $N\frac{\alpha}{2}$ and $N - N\frac{\alpha}{2}$.

The confidence bootstrap interval for Q_p has the following form:

$$P\left(X_{\frac{\alpha}{2}}^* < Q_p < X_{1-\frac{\alpha}{2}}^* \right) \approx 1-\alpha, \tag{7}$$

where statistics $X^*_{\frac{\alpha}{2}}$ and $X^*_{1-\frac{\alpha}{2}}$ are the percentiles of ranks $N\frac{\alpha}{2}$ and $N - N\frac{\alpha}{2}$, respectively.

The number of repetitions *N* is selected so as $\frac{N\alpha}{2}$ and $N - \frac{N\alpha}{2}$ are integers.

4. Semiparametric bootstrap methods of quantile estimation

Semiparametric bootstrap estimation methods are characterized by double bootstrap simulations, i.e. *n–k* values of bootstrap sample below the fixed threshold *u* are generated using empirical distribution F_n, but *k* values above this threshold are generated using the distribution which takes into account asymptotic properties of tail distribution (Pandey et al. 2003).

In this case the bootstrap distribution has the form:

$$F^*(x|u) = \begin{cases} (1 - F_n(u))F_0(x) + F_n(u), & \text{for} \quad x > u, \\ F_n(x), & \text{for} \quad x \le u, \end{cases} \tag{8}$$

where F_n is the empirical distribution and F_0 is the generalized Pareto distribution.

The generalized Pareto distribution $GPD(\xi, \beta)$ is expressed by the formula:

$$F_0(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}} & \text{for} \quad \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \text{for} \quad \xi = 0, \end{cases} \tag{9}$$

so the estimated distribution (8) has the following forms:

for $\xi \neq 0$:

$$\hat{F}(x|u) = 1 - \frac{k}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}} \quad \text{for } x > u, \tag{10}$$

and for $\xi = 0$:

$$\hat{F}(x|u) = 1 - \frac{k}{n}\exp\left(-\frac{x-u}{\hat{\beta}}\right) \tag{11}$$

where *k* is the number of elements of random sample greater than the fixed threshold and $\hat{\xi}, \hat{\beta}$ are estimators of parameters ξ, β.

Thus, *p*-quantile has the form:

$$X_{p;n} = \begin{cases} u + \dfrac{\hat{\beta}}{\hat{\xi}}\left(\left(\dfrac{n}{k}(1-p)\right)^{-\hat{\xi}} - 1\right) & \text{for } \hat{\xi} \neq 0, \\ u - \hat{\beta}\ln\left(\dfrac{n}{k}(1-p)\right) & \text{for } \hat{\xi} = 0. \end{cases} \tag{12}$$

The values of parameters $\hat{\xi}$ and β can be estimated by moments method, probability weighted moments method, or maximum likelihood method. The initial estimation of $\hat{\xi}$ can be obtained by generalized Hill estimator, moment estimator, Peng estimator or *W*–estimator (see Pekasiewicz 2015).

Two methods of estimation of parameters ξ, β are considered. In the first method, called semiparametric method I, parameters are estimated by probability weighted moments method (Landwehr, et al. 1979) and in the second one, called semiparametric method II – by modified probability weighted moments method, which is proposed in Pekasiewicz (2015).

In the method I, the estimators of parameters ξ, β have the following forms:

$$\hat{\xi}^{m(I)} = 2 - \frac{k\bar{Y}}{k\bar{Y} - 2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}}, \tag{13}$$

$$\hat{\beta}^{m(I)} = \frac{2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}\bar{Y}}{k\bar{Y} - 2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}}, \tag{14}$$

where $Y = X - u$.

In the semiparametric procedure with modified probability weighted moments method of estimation of $\hat{\xi}$ and β, the level crossing empirical distribution is used (Huang, Brill 1999):

$$F_k(y) = \begin{cases} 0 & \text{for } y < y_{(1)}^{(k)}, \\ \dfrac{1}{2}\left[1 - \dfrac{k-2}{\sqrt{k(k-1)}}\right] & \text{for } y_{(1)}^{(k)} \le y < y_{(2)}^{(k)}, \\ \dfrac{1}{2}\left[1 - \dfrac{k-2i}{\sqrt{k(k-1)}}\right] & \text{for } y_{(i)}^{(k)} \le y < y_{(i+1)}^{(k)},\ i = 2,3,...,k-1, \\ 1 & \text{for } y \ge y_{(k)}^{(k)}. \end{cases} \tag{15}$$

The estimators of parameters ξ, β are the following:

$$\hat{\xi}^{m(II)} = 2 - \frac{\bar{Y}}{\bar{Y} - 2\nu}, \tag{16}$$

$$\hat{\beta}^{m(II)} = \frac{2\nu\bar{Y}}{\bar{Y} - 2\nu}, \tag{17}$$

where

$$\nu = \frac{1}{k}\left(\frac{1}{2} + \frac{k-2}{2\sqrt{k(k-1)}}\right)Y_{(1)}^{(k)} + \frac{1}{k}\sum_{i=2}^{k-1}\left(\frac{1}{2} + \frac{k-2i}{2\sqrt{k(k-1)}}\right)Y_{(i)}^{(k)}. \tag{18}$$

5. Analyses of interval quantile estimation accuracy

The aim of the study is to compare the length of the confidence interval obtained by considered nonparametric and semiparametric methods. The accuracy of the best exact confidence interval is calculated by formula (4) and bootstrap procedures are analysed by simulation methods. The presented procedures are applied in the estimation of quantiles of orders higher than 0.9 for selected distributions.

The following distributions with fat tails are considered:

- Pareto $Pa(\theta, a)$, where $\theta, a > 0$,
- Student *t*-distribution S(*k*), where *k* is degrees of freedom.

Depending on parameters the distributions are characterized by the expected or non-expected value.

Quantiles of higher orders are estimated by the best exact nonparametric method, the percentile bootstrap method and two semiparametric bootstrap procedures (method I, method II). In the case of semiparametric bootstrap methods it is necessary to use information about the values from the tail distribution and the tail estimation using generalized Pareto distribution.

The construction of bootstrap nonparametric and semiparametric confidence intervals means that these methods can be studied and compared only using simulation analysis. The mean length of the confidence interval, and probability γ that the real value of the quantile is contained in the constructed interval are computed by repeating the estimation procedure 1000 times. This probability should be approximately equal to the predetermined confidence coefficient $1-\alpha$.

In the case of estimating higher order quantile, the changes in distribution parameters causes a significant change of the value of the quantile. The relationship between the values of selected quantiles and parameter *a* of Pareto

distribution is presented in Figure 1. In Figure 2 the relationship between selected quantile values and the degree of freedom of Student *t*-distribution is shown.

The results of the analysis of 0.99-quantile estimation for Pareto and Student *t*-distribution are presented in Table 3 and Table 4. The random samples must be rather big (in the tables - 1000 elements), which allows one to estimate generalized Pareto distribution parameters with small mean squared errors (the number of elements above the threshold is equals 100).

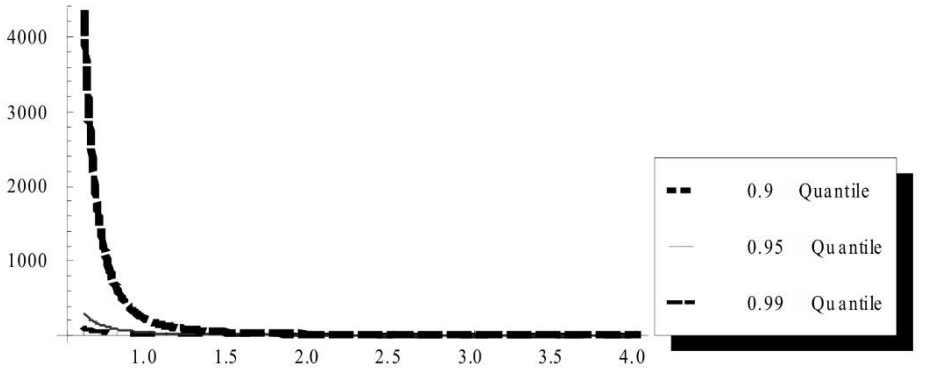


Figure 1. Relationship between higher order quantiles of Pareto distribution $Pa(2,a)$ and parameter a

Source: Own elaboration.

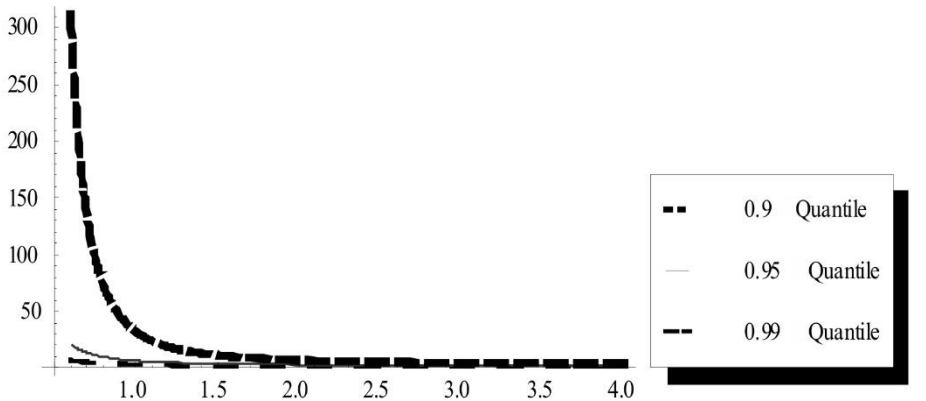


Figure 2. Relationship between higher order quantiles of Student *t*-distribution $S(k)$ and degree of freedom k

Source: Own elaboration.

Table 3. Accuracy d and probability γ of 0.99-quantile Pareto distribution estimation

Distribution	Best exact nonparametric method	Percentile method		Semiparametric method (I)		Semiparametric method (II)	
	d	d	γ	d	γ	d	γ
Pa(2, 1.25)	265.6520	86.7278	0.925	62.7672	0.921	61.8472	0.919
Pa(2, 1.5)	101.3190	37.0402	0.924	29.6164	0.926	30.0481	0.928
Pa(2, 1.75)	50.2623	20.0566	0.925	17.0014	0.946	16.9164	0.947
Pa(2, 2)	29.3498	12.6772	0.919	10.8831	0.926	12.6491	0.944
Pa(2, 2.25)	19.1083	8.7414	0.936	7.6362	0.943	7.6150	0.941
Pa(2, 2.5)	13.4312	6.3572	0.928	5.6160	0.942	5.7064	0.954
Pa(2, 2.75)	9.9807	4.8764	0.918	4.4066	0.956	4.4007	0.948
Pa(2, 3)	7.7489	3.8876	0.937	3.4652	0.949	3.5462	0.957
Pa(2, 3.25)	6.2154	3.1110	0.928	2.8326	0.954	2.9202	0.951
Pa(2, 3.5)	5.1188	2.6086	0,923	2.3731	0.949	2.4622	0.950
Pa(2, 3.75)	4.3070	2.2758	0.932	2,0576	0.945	2.0999	0.958
Pa(2, 4)	3.6884	1.9676	0.925	1.8047	0.953	1.8310	0.964

Source: Own calculation based on Mathematica 8.

The relative precision (the ratio of confidence length and the real value of quantile) in percentages is shown in Figure 3 and Figure 4.

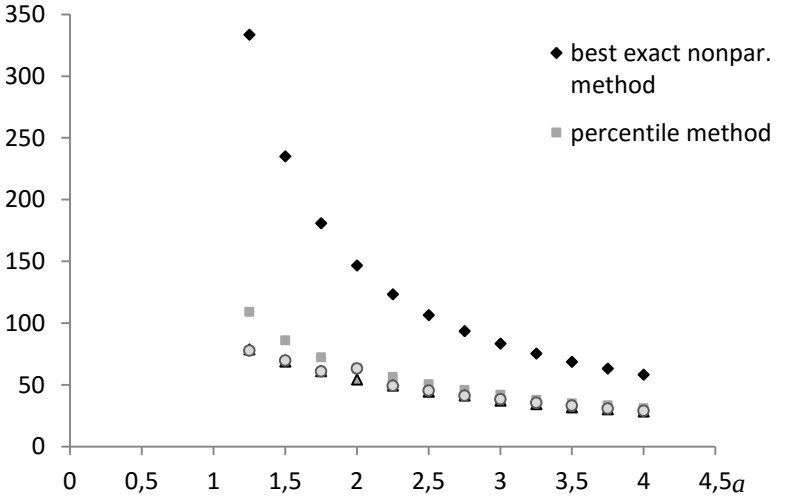


Figure 3. Relative precision of 0.99-quantile estimation for $Pa(2,a)$ distribution for considered methods

Source: Own calculation.

Table 4. Accuracy d and probability γ of 0.99-quantile Student t-distribution estimation

Distribution	Best exact nonparametric method	Percentile method		Semiparametric method (I)		Semiparametric method (II)	
	d	d	γ	d	γ	d	γ
$S(1.25)$	56.1500	18.8406	0.940	13.2522	0.916	13.0903	0.914
$S(1.5)$	26.4871	9.8648	0.921	7.9817	0.935	8.0148	0.940
$S(1.75)$	15.5615	6.3563	0.924	5.4952	0.940	5.5240	0.947
$S(2)$	10.4578	4.5764	0.910	4.0748	0.960	4.0828	0.957
$S(2.25)$	7.6799	3.4644	0.911	3.1792	0.945	3.2461	0.960
$S(2.5)$	5.9998	2.9216	0.926	2.6388	0.951	2.6384	0.953
$S(2.75)$	4.9027	2.4135	0.915	2.2014	0.950	2.2738	0.963
$S(3)$	4.1436	2.1148	0.940	1.9098	0.955	1.9402	0.948
$S(3.25)$	3.5942	1.8519	0.924	1.7452	0.959	1.7670	0.960
$S(3.5)$	3.1820	1.6880	0.932	1.5552	0.958	1.5813	0.954
$S(3.75)$	2.8635	1.5579	0.932	1.4434	0.963	1.4624	0.949
$S(4)$	2.6113	1.4072	0.921	1.3266	0.956	1.3604	0.961

Source: own calculation based on Mathematica 8.

The results of analysis imply that the application of bootstrap methods in estimation of quantile of Pareto distribution and Student t-distribution is more effective.

The choice of the quantile estimation method is important (see Figure 3 and 4) particularly for estimating the quantile heavy tailed distributions, i.e. $Pa(2, 1.25)$ or $S(1.25)$. It is associated with high values of distribution quantiles (see Figure 1 and 2).

In these cases the interval length obtained by the best exact nonparametric method (non-bootstrap procedure) is even three times longer than the bootstrap methods.

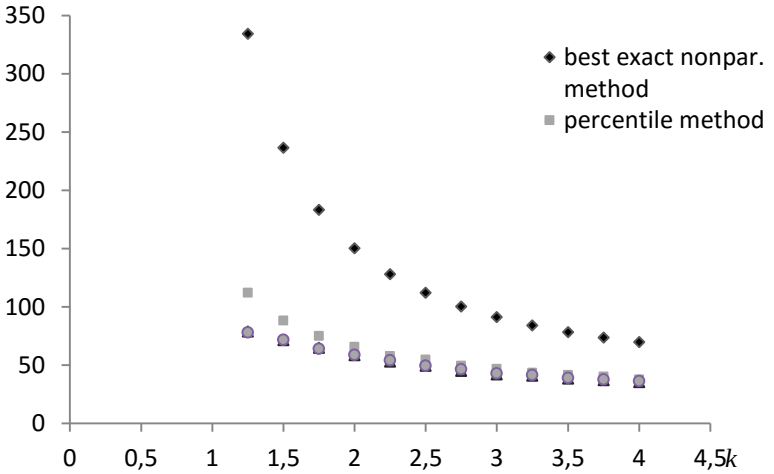


Figure 4. Relative precision of 0.99-quantile estimation for Student *t*-distribution for considered methods

Source: Own calculation.

5. Conclusions

In the paper different estimation procedures of higher order quantiles, including nonparametric and semiparametric methods, are considered. The application of bootstrap methods leads to confidence intervals which have smaller lengths than intervals derived from nonparametric, non-bootstrap methods. The best exact nonparametric confidence interval lengths are greater than the lengths of confidence interval obtained from the percentile bootstrap method. Semiparametric estimation methods allow one to get even shorter confidence intervals. Moreover, the probability that the confidence interval contains the real value of the distribution quantile is usually closer to the predetermined confidence level for semiparametric methods.

The generalized Pareto distribution parameters estimation method, which is used to approximate the tail distribution of the random variable, turns out to be less important in comparison with choosing the quantile estimation procedure.

The results obtained indicate that the choice of the estimation method is of greater importance when heavy tailed distribution quantiles are estimated.

The analysed procedures may be used to estimate measures based on higher order quantiles and may be applied in different economic and financial research.

REFERENCES

DOMAŃSKI, C., PRUSKA, K., (2000). Nieklasyczne metody statystyczne, [Non-classical Statistical Methods], Polskie Wydawnictwo Ekonomiczne, Warszawa.

EFRON, B., TIBSHIRANI, R. J., (1993), An Introduction to the Bootstrap, Chapman & Hall, New York.

HUANG, M. L., BRILL, P. H., (1999). A Level Crossing Quantile Estimation Method, Statistics & Probability Letters, 45, pp. 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), pp. 1055–1064.

PANDEY, M. D., VAN GELDER, P. H. A. J. M., VRIJLING, J. K., (2003). Bootstrap Simulations for Evaluating the Uncertainty Associated with Peaks-over-Threshold Estimates of Extreme Wind Velocity, Environmetrics, 14, pp. 27–43.

PEKASIEWICZ, D. (2015). Statystyki pozycyjne w procedurach estymacji i ich zastosowania w badaniach społeczno-ekonomicznych, [Order Statistics in Estimation Procedures and their Applications in Economic Research], Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

ZIELIŃSKI, R, ZIELIŃSKI, W., (2005). Best Exact Nonparametric Confidence Intervals for Quantiles, Statistics, 34, pp. 353–355.

ZIELIŃSKI, W., (2008). Przykład zastosowania dokładnego nieparametrycznego przedziału ufności dla VaR, [Example of Application of Exact Nonparametric Interval Confidence for VaR] Metody Ilościowe w Badaniach Ekonomicznych, 9, pp. 239–244.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 749–762

M-ESTIMATORS IN BUSINESS STATISTICS[1]

Grażyna Dehnel[2]

ABSTRACT

Recent years have seen a dynamic development in statistical methods for analysing data contaminated with outliers. One of the more important techniques that can deal with outlying observations is robust regression, which represents four decades of research. Until recently the implementation of robust regression methods, such as M-estimation or MM-estimation, was limited owing to their iterative nature. With advances in computing power and the growing availability of statistical packages, such as R and SAS, Stata, the applicability of robust regression methods has increased considerably.The aim of the study is to evaluate one of these methods, namely M-estimation, using data from a survey of small and medium-sized businesses. The comparison involves nine *M-estimators*, each based on a different weighting function. The results and conclusions are formulated on the basis of empirical data from the DG-1 business survey.

Key words: robust regression, M-estimation, business statistics, outliers

1. Introduction

Robust regression provides estimators which eliminate the influence of outliers. In the literature it is sometimes presented as a method designed to ignore outlying observations. It is often contrasted with methods aimed at detecting outliers. The fact is, however, that both detection and robust regression pursue the same objectives – the only difference is how they are achieved (Rousseeuw, Leroy, 1987). In the case of detection, the first step involves identifying outliers; only then are data corrected. In the case of robust regression, first a regression model is fitted to most of the data, then outliers can be detected, based on residual values.

While each of the two approaches has its benefits and drawbacks, it is the techniques of robust estimation that have been attracting growing interest recently. A number of approaches to robust regression have been proposed in an attempt to improve its performance. The starting point for the work was Ordinary Least Squares (OLS). One of its first modifications was *M*-estimation, which is

[1] The project is financed by the Polish National Science Centre, decision DEC- 2015/17/B/HS4/00905.
[2] Poznan University of Economics, Department of Statistics. E-mail: g.dehnel@ue.poznan.pl.

characterized by a low breakdown point by high efficiency. The next development was *S-estimation* and *LTS-estimation*, with a high breakdown point. The group of the latest methods includes *MM-estimation*. The *MM-estimator* was the first estimator with a high breakdown point and high efficiency under normal error (Stromberg, 1993). Some of these methods were developed in the 1970s and 80s of the 20th century, but because they rely on computationally demanding iterative procedures, they had limited applications. Nowadays a number of statistical packages are available, such as R or SAS, Stata, which facilitate the implementation of robust regression methods (Verardi, Croux, 2009). The growing interest in techniques of robust regression is also due to the fact that their application, unlike other methods, does not require earlier detection of outliers.

When choosing robust regression methods, one should keep in mind the properties of particular methods, which have been identified in the literature (Holland, Welsch 1977), (Huber, 1981), (Hampel *et al*., 1986), (Chen, Yin, 2002). One of the latest methods is *MM-estimation*. It is a combination of two different methods: efficient estimation of an *M-estimator* and *S-estimation* or a *LTS-estimation* with a high breakdown point. Hence, the ultimate quality of estimation depends on the quality of each of the two approaches. In each approach it is necessary to make additional decisions about the choice of parameters and functions. The present article is limited to an analysis of the properties of one of these approaches – *M-estimation*.

The study was aimed at evaluating properties of *M-estimators*, where different weighting functions were used. The evaluation was based on an empirical study using data on small and medium-sized enterprises in the transport section of the classification of economic activities (NACE Rev.2).

2. M-estimation

The class of *M-estimators* is a generalization of Maximum likelihood type estimators (MLE). *M-estimators* are classified as part of robust regression estimators of the so-called 1st generation. It is a group which is characterized by a low breakdown point in the case of *x-outliers*. The M-estimator was introduced by Huber in 1964 (Huber, 1964). It is a robust equivalent of the approach represented by the least squares method (Chen, 2007). The loss function of the least squares method is replaced by another loss function $\rho(\cdot)$, which is less sensitive to extreme residual values

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}(\theta)\right) \tag{1}$$

where: ρ - loss function

s - scale parameter

$r_i = y_i - X\theta$.

To ensure observed response values have comparable variation with respect to the values of the dependent variable, residuals are standardized using a dispersion measure *s*. The use of a classic measure of dispersion for the purposes of standardization, in the presence of outliers, results in overestimated values. For this reason, standard deviation is replaced with other measures of dispersion, such as median absolute deviation (MAD) or interquartile range (IQR).

The objective function meets the following conditions (Banaś, Ligas, 2014):

- non-negativity
- equals zero when its argument equals zero ($\rho(0)=0$)
- is symmetric (even function) ($\rho(r_i)=\rho(-r_i)$),
- monotonicity in $|r_i|$ ($\rho(r_i)\geq\rho(r_j)$ for $|r_i|>|r_j|$).

Assuming the scale parameter *s* is known, an estimate of the estimator θ_M is obtained by solving a system of *p* equations with respect to vector θ expressed as a product of independent variables and partial derivatives of the ρ function:

$$\sum_{i=1}^{n}\Psi\left(\frac{y_i-\sum_{k=1}^{p}x_{ik}\theta_k}{s}\right)x_i=0 \tag{2}$$

where: Ψ - influence function, a derivative of ρ function

p – the number of variables *x*.

In addition to the loss function, the influence function is another one that characterizes *M-estimators*. It helps to assess how a single observation affects the value of the estimator. Equation (2) is typically solved by means of *iteratively reweighted least squares* (IRLS) with weights given by the following formula (Trzpiot, 2013):

$$w_i=\Psi\left(\frac{y_i-\sum_{k=1}^{p}x_{ik}\theta_k}{s}\right)\Big/\left(\frac{y_i-\sum_{k=1}^{p}x_{ik}\theta_k}{s}\right) \tag{3}$$

where: w_i - weighting function.

The weighting function w_i, which is a ratio of the influence function and the residual, is the third function which characterizes *M-estimation*. The weighting function meets the following conditions (Banaś, Ligas, 2014):

- is continuous, symmetric,
- decreases when the residual increases,
- is equal to one when its argument is zero,
- decreases to zero for the argument increasing to +/- infinity.

The values of weights depend on what function Ψ is chosen to correspond with function ρ. In the literature of the subject many variations of *M-estimators*

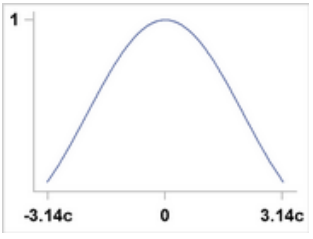
are suggested, with different variants of function Ψ (Fair, 1974), (Holland, Welsch 1977), (Huber, 1981), (Hampel *et al*., 1986), (Chen, Yin, 2002), (Banaś, Ligas, 2014). Curves of the weighting functions are different, but in *M-estimation* one always attempts to minimize or eliminate the influence of outliers. For this reason, all the proposed weighting functions cut off or reduce the influence of large residuals on the estimation of function parameters and/or scale parameter.

The selection of function Ψ is made depending on what weight we want to assign to outliers, among other things. In order to describe some of their properties, evaluation and usefulness, and their influence on estimation results, nine different weighting functions are analysed in this article: Andrews', Tukey's (bisquare), Cauchy's, Fair's, Hampel's, Huber's, logistic, Talworth's and Welsch's weighting functions (see Fig. 1).

Weighting functions have tuning factors, which can be modified. Using tuning factors, it is possible to reduce the impact of outliers with large residuals, but this is achieved at the cost of reducing the estimator efficiency. In the study described in this article, the tuning factors of the weighting function were set in such a way as to ensure 95% efficiency of estimates of M-estimators, see Table 1.

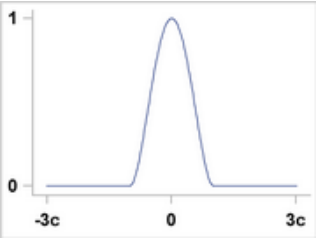
Andrews

$$W(x,c)=\begin{cases}\frac{\sin(\frac{x}{c})}{\frac{x}{c}} & \text{if } |x|\le \pi c\\ 0 & \text{otherwise}\end{cases}$$



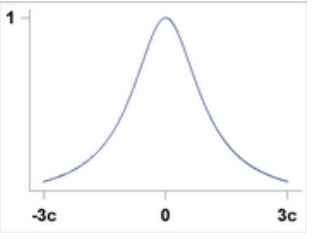
Tukey

$$W(x,c)=\begin{cases}(1-(\frac{x}{c})^2)^2 & \text{if } |x|<c\\ 0 & \text{otherwise}\end{cases}$$



Cauchy

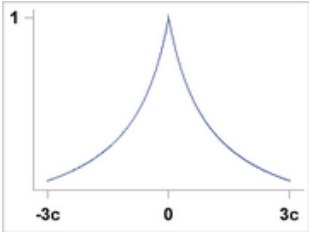
$$W(x,c)=\frac{1}{1+(\frac{|x|}{c})^2}$$



for c=2

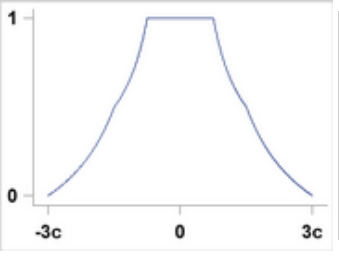
Fair

$$W(x,c)=\frac{1}{(1+\frac{|x|}{c})}$$



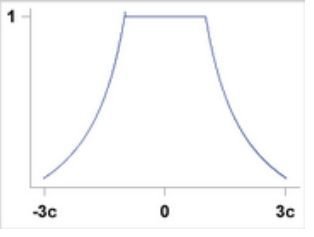
Hampel

$$W(x,a,b,c)=\begin{cases}1 & |x|<a\\ \frac{a}{|x|} & a<|x|\le b\\ \frac{a}{|x|}\frac{c-|x|}{c-b} & b<|x|\le c\\ 0 & \text{otherwise}\end{cases}$$



Huber

$$W(x,c)=\begin{cases}1 & \text{if } |x|<c\\ \frac{c}{|x|} & \text{otherwise}\end{cases}$$



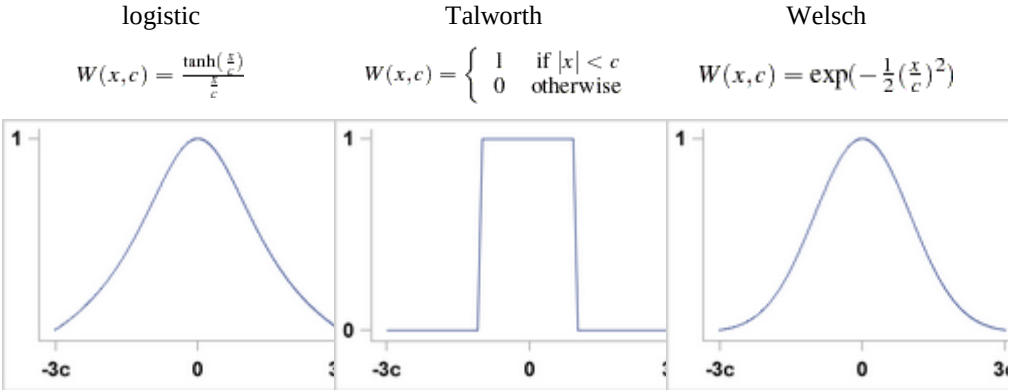


Figure 1. Weighting functions of *M-estimator*

Source: Based on SAS INSTITUTE INC. (2014).

The initial value of $\hat{\theta}_0$ is estimated based on the OLS method. In each iteration *t*, one uses values of residual and weights obtained in iteration *t-1* until convergence is achieved (Alma, 2011). After each iteration, it is also necessary to conduct standardization.

In practice the scale parameter *s* is unknown. One simple and very resistant possibility in these cases is to use the median absolute deviation estimator (Huber, 1964, Ripley, 2004; Trzpiot, 2013) Another possibility is to estimate scale *s* in an MLE-like way (Venables, Ripley, 2002).

Table 1. Tuning factors of the weighting function

Weighting function	Tuning factors a, b, c
Andrews'	1.339
Bisquare	4.685
Cauchy's	2.385
Fair's	1.4
Hampel's	4, 2, 8
Huber's	1.345
Logistic	1,205
Talworth's	2.795
Welsch's	2,985

Source: Based on SAS INSTITUTE INC. (2014).

The *M-estimator* is only resistant to outliers in the *y-direction*; it is not resistant to leverage points. This affects the range of potential applications. Hence, the estimator is frequently used but only in situations where leverage points are not a problem. Its breakdown point is not high and is equal to 1/n. The M-estimator is conditional bias – conditional on the proportion of the outlier in the sample (Cox *et al.*, 1995).

3. Evaluation of estimates obtained in the empirical study

A preliminary evaluation of estimates obtained using *M-estimators* was conducted in terms of the goodness of fit of the model, represented by the coefficient of determination. The robust version of the coefficient of determination is defined as:

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} \tag{4}$$

where ρ is the loss function for the robust estimate, $\hat{\mu}$ is the robust location estimator, and $\hat{s}$ is the robust scale estimator in the full model.

Properties of the estimators analysed in the study were evaluated using the bootstrap method. 1000 iterations of drawing samples were made, which were then used to calculate:

- Relative estimation error (REE)

$$CV\left(\hat{Y}_d\right) = \frac{\sqrt{Var\left(\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)} = \frac{\sqrt{\frac{1}{999}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d} - \hat{Y}_d\right)^2}}{E\left(\hat{Y}_d\right)} \tag{5}$$

- Mean absolute relative bias (ARB)

$$ARB\left(\hat{Y}_d\right) = \frac{1}{1000}\left|\sum_{b=1}^{1000}\frac{\hat{Y}_{b,d} - Y_d}{Y_d}\right| \tag{6}$$

- Relative root mean square error (RMSE)

$$RMSE\left(\hat{Y}_d\right) = \frac{\sqrt{\frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d} - Y_d\right)^2}}{Y_d} \tag{7}$$

4. The description of the study

The empirical study was based on information from official statistics collected in a business survey known as DG-1. It is the largest survey in Polish short-term

business statistics. It collects data from businesses employing over 9 people. The survey collects data from all medium-sized and large enterprises and a 10% sample of small businesses. It is conducted on a monthly basis. Its objective is to collect up-to-date information about basic indicators of economic activity of enterprises. In the empirical study only data about small and medium-sized companies were used (with the number of employees ranging from 10 to 250), which conducted their business activities in December 2011. In the model considered in the study revenue was the dependent variable. Independent variables came from an administrative register. Three independent variables were used in the model: profit, cost and the number of employees. A 10% sample of small and medium-sized companies from the DG-1 survey was treated as the general population. The domain of study was created by cross-classifying the administrative division into provinces with the NACE category of business activity. Results of the study were limited to domains included in one NACE section: *transport*. The section was chosen on the basis of the assessment of the goodness of fit of the regression model to the empirical data. The main motivation for the choice of the section was to ensure that domains it contained were characterized by the presence of outliers, which considerably reduced the quality of the classical model of regression (see. Table 2).

The first stage of the analysis involved assessing the distribution of businesses in terms of variables included in the model. Values of the basic descriptive statistics for all the variables were characterized by high variability and strong asymmetry. In the case of the variable 'Revenue', the coefficient of variation amounted to as much as 405%, while skewness was as high as 5.63.

Table 2. Statistical characteristics of the distribution of the *Revenue* variable (in thousand PLN by province and section 'Transport', 2011

Province	**CV**(%)	**Skewness**	$\mathbf{R^2}$	**Percentage of outliers** (%)	**N**
Dolnośląskie	90	1.24	0.537	7.1	28
Kujawsko-Pomorskie	100	1.45	0.945	16.7	24
Lubelskie	231	4.22	0.995	12.0	25
Lubuskie	303	3,99	0.139	15.0	20
Łódzkie	106	2.47	0.991	6.7	30
Małopolskie	327	5.62	0.999	11.8	34
Mazowieckie	405	5.63	0.999	6.8	73
Opolskie	89	0.91	0.984	28.6	14
Podkarpackie	72	0.27	0.982	18.8	16
Podlaskie	286	3.45	0.999	8.3	12
Pomorskie	138	2.38	0.980	3.0	33
Śląskie	261	5.55	0.992	9.4	64
Świętokrzyskie	143	2.52	0.998	13.0	23
Warmińsko-Mazurskie	65	0.76	0.892	15.4	13
Wielkopolskie	98	1.87	0.993	8.2	49
Zachodniopomorskie	138	2.21	0.970	13.3	30

Source: Own calculations based on DG1 survey.

In addition, Student's t-test and Cook's D confirmed the presence of outliers. These properties indicated the need for the use of robust regression method.

The percentage share of outliers, as well as the value of the coefficient of determination R^2 are presented in Table 2. The assessment of these two parameters indicates that the percentage share of outliers is not correlated with the value of the coefficient of determination. In the case of both more and less numerous sections, even a relatively large number of outliers does not necessarily have a negative impact on the model fit. On the other hand, individual outliers may have a large influence on the quality of the model, for the impact of outliers depends not only on their number but also on their type **(**outliers in the *x-direction*, outliers in the *y-direction*) and their distance from typical observations. The graphic presentation showing the relationship between the type of outliers and the model quality only shows domains with the lowest values of the coefficient of determination, that is for provinces of *Dolnośląskie, Lubuskie, Warmińsko-Mazurskie, Zachodniopomorskie* (see. Fig. 2)

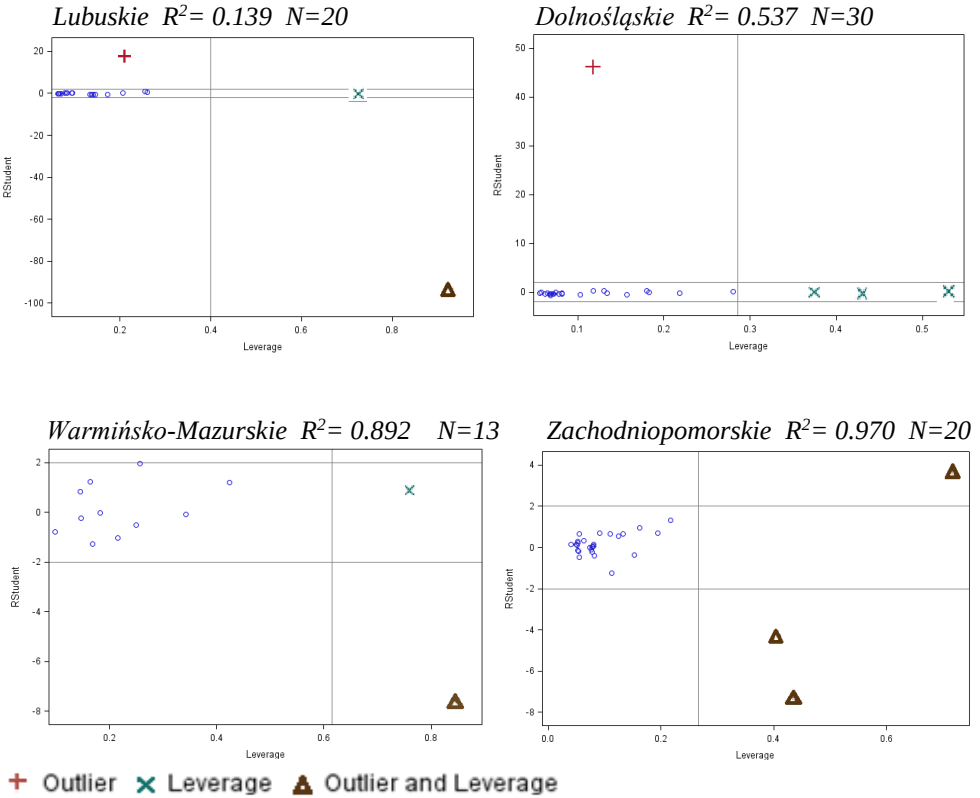


Figure 2. Outlier and Leverage diagnostic for *transport* in selected provinces

Source: Calculations based on the DG1 survey and the tax register of December 2011.

5. Empirical results of the study

The aim of the study was to compare the properties of *M-estimators*, in which different weighting functions were used. 9 types of *M-estimators* were analysed. The analysis was divided into two parts. The first part involved assessing the quality of the model's goodness of fit based on the robust version of the coefficient of determination and estimation errors of the equation parameters.

Values of the coefficient of determination are shown in Fig. 3. Differences in values obtained for each type of *M-estimator* reflect their sensitivity to the presence of different kinds of outliers (in the *x-direction* or in the *y-direction)* and their distance from the bulk of the data. The analysis of the results suggests that the use of *M-estimation* improves the goodness of fit of the model only when *y-outliers* are present. In the case of *x-outliers*, the application of *M-estimation* resulted in lower values of the coefficient of determination (compared to OLS), see Table 2 and Fig. 3.

In the domains analysed in the study, the highest values of the coefficient of determination were recorded for Fair's and Huber's functions, while the lowest ones for Cauchy's and Hampel's functions. Also, Talworth's and Tukey's functions are noteworthy. As can be seen from the results, the application of these functions in domains where the influence of outliers on the model quality is large results in a considerable improvement in efficiency and the robustness of *M-estimators*. This is due to the fact that they completely ignore observations for which large residuals were recorded.

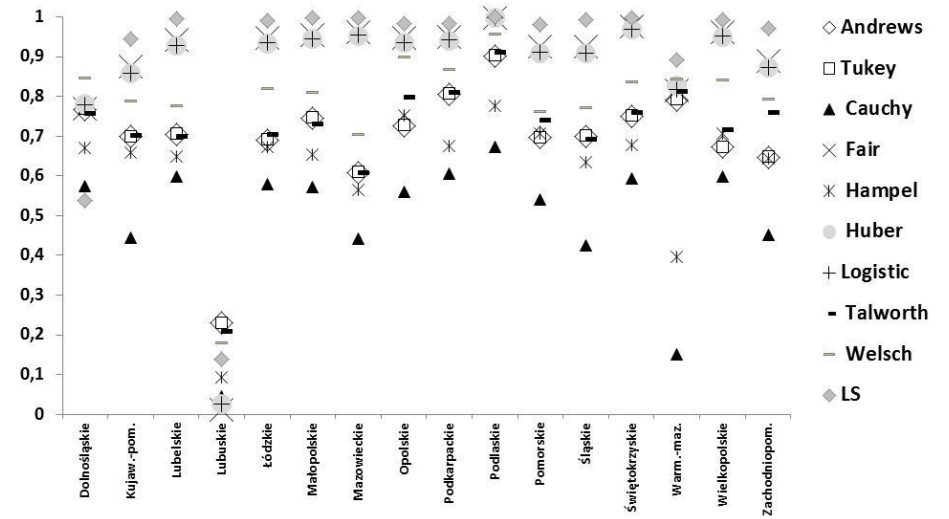


Figure 3. The coefficient of determination for the regression models of *Transport* in provinces

Source: Calculations based on the DG1 survey and the tax register of December 2011.

The study also investigated the values of model parameters and their estimation errors (standard errors) for the following variables: *profit*, *cost* and *the number of employees*, see Fig. 4. Both the estimated parameters and standard errors indicate high similarity of estimates in the domains of interest for all the weighting functions considered. The only case in which there were significant differences in estimates of the equation parameters (the slope) as well as the standard error was the number of employees in the province of Lubuskie. It is noteworthy that this situation applies to domains with the lowest values of the coefficient of determination for all kinds of the *M-estimator* ($R^2 \in (0{,}03;0{,}23)$). Additionally, in this domain there is a leverage point far removed from the bulk of the data, see Fig. 2.

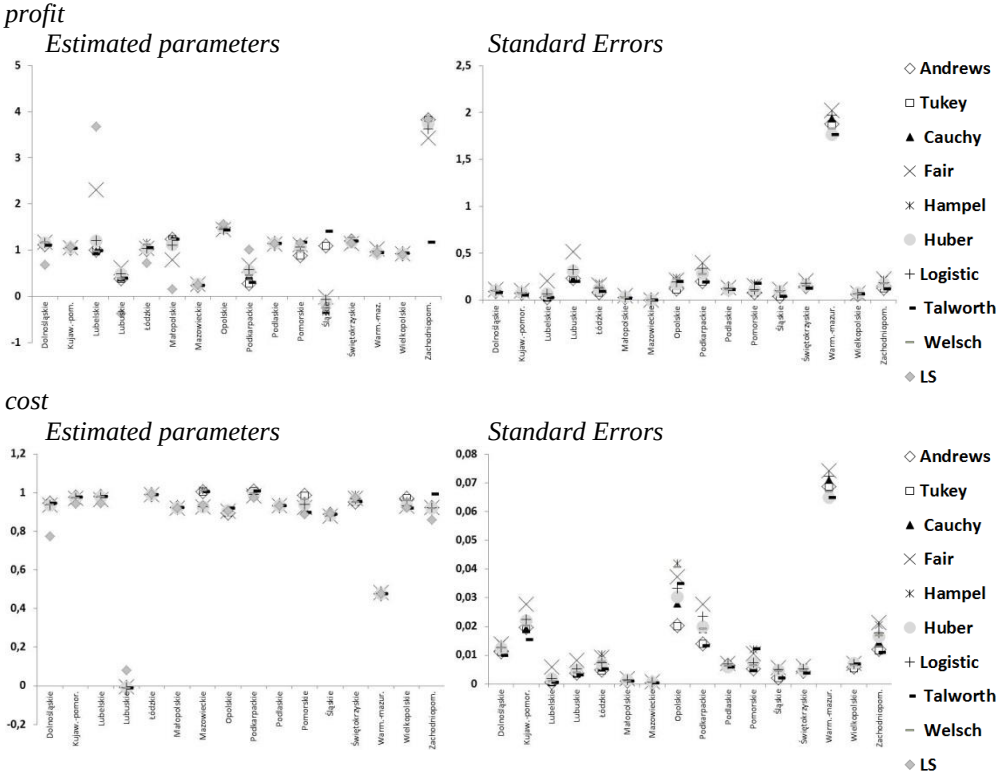


Figure 4. Estimated parameters and Standard Errors of the weighting functions for profit, cost and the number of employees

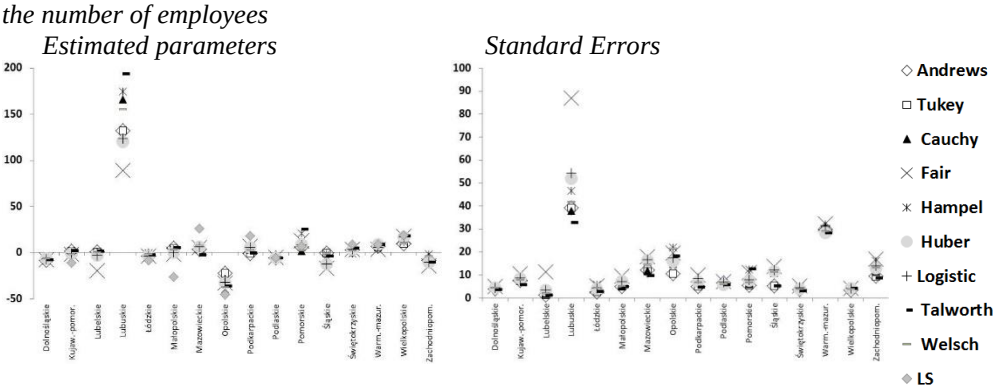


Figure 4. Estimated parameters and Standard Errors of the weighting functions for profit, cost and the number of employees (cont.)

Source: calculations based on the DG1 survey and the tax register of December 2011.

The second stage of the analysis consisted in comparing the properties of parameter estimators for the regression equations derived on the basis of the weighting functions. The bootstrap method was applied to determine measures for the assessment of the efficiency, bias and MSE, see Fig. 5. In the case of the variable *'The number of employees*' mean absolute relative bias (ARB) does not exceed 30%, while the relative estimation error is below 20%. This variable is least correlated with the variable of interest.

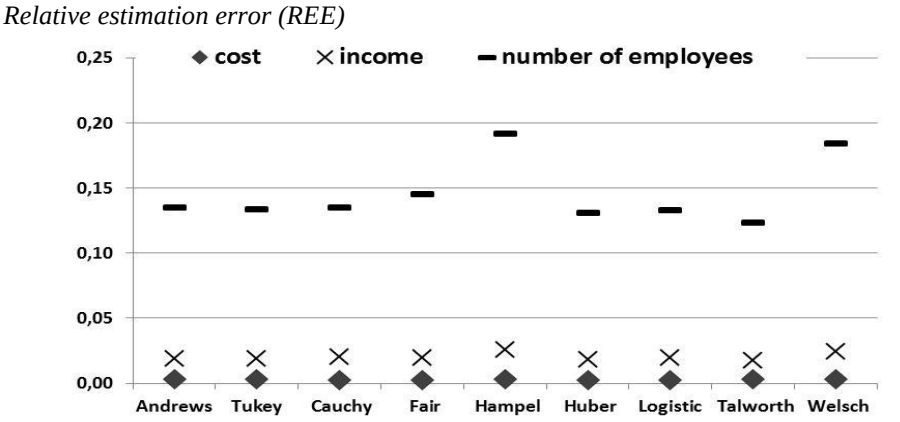
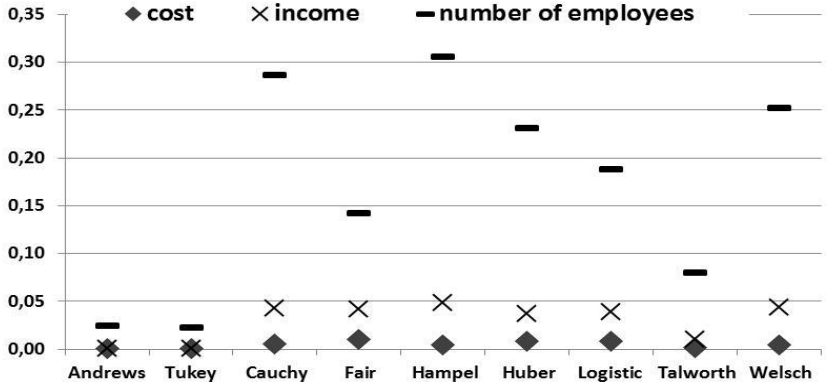


Figure 5. Performance criteria of estimates of equation parameters for profit, cost and the number of employees

Mean absolute relative bias (ARB)



Relative root mean square error (RMSE)

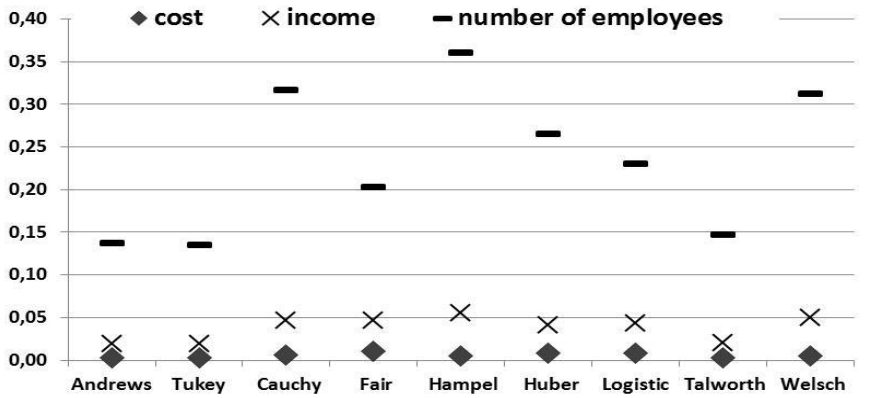


Figure 5. Performance criteria of estimates of equation parameters for profit, cost and the number of employees (cont.)

Source: Calculations based on the DG1 survey and the tax register of December 2011.

For the two remaining variables included in the model (*cost* and *profit*), the relationships between the values of characteristics of different types of *M-estimators* are very similar, both in terms of efficiency, and bias, which has a direct influence of MSE. The values of REE and ARB are close to zero. The amount of bias and efficiency for most cases is almost insignificant.

6. Conclusion

- The use of the *M-estimator* in the presence of outliers can considerably improve the quality of the model's fit compared to the classical method of estimation – it largely depends on the type of outliers. The *M-estimator* is only resistant to *y-outliers* but is not resistant to *leverage points*. It should therefore be used in situations where there are no *leverage points*.
- In practical applications of *M-estimation*, the selection of function Ψ is not a key choice for obtaining good robust estimates. The adoption of each of the nine weighting functions analysed in the study yielded similar results from the viewpoint of the values of estimated parameters and their standard errors. The least adequately fitted models were those based on Cauchy's and Hampel's functions. The best fit was obtained for the models based on Fair's and Huber's functions; one drawback in their case was the relatively high level of standard errors.
- The largest gain in efficiency and robustness of *M-estimators* was obtained when Talworth's and Tukey's functions were used. This result was particularly visible for domains in which the influence of outliers on the quality of the classical LS model was very strong. Owing to the curve shapes of Talworth's and Tukey's functions, observations with large residuals are ignored.

REFERENCES

ALMA, Ö. G., (2011). Comparison of Robust Regression Methods in Linear Regression, [in:] Int. J. Contemp. Math. Sciences, Vol. 6, No. 9, pp. 409–421.

BANAŚ, M., LIGAS, M., (2014). Empirical tests of performance of some M-estimators, Geodesy And Cartography, Vol. 63, No. 2, pp. 127–146.

CHEN, C., (2007). Robust Regression and Outlier Detection with the ROBUSTREG Procedure, SUGI, http://www2.sas.com/proceedings/sugi27/ pp.265–27.pdf.

CHEN, C., YIN, G., (2002). Computing the Efficiency and Tuning Constants for M-Estimation, Proceedings of the 2002 Joint Statistical Meetings, 478–482.

COX, B. G., BINDER, A., CHINNAPPA, N. B., CHRISTIANSON, A., COLLEDGE, M. J., KOTT, P. S., (1995). Business Survey Methods, John Wiley and Sons.

FAIR, R. C. (1974). On the robust estimation of econometric models, Ann. Econ. Social Measurement, 3, pp. 667–678.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. STAHEL, W. A., (1986). Robust Statistics: The Approach Based on Influence Functions, Wiley, New York.

HOLLAND, P., WELSCH, R., (1977). Robust Regression Using Interactively Reweighted Least-Squares, Commun. Statist. Theor. Meth., 6, 813–827.

HUBER, P. H., (1964). Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35, pp. 7–101.

HUBER, P. H., (1981). Robust Statistics, New York: John Wiley and Sons.

RIPLEY, B. D., (2004). Robust Statistics, M.Sc. in Applied Statistics MT2004, 1992-2004, https://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf.

ROUSSEEUW, P. J., LEROY, A. M., (1987). Robust Regression and Outlier Detection. Wiley-Interscience, New York.

SAS INSTITUTE INC., (2014). SAS/STAT® 13.2 User's Guide. The Robustreg Procedure Cary, NC: SAS Institute Inc.

STROMBERG, A. J., (1993). Computation of high breakdown nonlinear regression parameters, [in:] Journal of the American Statistical Association, 88 (421).

TRZPIOT, G., (2013). Wybrane statystyki odporne [Selected resistant statistics], [in:] Studia Ekonomiczne, No. 152, pp. 162–173, Uniwersytet Ekonomiczny w Katowicach.

VENABLES, W. N., RIPLEY, B. D., (2002). Modern Applied Statistics with S-PLUS. Springer-Verlag.

VERARDI, V., CROUX, C., (2009). Robust regression in Stata, [in:] The Stata Journal, 9, No. 3, pp. 439–453.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 763–780

INFORMATIVE VERSUS NON-INFORMATIVE PRIOR DISTRIBUTIONS AND THEIR IMPACT ON THE ACCURACY OF BAYESIAN INFERENCE

Wioletta Grzenda[1]

ABSTRACT

In this study the benefits arising from the use of the Bayesian approach to predictive modelling will be outlined and exemplified by a linear regression model and a logistic regression model. The impact of informative and non-informative prior on model accuracy will be examined and compared. The data from the Central Statistical Office of Poland describing unemployment in individual districts in Poland will be used. Markov Chain Monte Carlo methods (MCMC) will be employed in modelling.

Key words: Bayesian approach, regression models, a priori information, MCMC.

1. Introduction

For data mining techniques, classification and regression methods play an important role. The choice of an appropriate model is the basis of data analyses. The key advantage of the Bayesian approach is the ability to include additional information that is external to the sample in the modelling process (Lancaster, 2004). In Bayesian analysis, statistical inference is based on posterior distributions, which combine prior information with sample-based information. The impact of prior information on estimation model parameters in the parametric survival models has been investigated in (Grzenda, 2013), among others. In modelling, taking into account prior information has also an influence on the predictive power of a model.

The Bayesian model selection criteria frequently correspond to finding a model, which is characterised by a maximum posterior probability while considering model selection in the context of decision problems. The primary objective of this paper is to analyse the impact of prior information on the predictive power of a model using selected measures assessing the accuracy of prediction. Particular attention is paid to the selection of informative versus non-

[1] Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography. E-mail: wgrzend@sgh.waw.pl.

informative prior distribution. This is because the appropriate selection of a priori distribution may result in more accurate models.

Moreover, in this paper, the impact of informative and non-informative prior distributions on the accuracy of both classification and regression have been investigated. What should be emphasised in this context is that in Bayesian methods (Congdon, 2006; Gelman et al., 2000) parameters of a model are treated as random variables. Let θ denote the estimated parameter, and $\mathbf{x}$ observed data. The initial knowledge about the parameter θ is represented by prior distribution $p(\theta)$. The Bayesian inference approach is based on posterior distribution, which is determined in the following way (Bolstad, 2007):

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta) p(\theta)}{\int p(\mathbf{x} \mid \theta) p(\theta) d\theta}.$$

This equation is expressed in an equivalent proportional form:

$$p(\theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \theta) p(\theta).$$

In Bayesian approach, posterior distribution includes all available knowledge about the unknown parameter. This is prior information and information derived from data. The posterior distribution can be summarized by one statistic. Most frequently, this is the posterior mean as it minimizes a posterior mean square error. It is given by the formula:

$$E(\theta \mid \mathbf{x}) = \int \theta p(\theta \mid \mathbf{x}) d\theta.$$

Frequently, instead of a single parameter θ, the parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]^T$ is considered. The inference about any element of vector $\boldsymbol{\theta}$ is performed using marginal distribution, which is obtained by integrating the joint posterior distribution over the remaining coordinates. Given the complexity of calculations, the Markov chain Monte Carlo methods (MCMC) are used in practice (Congdon, 2006). The most famous algorithm among these methods is the Metropolis algorithm. In this paper the adaptive rejection Metropolis sampling algorithm (ARMS), which is a generalization of the Metropolis algorithm, has been used.

In this study, multiple regression models and logistic regression models have been estimated with informative and non-informative prior distributions. Based on obtained posterior distributions of model parameters, the posterior means have been calculated. These posterior means have been used as estimates for unknown parameters of the model (Lancaster, 2004). Next, the selected measures for model accuracy have been determined and compared. Many statistics can be used to measure the accuracy of models (Japkowicz and Shah, 2011; Provost and Fawcett, 2013). In the case of classification, the key measures are the incorrect

classification rate and confusion matrix, while in the case of regression the mean square error, median square error and maximum absolute error are frequently used. The predictive power of competing models can be compared based on Lift and ROC curves.

2. The scope of research

In this paper data from the Central Statistical Office of Poland, describing the districts in Poland, has been used. The object of this study is unemployment rate in districts in Poland in the year 2014. The characteristics of posterior distribution obtained for the previous year have been suggested as prior information for modelling data for the next year. Therefore, two sets of data have been created to model the unemployment rate in two successive years: 2013 and 2014. The number of observations in both data sets is the same, namely 380.

The examined feature is the unemployment rate in districts in Poland; in August 2013 the average was 15.92%, whereas in August of the next year the average was 14.32%. Moreover, for the purpose of this research, i.e. the investigation of classification accuracy, a binary variable *unemployment* has been created based on the continuous variable *unemployment_rate*. The variable *unemployment* differentiates the districts into those with low unemployment below 10% and the remaining ones. In 2013 there were 61 (16.05%) districts with the unemployment rate below 10%, whereas in 2014 there were 93 districts (24.47%). The unemployment may be defined and explored in many ways, but it is worth emphasising that a significant spatial diversity of the unemployment rate is observed in Poland (Gołata, 2004). The subject matter of this study is registered unemployment, including the unemployed registered in the district labour offices and seeking employment through these offices.

The preliminary data analysis including variable significance assessment, model adjustment to fit the observed data, model correctness verification and predictive power assessment (Lancaster, 2004) reduced the initially proposed set of variables to the following variables:

- salary – the amount of average monthly gross wages and salaries in thousand zlotys (mean=3.42, min=2.54, max=6.81);
- number_children – the number of children aged 3-5 per one place in nursery school (mean=1.48, min=0.77, max=5.11);
- flats – the number of flats ready for occupancy per 1000 residents (mean=2.92, min=0.16, max=15.26);
- EU_funds – the total value of contracts signed for financing in million zlotys per 1000 residents (mean=10.5, min=1.83, max=107.74);
- farm – the average area of individual farm (farms over 1 hectare have been investigated): 1 - less than 10 hectares (49.21%), 2 - from 10 to 15 hectares (30.79%), 3 - 15 hectares and over (20%);

- innovation – the average share of innovative companies in the total number of companies in %: 1 - less than 13% (35.26%), 2 - from 13 to 15% (49.21%), 3 - 15% and over (15.53%).

Variable characteristics for 2014 have been given in parentheses. The characteristics of the districts such as education, road infrastructure or population density, which were investigated in other studies such as (Gołata, 2004) have turned out to be statistically insignificant.

3. The multiple regression models

3.1. Bayesian multiple regression model

Let $\mathbf{y} = [y_1, \ldots, y_n]^T$ be the vector of observed values of the dependent variable $\mathbf{X}$, $(n \times k)$ be a matrix of independent variables, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$ be a vector of regression coefficients. The classical linear regression model can be expressed as follows (Draper and Smith, 1981):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ denotes an error vector, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$.

In Bayesian approach (Gelman et al., 2000; Gill, 2008), the regression coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$ are random variables. Let $p(\boldsymbol{\beta})$ denote their joint prior distribution and let us assume that the elements of vector $\boldsymbol{\beta}$ are independent. Then, we have the following the likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right].$$

Then, based on Bayes' theorem, posterior distribution is given by:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) \propto L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) p(\boldsymbol{\beta}) p(\sigma^2).$$

For model parameters, various prior distributions can be selected. The Bayesian approach with an informative prior allows us to incorporate additional information. If we do not have such information, then a non-informative prior can be selected. For regression coefficients $\boldsymbol{\beta}$, the most frequent normal prior distribution is selected: $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$. Assuming that the average equals 0 and there is a suitably small variation, a non-informative prior distribution is obtained: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^6 \mathbf{I})$. For the parameter σ^2, inverse gamma distribution is selected most frequently.

3.2. Key accuracy indicators of multiple regression models

While examining the accuracy of the model it is important to determine how the estimated values differ from actual values present in the training data. There are many ways to calculate the error, i.e. the difference between the estimated and actual values. The most natural one (Provost and Fawcett, 2013) is determining the absolute error:

$$AE = \left| y_i - \hat{y}_i \right| .$$

The maximum absolute error is a useful measure of prediction accuracy in the case of extreme values:

$$MAE = \max_i \left| y_i - \hat{y}_i \right| .$$

The sum of the squares error is a commonly used criterion for model accuracy. It is a natural consequence of estimating parameters of the classical regression model using least squares methods. This measure expresses the total value of the estimate error when the regression equation is used:

$$SSE = \sum_i \left(y_i - \hat{y}_i \right)^2 .$$

The degree of regression fit as an approximate linear relationship between the dependent variable and the explanatory is given by the coefficient of determination:

$$R^2 = \frac{SSR}{SST},$$

where $SSR = \sum_i \left(\hat{y}_i - \bar{y} \right)^2$ denotes the sum of squares regression and SST=SSR+SSE the sum of squares total, respectively. Finally, the mean squared error is defined as:

$$MSE = \frac{SSE}{n - m - 1},$$

where n is the number of observations, and m is the number of explanatory variables.

3.3. The specification and estimation of Bayesian multiple regression models

In this section, the multiple regression models with informative and non-informative prior distributions are discussed. In the first model developed for data from 2014, a priori distribution that has a minimal impact on posterior distribution has been used for all model parameters (Gelman et al., 2000). Therefore, non-informative independent normal prior distributions with mean equalling 0 and

variance 10^6 for regression coefficients, as well as inverse gamma distribution for the parameter σ^2, have been used. In all investigated models, the number of burn-in samples is assumed to be 2000 and posterior samples equals 10000 in order to minimize the effect of initial values on posterior inference. The highest posterior density (HPD) intervals for all parameters in all models have been determined for α=0.05. The characteristics of prior distributions and posterior distributions of the first model parameters for data from 2014 are presented in Table 1.

Table 1. The prior and posterior distributions

Parameter	Model 1					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	0	10^6	25.7691	2.5752	20.4681	30.5697
Salary	0	10^6	-2.8037	0.5352	-3.8700	-1.7685
Number_children	0	10^6	3.3589	0.4622	2.4530	4.2562
EU_funds	0	10^6	-0.1090	0.0286	-0.1643	-0.0521
Farm1	0	10^6	-3.7221	0.6291	-4.9466	-2.4826
Farm2	0	10^6	-5.2634	0.9361	-7.1080	-3.4612
Innovation1	0	10^6	-2.1867	0.8343	-3.8149	-0.5576
Innovation2	0	10^6	-2.9499	1.0464	-5.0837	-0.9786
Dispersion	IG		21.1553	1.5442	18.2595	24.2207

Based on the highest posterior density intervals (Bolstad, 2007), all variables are statistically significant for α=0.05. The convergence of generated Markov chain has been verified by several tests and graphically. The result of Geweke's test (Geweke, 1992) is included in Table 2. The graphs for generated chains are presented in the Figures 1-9. The results show no indication that the Markov chain has not converged for all the parameters of the investigated model at any significant level. Moreover, the Monte Carlo standard error (MCSE) is presented.

Table 2. Geweke convergence diagnostics and MCSE

Parameter	Model 1		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.6547	0.0980	0.0969
Salary	-0.8009	0.4232	0.0078
Number_children	-0.7595	0.4475	0.0052
EU_funds	-0.9591	0.3375	0.0003
Farm1	-0.3756	0.7072	0.0176
Farm2	-1.4492	0.1473	0.0469
Innovation1	-1.9345	0.0531	0.0399
Innovation2	-1.7811	0.0749	0.0568
Dispersion	1.8983	0.0577	0.0162

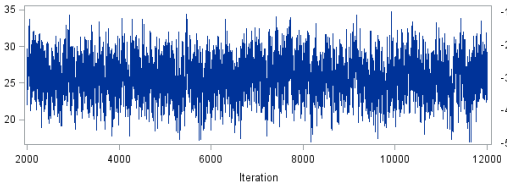


Figure 1. Trace Plots for *Inercept*

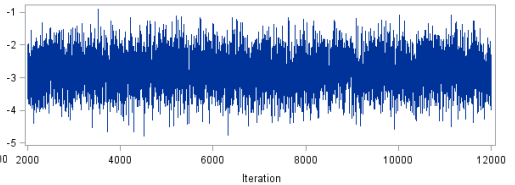


Figure 2. Trace Plots for *Salary*

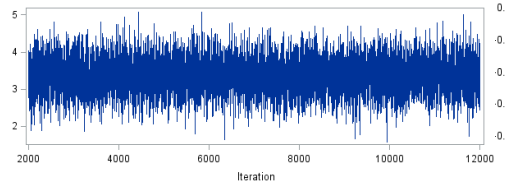


Figure 3. Trace Plots for *Number_children*

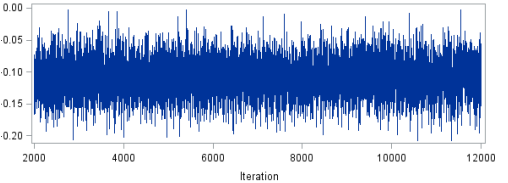


Figure 4. Trace Plots for *EU_funds*

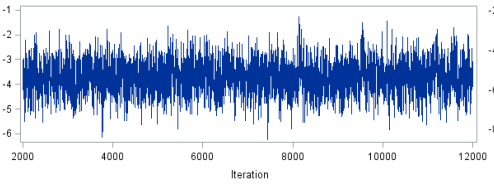


Figure 5. Trace Plots for *Farm1*

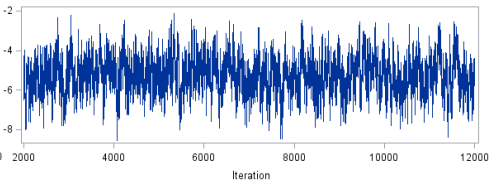


Figure 6. Trace Plots for *Farm2*

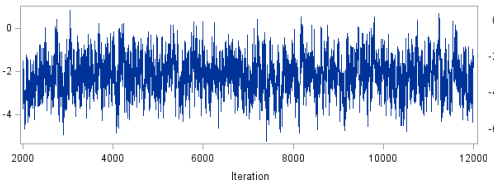


Figure 7. Trace Plots for *Innovation1*

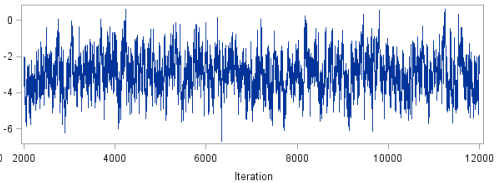


Figure 8. Trace Plots for *Innovation2*

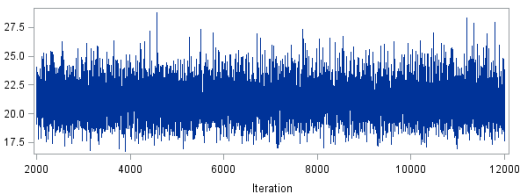


Figure 9. Trace Plots for *Dispersion*

Next, the multiple regression models for data from 2013 have been estimated. In order to obtain objectively correct results, non-informative prior distributions have been assumed in the model for data from 2013, as in the previous model. The obtained characteristics of posterior samples for data from 2013 have been used as prior information for the regression coefficients in the model for data from 2014. The prior and posterior distributions for second model parameters for data from 2014 are given in Table 3. All variables are again statistically significant for α=0.05.

Table 3. The prior and posterior distributions

Parameter	Model 2					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	26.8856	2.3281	25.5936	1.2719	23.2001	28.2022
Salary	-3.0471	0.5471	-2.9138	0.3281	-3.5662	-2.2840
Number_children	3.2935	0.4026	3.3003	0.2926	2.7320	3.8825
EU_funds	-0.1058	0.0283	-0.1099	0.0197	-0.1479	-0.0711
Farm1	-5.6123	0.6687	-4.5617	0.4342	-5.4003	-3.7021
Farm2	-4.6628	0.7196	-4.8026	0.4872	-5.7637	-3.8706
Innovation1	-2.7432	0.7040	-1.7124	0.4616	-2.6566	-0.8517
Innovation2	-0.3766	0.5933	-1.3999	0.4505	-2.2689	-0.5033
Dispersion	IG		21.2691	1.5815	18.2931	24.4650

The result of Geweke's test (Geweke, 1992) and the Monte Carlo standard error are included in Table 4. The results show no indication that the Markov chain has not converged for all the parameters of the investigated model at the significance level 0.01. The values of Monte Carlo standard errors for all model 2 parameters have been lower than in model 1.

Table 4. Geweke convergence diagnostics and MCSE

Parameter	Model 2		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	0.5981	0.5498	0.0154
Salary	0.1330	0.8942	0.0034
Number_children	-0.6970	0.4858	0.0029
EU_funds	-0.2500	0.8026	0.0002
Farm1	-2.1693	0.0301	0.0078
Farm2	-0.7126	0.4761	0.0095
Innovation1	0.4638	0.6428	0.0087
Innovation2	0.3302	0.7412	0.0096
Dispersion	1.6572	0.0975	0.0161

For both models, a deviance information criterion (DIC) has been calculated. For the first model DIC is 2245.575, for the second one it is 2244.557, the results do not differ significantly.

The estimated posterior means have been selected as estimation for the unknown model parameters for both models. With these assumptions, selected measures of model accuracy have been calculated (Table 5).

Table 5. Precision and accuracy of models

Statistics		**Model 1 Non-informative Priors**	**Model 2 Informative Priors**
Maximum Absolute Error	*MAE*	21.621	20.816
Sum of Squares Error	*SSE*	16242.310	15487.660
Sum of Squares Regression	*SSR*	13205.760	12639.100
Coefficient of Determination	R^2	0.448	0.449
Mean Squared Error	*MSE*	43.662	41.632

The obtained results indicate that the model with informative prior is the one with greater prediction accuracy. Moreover, these results indicate that about 45% of *unemployment_rate* variable variance is explained by the estimated models.

The estimated values of multiple regression models show that only the variable describing the number of children aged 3-5 per one place in nursery school has a positive impact on unemployment in the analysed districts. Thus, the greater the number of children per one place in nursery school, the higher the unemployment levels. The results also indicate that the lower salaries in districts, the higher unemployment. The unemployment rate in a given district also depends on the total value of EU contracts signed for financing – the smaller the value of grants, the higher unemployment. Moreover, the study found that the more fragmented farms and the lower the average share of innovative companies in the total number of companies, the higher the unemployment rates.

4. The logistic regression models

4.1. Bayesian logistic regression model

The logistic regression models (Finney, 1972; Hosmer and Lemeshow, 2000) are very often used in the study of socio-economic phenomena when a binary dependent variable is considered. These models are also applied to estimate the probability of belonging to a given class in classification tasks (Japkowicz and Shah, 2011).

Let us consider a dependent variable that takes only two values. Let $y_i = 1$ indicate the presence, and $y_i = 0$ the absence of the event, for $i = 1, \ldots, n$. Moreover, let p_i denote the probability that $y_i = 1$, $p_i = P(y_i = 1)$. Let $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ik}]^T$ be a vector of independent variables, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]$ be a vector of regression coefficients. Let $\text{logit}(p_i) = \boldsymbol{\beta}\mathbf{x}_i$, then the classical logistic regression model can be expressed as follows:

$$p_i = \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{x}_i)}.$$

The likelihood function over a data set for *n* subjects is:

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^{n} \left[(p_i)^{y_i} (1 - p_i)^{(1-y_i)} \right] = \prod_{i=1}^{n} \left[\left(\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{x}_i)} \right)^{y_i} \left(1 - \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{x}_i)} \right)^{(1-y_i)} \right]$$

In this paper, Bayesian logistic regression models are investigated (Albert and Chib, 1993; Congdon, 2006; Gelman et al., 2000). Assuming normal prior distribution $\beta_j \sim N(\mu_j, \sigma_j^2)$ for regression coefficients, and each of them being independent from the other, the posterior distribution is given by:

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{n} \left[\left(\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{x}_i)} \right)^{y_i} \left(1 - \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{x}_i)} \right)^{(1-y_i)} \right]$$

$$\cdot \prod_{j=0}^{k} \left[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2} \left(\frac{\beta_j - \mu_j}{\sigma_j} \right)^2 \right\} \right].$$

4.2. Key accuracy indicators of logistic regression models

The evaluation of the accuracy of a logistic regression model can be performed in many ways (Hosmer and Lebeshow, 2000). If the purpose of the modelling is to obtain the best possible classification (Provost and Fawcett, 2013), the following measures are the most common: the confusion matrix or classification table, the accuracy rate or interchangeably misclassification error rate. Graphically, the classification accuracy can be verified with ROC curve and LIFT curve (Japkowicz and Shah, 2011). Models with good classification capacity should be characterized by a high accuracy and a low rate of

misclassification. The results for the logistic regression model can be summarized in a classification table:

		Observed	
		POSITIVE	NEGATIVE
Predicted	YES	True positive	False positive
	NO	False negative	True negative

The basic measure for assessing the accuracy of the model in terms of classifying individual observations into the groups designated by the dependent variable is the accuracy of classification, i.e. the percentage of correct decisions:

$$Accuracy = \frac{Number\ of\ \ correct\ decisions\ made}{Total\ \ number\ of\ \ decisions\ made}$$

Alternatively, the misclassification error rate is calculated:

$$MISC = 1 - Accuracy$$

Based on the table such measures as sensitivity or true positive rate (TPR) and specificity or true negative rate (TNR) are often calculated:

$$TPR = \frac{true\ \ positive}{true\ \ positive + \ false\ negative},$$

$$TNR = \frac{true\ negative}{true\ negative + \ false\ \ positive}.$$

To determine the ROC curve, FPR (false positive rate) is calculated as 1-TNR. The ROC curve is formed by presenting FPR values on the axis *X*, and TPR values on the axis *Y*.

Model adjustment in terms of data and the prognostic effectiveness of competing models can also be compared using the LIFT curve. For a given model the LIFT curve compares the predictive model to no model (pick randomly):

$$\frac{True\ \ positive\ of\ \ Model}{True\ \ positive\ of\ no\ \ Model}.$$

4.3. The specification and estimation of Bayesian logistic regression models

Similarly to multiple regression models, Bayesian logistic regression models with non-informative and informative prior distributions were compared. The general assumptions regarding Bayesian estimation for logistic regression models were the same as in the case of multiple regression models. A model for the data from the year 2014 has been estimated, using non-informative normal prior distribution for all regression coefficients (Model 3). In Table 6, prior distribution settings and posterior distribution statistics for Model 3 are shown. For α=0.05, all variables are statistically significant except one level of *Farm* variable.

Table 6. The prior and posterior distributions

Parameter	Model 3					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	0	10^6	-6.9112	1.7665	-10.5488	-3.6320
Salary	0	10^6	1.2748	0.3379	0.6280	1.9384
Number_children	0	10^6	-1.9011	0.4606	-2.8083	-1.0316
Flats	0	10^6	0.2332	0.0856	0.0652	0.4013
EU_funds	0	10^6	0.0535	0.0196	0.0170	0.0934
Innovation1	0	10^6	1.4576	0.4885	0.5292	2.4134
Innovation2	0	10^6	2.3524	0.6803	1.0295	3.6720
Farm1	0	10^6	0.2196	0.4166	-0.5637	1.0672
Farm2	0	10^6	2.3576	0.6328	1.0999	3.5671

In Table 7, the results of Geweke's test (Geweke,1992) and the values of Monte Carlo standard error are shown. The study failed to reject the null hypothesis that the chains generated for individual model parameters converge at any level of significance. The figures depicting generated chains confirmed the inference regarding the convergence of these chains.

Table 7. Geweke convergence diagnostics and MCSE

Parameter	Model 3		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.8952	0.0581	0.0764
Salary	-1.2440	0.2135	0.0060
Number_children	-0.0992	0.9210	0.0067
Flats	-0.6526	0.5140	0.0011
EU_funds	-1.4819	0.1384	0.0002
Innovation1	-1.8530	0.0639	0.0244
Innovation2	-1.7862	0.0741	0.0424
Farm1	-1.0598	0.2892	0.0142
Farm2	-1.5770	0.1148	0.0381

Next, Bayesian logistic regression model has been estimated using informative prior distributions. The estimation was performed in the same way as for multiple regression models. The model is hereinafter referred to as Model 4. The results of the model estimation are provided in Table 8.

Table 8. The prior and posterior distributions

Parameter	Model 4					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	-5.3563	1.8043	-5.4033	0.9400	-7.2131	-3.5243
Salary	1.0686	0.3208	1.0654	0.2048	0.6661	1.4633
Number_children	-2.9115	0.6143	-2.2641	0.3597	-2.9848	-1.5786
Flats	0.3105	0.0949	0.2551	0.0629	0.1367	0.3831
EU_funds	0.0516	0.0198	0.0520	0.0138	0.0244	0.0782
Innovation1	1.8938	0.5578	1.3484	0.3230	0.7076	1.9705
Innovation2	1.2809	0.5354	1.7586	0.3657	1.0718	2.5072
Farm1	1.0792	0.6046	0.4214	0.3319	-0.2251	1.0632
Farm2	1.7633	0.6238	2.0049	0.3765	1.2680	2.7459

Table 9 provides the results of Geweke's test and MCSE values for Model 4. The results show that the study failed to reject the null hypothesis that the chains generated for individual model parameters converge at any level of significance. The MCSE values for all parameters of Model 4 are lower than the corresponding values for Model 3.

Table 9. Geweke convergence diagnostics and MCSE

Parameter	Model 3		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.8952	0.0581	0.0139
Salary	-1.2440	0.2135	0.0021
Number_children	-0.0992	0.9210	0.0049
Flats	-0.6526	0.5140	0.0007
EU_funds	-1.4819	0.1384	0.0001
Innovation1	-1.8530	0.0639	0.0076
Innovation2	-1.7862	0.0741	0.0111
Farm1	-1.0598	0.2892	0.0078
Farm2	-1.5770	0.1148	0.0107

For the third model, the DIC value equals 336.117, while the value of the same indicator for Model 4 is lower and equals 331.776. Therefore, Model 4 is a better model out of the two models.

The average values of the posterior distributions of the third and fourth model were used as the estimation for unknown model parameters. Next, the performance indicators for both models were calculated and compared. The lower misclassification error rate observed in the case of the model with informative prior distribution indicates that it is a model of higher accuracy (Table 10).

Table 10. Geweke convergence diagnostics and MCSE

Statistics		Model 3 Noninformative Priors	Model 4 Informative Priors
Accuracy Rate	*AR*	67.11	70.53
Misclassification Error Rate	*MICS*	32.89	29.47
True Positive Rate	*TPR*	0.495	0.516
True Negative Rate	*TNR*	0.728	0.766

ROC curves for models with informative and non-informative prior distributions are provided in Figure 10. This confirms that the model based on informative prior distribution is a model with better classification properties.

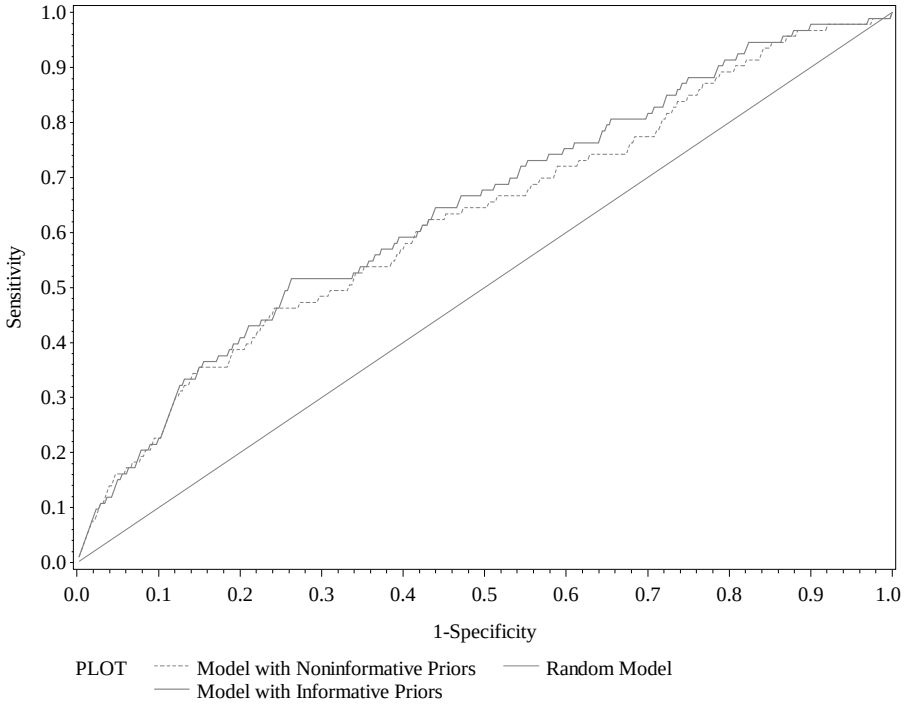


Figure 10. The ROC curve

LIFT curves for the model with non-informative and the model with informative prior distributions are shown in Fig. 11. The LIFT trend indicates that a model matches well the data (Tufféry, 2011). For every decile, the LIFT curve developed for the model based on informative prior distributions is located above the curve formed for a model created with non-informative prior distributions. Therefore, the model using informative prior distributions demonstrates better classification capabilities.

To sum up, all the analysed accuracy indicators show that the logistic regression model with informative prior distributions yields better classification capabilities.

Moreover, the estimation of logistic regression parameters shows that larger values of all the variables except for *number children* increase the probability of the unemployment rate in the district being below 10%. The higher the salaries, the bigger the number of flats ready for occupancy, and the larger the EU funds, the higher the chances of a low unemployment rate in the district. Moreover, the less fragmented farms and the bigger proportion of innovative enterprises, the higher the probability of the unemployment level in the district being below 10%.

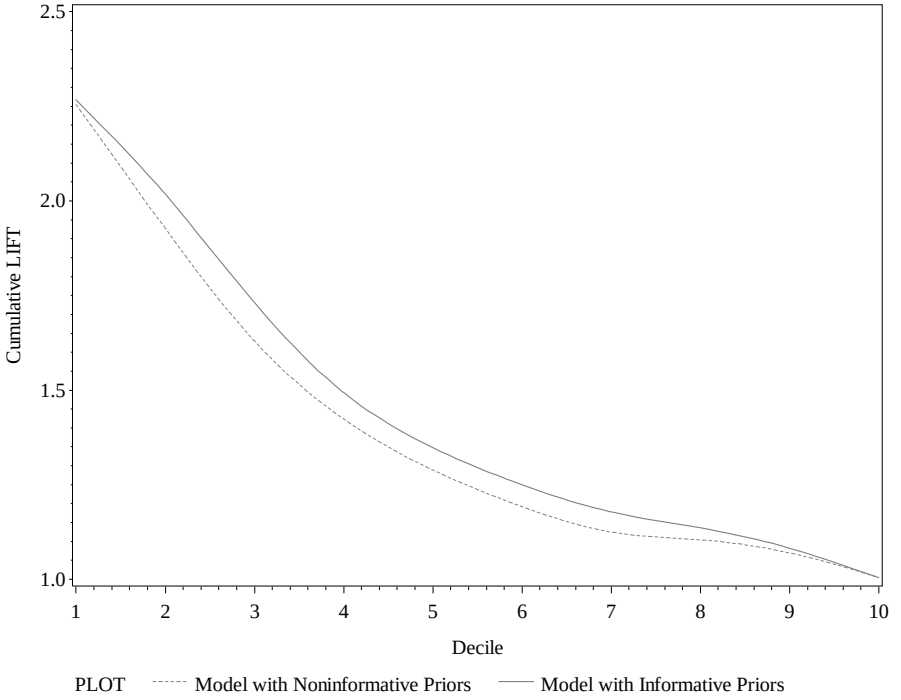


Figure 11. The LIFT curve

5. Summary

In this paper multiple regression models and logistic regression models have been investigated. Both categories of models are directly related and can be used for prediction, but they use different target variables. The primary objective of the study was to analyse how and to what extent prior information can influence the precision of regression and classification while using real data sets. First and foremost, the predictive analysis has been performed. This is because the outcomes of explanatory modelling cannot always be applied for predictive modelling (Provost and Fawcett, 2013).

To sum up, the predictive accuracy of models developed with non-informative and informative a priori distributions has been compared. The impact of prior information on the values of selected performance indicators developed for the models estimated with non-informative and informative a priori distributions has been shown. These results indicate that the accuracy of models estimated with informative a priori distributions is higher. Therefore, when additional out-of-sample knowledge is available, the appropriate selection of a priori distribution can improve the accuracy of regression and classification models.

REFERENCES

ALBERT, J. H., CHIB, S., (1993). Bayesian analysis of binary and polychotomos response data. Journal of the American Statistical Association, 88, 669–679.

BOLSTAD, W. M., (2007). Introduction to Bayesian statistics, USA: Wiley & Sons.

CONGDON, P., (2006). Bayesian Statistical Modelling, 2nd ed., UK: John Wiley & Sons Inc.

DRAPER, N., SMITH, H., (1981). Applied Regression Analysis, 2nd ed., New York: John Wiley & Sons.

FINNEY, D. J., (1972). Probit Analysis, London: Cambridge University Press.

GEWEKE, J., (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo J., Berger J., Dawiv A., Smith A. Bayesian Statistics, 4, 169–193.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B., (2000). Bayesian data analysis, London: Chapman & Hall/CRC.

GILL, J., (2008). Bayesian Methods, A Social and Behavioral Science Approach, USA: Chapman&Hall/CRC.

GOŁATA, E., (2004). Indirect Estimation of unemployment for the local labour market, Poznan: Publisher Academy of Economics in Poznan (in Polish).

GRZENDA, W., (2013). The significance of prior information in Bayesian parametric survival models. Acta Universitatis Lodziensis, Folia Oeconomica, 285, 31–39.

HOSMER, D. W., LEMESHOW, S., (2000). Applied Logistic Regression, New York: Wiley.

JAPKOWICZ, N., SHAH, M., (2011). Evaluating Learning Algorithms. A Classification Perspective, New York: Cambridge University Press.

KOOP, G., (2003). Bayesian Econometrics, Chichester, UK: Wiley.

LANCASTER, T., (2004). An Introduction to Modern Bayesian Econometrics, Oxford, UK: Blackwell Publishing.

PROVOST, F., FAWCETT, T., (2013). Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking, USA: O'Reilly Media, Inc.

TUFFÉRY, S., (2011). Data Mining and Statistics for Decision Making, Chichester, UK: Wiley.

VEHTARI, A., OJANEN, J., (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6, 142–228.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 781–787

EDWARD ROSSET (1897-1989) - THE NESTOR OF POLISH DEMOGRAPHERS AND STATISTICIANS

Czesław Domański[1]

On 26 of September 2015 a ceremony was held to mark 70th anniversary of foundation of the University of Lodz and 50th anniversary of establishing the Faculty of Economics and Sociology. On that occasion a commemorative plaque was unveiled to pay tribute Professor Edward Rosset - a Great Scholar and one of the Founding Father of the University of Łodz.

Professor Edward Rosset was born on 4 of November, 1897 in Łodz and died there on 2 June, 1989.

Childhood years of Edward Rosset fell to the time of a very intensive development of textile industry, where his father found employment selling textiles manufactured by local, privately-owned companies to the Russian markets. Rosset's mother, who studied the piano under Ignacy Paderewski, was a talented pianist, composer and graduate of the Academy of Music in Warsaw. The future scholar was growing up in a large family - he had five siblings (a brother and four sisters, all of whom died during the Nazi occupation of Poland). In the year 1916 he passed the examination for the secondary school certificate (Matura) in Philological Secondary School of Bogumił Braun. In the years 1917-1922 he completed his studies at the Faculty of Law and Political Sciences of Warsaw University.

After graduating in 1922 Edward Rosset, aged 25, returned to Łodz and took up his first post at the Department of Statistics of the Municipal Office. He was soon promoted, becoming the head of the Department, and he held the position until the end of 1940s, except for the period of war. Rosset made his professional debut at the Department of Statistics when he co-authored the publication entitled „Statistics of the City of Łodz 1918-1920". Edward Grabowski, the professor of statistics at the Polish Free University and Rosset's predecessor as the head of the Department, was the chief editor of the publication.

Published for the first time in 1922 and then for subsequent years, "Statistical Yearbook of the City of Lodz" was Edward Rosset's important accomplishment. It was a continuation of the earlier-mentioned „Statistics of the City of Lodz 1918-1920". Until the year 1929 yearbooks were comprehensive volumes which contained detailed descriptions of the most significant socio-economic and natural

[1] University of Lodz, Chair of Statistical Methods. E-mail: czedoman@uni.lodz.pl.

phenomena, presented in two languages – Polish and French. From 1930 onwards, a shorter version entitled „Little Statistical Yearbook" was appearing until the outbreak of the World War Two, as a consequence of the Great Economic Crisis. After the war three more volumes of "Statistical Yearbook of the City of Lodz" were published for the years 1945, 1946 and 1947.

Following his life motto „there is a problem hidden behind every number, it only needs to be discovered", Rosset did not limit his activities to editing source materials.

The problems he tackled in his works in the pre-war period could be classified into the following areas:

- manifestations of social pathology in big cities,
- living standards and health conditions of inhabitants of Łodz,
- political profile of inhabitants of Łodz.

A separate group of research areas followed directly from Edward Rosset's deep interest in the influence of war on population relations and processes, as well as his interest in the post-war revival of Estonia and other Baltic States.

One of the publications which deal with the problems of social pathologies is „Alcoholism in Łodz in the Light of Statistical Research". It was published in 1925 and its extended version „Alcoholism in Polish Cities". Similarly painful problems of big cities were discussed in the work published in 1931 and entitled „Prostitution and Venereal Diseases in Łodz", and an interesting fact was that it used police sources. Considering the problem of venereal diseases, Edward Rosset expressed the opinion that they reduced the reproductive capacity, and thus could become a significant depopulation factor.

It is definitely worth paying attention to three of his works published at the turn of 1920s and 1930s, viz. „Political Profile of the City of Łodz in the Light of Election Statistics" (1927), „Proletariat of Łodz in the Light of Demographic Research" (1930) and „Łodz in the Years 1860-1870. A Historical-Demographic outline" (1928). The first of the publications offers an analysis of statistics of the four consecutive parliamentary elections in the years 1906-1912, preceding the regaining of independence, and elections for City Councils, the Sejm and the Senate (the Lower and Upper Houses of Parliament) in independent Poland.

Undoubtedly, one of the highlights of Edward Rosset's scientific career was his participation in the international demographic congress, which was held in Rome in 1931. It was a pan-European meeting of demographers organized by the Italian Committee for Population Research and chaired by professor Corrado Gini. Edward Rosset presented two papers on this forum, i.e. „Demographic laws of war" and „Venereal diseases and war". Many years later the Author wrote: „The first of the two papers caused quite a stir during the Congress. The audience were impressed by the original approach I took, namely the generalization of demographic phenomena caused by war and formulation of rules based on them, which I called the demographic laws of war. Not less impressed by the book was the demographic and statistical literature all over the world". The paper in

question became for Edward Rosset a pass to the European demographic circles, and a clear evidence of his position was the fact that he became a member of the Italian Committee for Population Research.

As a practitioner-statistician, Edward Rosset was invited in the 1929/1930academic year to join a group of collaborators of the Łodz branch of the Polish Free University. Initially, he worked as an assistant lecturer giving lectures in statistics, demography and population policy, but shortly before the war Professor Zofia Daszyńska-Golińska submitted a request to the University authorities to appoint Rosset to the post of assistant professor.

After the outbreak of the Word War Two Edward Rosset lost his job, and was forced to leave his flat and home town. He moved illegally to Warsaw, with his wife Zofia and their two children – a daughter Irena and a son Stefan, and he spent the whole period of the Nazi occupation in a hideaway suffering extreme poverty. „We were living illegally"- he recalled after many years- „as we were refused by the occupier not only the right to honour and dignity but also the right to live. For all this time I was poring over the textbook. It was in the scientific activity that I found salvation from the threat of a nervous breakdown. Andrzej Grodek – a professor at the Main College of Trade, despite the fact that he was putting his personal safety at risk, granted me an access to the college library. I was set a limit – eight volumes per week - and I used this allowance eagerly. From the rich collection I was choosing methodically all those items which I found useful for my future university lectures". It is worth adding here that Edward Rosset could read fluently in five foreign languages (English, Russian, German, French and Italian). Although the notes he made at the time perished during the 1944 Warsaw Uprising, the knowledge the Professor acquired was stored up in his excellent memory.

The motivation for intensive, secret work despite the hardships of war can be found in another passage from Rosset's memories: „I had little hope to survive the war, but I said to myself if heavens allowed me to see the day of freedom then I had to be ready to take up my dream job of the professor of statistics at the University of Łodz".

Extensive reading of demographic, statistical and economic studies was not the only activity that the Scholar was engaged in during the years of occupation. He was also commissioned by the authorities of the Polish resistance to conduct an expert analysis (signed with a pseudonym Edward Nieżyński, engineer) which examined the occupational structure of the rural population on the territory to be incorporated into Poland after the war.

In the early 1945 Professor E. Rosset became seriously involved in organizational, didactic and scientific activities. The dreams which he had had in the gloomy days of the Nazi occupation finally came true. In 1940s he established two departments of Statistics – one was a unit of the newly-founded University of Lodz, the other was a part of the School of Economics. In 1961, when the two institutions merged, a Department of Demography and Statistics was established

as a part of the Faculty of Economics and Sociology. The Professor held the post of the head of the Department until his retirement in 1968.

The academic career of Professor Edward Rosset can be divided into the following stages: the position of an assistant in the Department of Statistics of the Polish Free University in the years 1929-1939, the post of an assistant professor of the University of Łodz, receiving a doctor's degree on the basis of the dissertation „The Demographic Laws of War", the position of adjunct professor at the Higher School of Economics and the University of Łodz in the year 1954, being awarded the title of associate professor in 1958, and full professor in 1963. In the Higher School of Economics the Professor held the following positions: vice-dean, dean, vice-Rector and the Rector. In the years 1961-1965 he was the vice-Rector of the University of Łodz.

The contribution made by Professor Rosset to the organization and development of demographic research in Poland is hard to be overestimated. He acted as a promoter of this research on behalf of the Polish Academy of Sciences, whose corresponding member he became in 1962, and the real member in 1976. In the period of 1978-1983 he held the post of the deputy chairman of the Lodz Branch of the Polish Academy of Sciences. He was the initiator of establishing the Committee for Demographic Sciences of the Polish Academy of Sciences, and he chaired it for several terms of office. Towards the end of his life he became the honorary chairman of the Committee. In 1963, thanks to the Professor's effort, the journal „Demographic Studies " was created and it became not only the organ of the Committee for Demographic Sciences but also the forum for publishing papers of Polish and foreign demographers. Edward Rosset was the editor-in-chief of the journal for 25 years.

The scientific output of Edward Rosset in the post-war period is extremely rich and varied. In contemporary demographic science there are practically no fundamental problems which were not examined by the Professor at some point of his research. The total number of all his published works (papers, monographs, reviews, reports, expert analyses) exceeded 300, including 16 books, and taking into account unpublished works we get an impressive number of 401 items.

His fundamental work "Ageing Process of Population", which was published in 1959, deals with the reasons and various consequences of the process of ageing societies. The work became internationally renowned soon after it came out and it was translated into foreign languages: English (Ageing Process of Population, 1964) and Russian (Process starienia nasielenia, 1968). The importance of the book is manifested in its numerous citations. Biographers and reviewers stress the fact that the everlasting value of the book consists in addressing various problems which future societies will be faced with, and will result from a constantly growing share of older people in human populations.

The problems discussed in „Ageing Process of Population" are developed in a comprehensive demographic study „Old people" (1969). Analytical elements included by the Author led him to the conclusion that the Polish society would soon reach the border of the old age with all the possible consequences of the fact:

social, medical and economic. „Life expectancy" (1979) is in a way connected with the problems of population ageing, as it provides an extensive analysis of reasons for an increase in life expectancy, which leads, in turn, to a growing share of the elderly in the society. On the other hand, the work can also be perceived as the one which deals indirectly with the problem of population reproduction.

Demographic processes and structures on a national level were analysed by Professor Rosset from the past – present – future perspective in several of his books. First of all, we need to mention here:„Demography of Poland" in two volumes (1975), „Demographic portrait of Poland" (1965), „Demographic prospects of Poland" (1962) and „Poland of the year 1985 – a demographic vision" (1965).

The author presented there some trends in population change and thoughtfully anticipated social and economic consequences of the ongoing demographic changes.

The process of population reproduction, which is the core issue of the second volume of „Demography of Poland", continued to be analysed - yet only with reference to marriage formation and breakdown - in the last great monograph published during the Professor's lifetime entitled "Divorces" (1986). Although the author does not totally reject divorce as a way of resolving extreme conflicts in marriage, in his book he appears to be an advocate of permanent relationships and he perceives the massive scale of divorce in many contemporary societies as a manifestation of social pathology. „Social evil does not cease to be evil just because it has become prolific" – he writes in the epilogue to the book, and goes on to say: „How can we assume that family is not in crisis when more and more people complain about their broken lives caused by marriage breakdown, and there are growing numbers of those whose bad experience of broken marriage makes them give up once and for all the idea of ever re-marrying, and when the number of "divorce orphans" is constantly growing. Obviously, the crisis does not mean that the institution of marriage is bound to become extinct. However, the threat of being extinct is real and thus it forces us to make more effort to strengthen the family so that the worst scenario does not happen".

Professor Rosset never forgot his home town, and its problems always remained within the scope of his interests. In 1962, on his initiative and under his editorial guidance an extensive monograph „Łódź in the years 1945-1960" was published. E. Rosset includes a chapter „Population relations" there, in which he provides an original estimation of population losses suffered by Łodz during the World War Two. Another monograph "Textile workers of Łodz", edited by E. Rosset and published in 1964, was also of regional character.

The theory of demography constitutes a separate area in the research work of Professor Rosset. The first signs of his interest in theoretical problems can be traced back in the earlier mentioned "Demographic Laws of War". A comprehensive monograph entitled „Demographic Explosion" (1978) gives the reader a chance to find a reflection of the Professor's interest in the theory of demographic transition. The idea is further developed in the work called "The

Theory of Demographic Transition - its Logic, Techniques and Prospects" (1987). The same research area is represented by the book „The Doctrine of Optimum Population in Historical Development" (1983). It presents the notion of population optimum in different periods of history and taking into account different criteria.

While studying the works of Edward Rosset one can only admire his great erudition, which is reflected in numerous references and aptly selected quotations enabling the reader not only to follow the author's ideas but also to get familiar with the world literature on the subject. The great merit of his writing is the exceptional beauty and precision of the Polish language. He also paid a lot of attention to clarity and logic of writing as he was of the opinion that the reader should be able to follow without difficulty even the most complex problems described by the author.

Apart from the aforementioned Demographic Congress in Rome in 1931, Professor Rosset participated in the following meetings: International Demographic Symposium in Smolenice (the present day Czech Republic, 1961), Budapest (1962), Zakopane (1964) and Liége (1973), among other things. He was also an organizer of the National Demographic Conference in Zakopane (1966).

Throughout his academic career – at the High School of Economics and the University of Łodz - and even after he had retired, Professor Rosset maintained close contacts with demographers from Eastern and Western Europe and was held in high esteem by his colleagues. Some of them were greatly honoured to be called his disciples. As a visiting professor he gave lectures at the Vienna University and universities of Belgrade, Berlin, Bucharest, Florence, Moscow, Prague, Pescu, Rome and Sofia.

In 1978 Edward Rosset received an honorary doctorate from the University of Lodz in recognition of his remarkable scientific achievements and work for his alma mater.

Numerous medals awarded to Edward Rosset throughout his long life are the best evidence of the appreciation he was given by both academic circles and the authorities. The Professor himself highly valued the Officer's Cross of Revival of Poland which he received in 1929 to commemorate the 10th anniversary of Poland's independence.

In the interwar period he was also honoured with Rother awards: Defender of Poland 1918-1921, Medal for Long-term Service (1928), Medal of Tenth Anniversary of Poland's Independence (1928) and the Estonian Officer's Cross of the Red Cross. After the war he was awarded with: The Golden Cross of Merit (twice: 1946, 1955), the Order of the Work Flag I Class (1976) and the Commander's Cross with the Star of the Order of Revival of Poland (1986).

REFERENCES

BARTKIEWICZ, Z., (1911). The Evil City.

DOMAŃSKI, CZ., (1991). Edward Rosset (1897-1989), Przegląd Statystyczny, No. 2.

FLATT, O., (1853). The Description of the City of Łódz, Warszawa, pp. 149.

KOWALESKI, J. T., OBRANIAK, W., (1997). Professor Edward Rosset, [in:] Profiles of the Statisticians of Łódz, No. 37, Łódzkie Towarzystwo Naukowe, Łódź, pp. 7–17.

KOWALESKI, J. T., OBRANIAK, W., (2012). Rosset Edward (1897-1989), [in:] Polish Statisticians, GUS, PTS Warszawa, pp. 304–313.

REYMONT, W., (1899). The Promised Land, Nakład Gebethnera i Wolffa, Warszawa.

ROSSET, E., (1962). Łódź in the years 1945-1960, rozdział „Population relations".

OKÓLSKI, M., WELFE, W., (1987). Edward Rosset – the Scholar of the Past Generation, the Modern Man, Ekonomista, No. 1, pp. 23–41.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 789—792

REPORT

Quality of Life and Spatial Cohesion: Interaction of Development and Well-being in the Local Context (QLiSC), 17-18 November, Warsaw, Poland

The conference organized by the Central Statistical Office of Poland (CSO) and the Cardinal Stefan Wyszynski University in Warsaw (CSWU) was held on 17-18 November at the CSWU campus. The conference was sponsored by the European Union Cohesion Fund (as a part of the EU Technical Assistance Operational Programme 2014-2020).

It was opened by the CSWU Rector, **Rev. Prof. Dr. Stanisław Dziekoński**, whose welcome speech was followed by those from the members of the Honorary Committee: **Dr. Grazyna Marciniak** – Vice-President of the Central Statistical Office (CSO) of Poland, on behalf of the CSO President, **Dr. Dominik Rozkrut**; **Rev. Archbishop Henryk Hoser** SAC (Archdiocese of Warsaw-Praga); **and Rev. Prof. Dr. Slawomir Zareba** – Dean of the History and Social Science Faculty of the CSWU; and by **Prof. Dr. Wlodzimierz Okrasa** – head of the conference Scientific Organizing Committee (and Moderator of the conference).

The conference was devoted to the review of the current research experience, along with the conceptual, measurement, and practical issues concerning relationships between the quality of life and well-being (of individuals, families and households) and the development of territorial units (municipalities or communes, like *gmina* in Poland, or regions, etc.), and to exchanging ideas for a possible future research agenda.

Two key note speakers, **Prof. Dr. Graham Kalton** (Westat and the University of Maryland, USA) and **Prof. Dr. Filomena Maggino** (University of Florence, Italy) overviewed some of the frontier issues from methodological and conceptual standpoints, respectively: G. Kalton talked about *Estimating Quality of Life and Wellbeing at the Local Level – The Role for Small Area Estimation;* F. Maggino about *Quality of Life and Social Cohesion: Defining and Measuring Subjective Aspects.*

As it was expected, a wide range of papers on substantive and methodological problems approached from the multidisciplinary perspective was presented at the conference, with reference often made to the 'locality' and space.

The first (plenary) session - *Spatial aspects of wellbeing and cohesion* - chaired by **Graham Kalton** was composed of the following presentations:

- Włodzimierz Okrasa: *Community Cohesion and Individual Wellbeing - multilevel spatial approach;*
- Jan Kordos: *Small Area Estimation for the Quality of Life and Spatial Cohesion Indicators;*
- Zhanjun Xing and Quan Zhang: *Spatial cohesion and subjective well-being of the elderly;*
- Krzysztof Zagórski: *Individual and regional relative income and 'life satisfaction'.*

As a session's satellite event, a workshop „What we do not know about statistics?" was organized by directors of departments at the CSO, **Dominika Rogalinska and Renata Bielak**. It was envisioned as a way to familiarize the CSWU students with modern tools and forms of using statistical information in social research - the workshop was well attended.

The two parallel sessions were devoted to: *Objective and subjective dimensions of the spatial characteristics of quality of life* and *Welfare / wellbeing and local cohesion*

During the first one, chaired by **Czesław Domański**, the following papers were presented:

- Elżbieta Bojanowska: *The development of social services in the local environment from the perspective of the state social policy;*
- Tomasz Panek: *Dilemmas of measuring quality of life;*
- Przemysław Śleszyński, Jerzy Bański, Marek Degórski and Tomasz Komornicki: *Delimitation of the natural, social and economic problem areas in Poland;*
- Anna Szukiełojć-Bieńkuńska, Tomasz Piasecki and Karol Sobestjański: *Territorial differentiation of subjective wellbeing in Poland.*

The second session, chaired by **Krzysztof Wielecki**, encompassed presentations:

- Artur Czech and Teresa Słaby: *Spatial cohesion of Polish households – meanders of taxonomic analysis;*
- Anna Barwińska-Małajowicz and Bogusław Ślusarczyk: *The functioning of the labor market and the quality of life at the local level;*
- Marcin Szymkowiak and Łukasz Wawrowski: *The measurement of poverty at the low levels of spatial aggregation;*
- Marek Degórski: *Quality of life and the provision of ecosystem.*

The second day of the conference started with a lecture of **Janusz Czapiński**, an invited speaker, who talked about *Subjective wellbeing – interpretations in theory and empirical research.*

The following two sessions focused on different aspects of inequalities. The session on *Spatial inequalities of quality of life* was chaired by **Filomena Maggino** – it included five presentations:

- Czesław Domański and Alina Jędrzejczak: *Measurement of income inequality at the sub-national level using the SAE techniques;*
- Semen Matkovskyy and Svitlana Zimovina: *Quality of life in transborder areas – health and material conditions of life;*
- Daniel Skobla: *Exploring the living conditions of the Roma population in Slovakia: the atlas of Roma communities;*
- Marek Cierpiał-Wolan: *The well-being paradox in transborder areas;*
- Second Bwanakare: *Cross-Entropy Econometrics and Income Redistribution Trough a Social Accounting Matrix: The Polish Case Study*.

Another session, *Factors of differentiations household – community – subregion*, chaired by **Marek Cierpiał-Wolan** included the following presentations:

- Edyta Mazurek: *The impact of tax-free amount on the financial situation of households and the state budget;*
- Rev. Wojciech Sadłoń: *Quality of life and religion;*
- Tomasz Kasprzak: *The influence of Euroregions on local development – the case of the Euroregion Cieszyn Silesia;*
- Andrzej Młodak and Tomasz Józefowski: *Taxonomic poverty measures for subregions;*
- Tomasz Komornicki: *The improvement of spatial accessibility and quality of life.*

The conference was concluded with Panel of experts, with the participation of Wlodzimierz Okrasa as a moderator, and Graham Kalton, Filomena Maggino, Zhanjung Xing, Grazyna Marciniak, and with active involvement of the audience.

In addition to regular sessions, a poster session was also organized, with high quality poster presentations by:

- Grzegorz Gudaszewski, Dariusz Winkler: *The development characteristics of communes (gminas) and the spatial distribution of the beneficiaries of the program "500-plus";*
- Marta Kusterka-Jefmańska: *Quality of life and sustainable urban development;*
- Anna Pluskota: *Empowerment and wellbeing of local communities – studies of dependency;*
- Krzysztof Błoński and Bartłomiej Jefmański: *The differentiations of the quality of life in the Wałcz county;*
- Cyprian Kozyra: *Application in Poland of convergence and divergence models of the local tax charges incurred by the residents of the local governance units;*

- Przemysław Śleszyński: *Uncontrolled urbanization and its socio-economic costs;*
- Piotr Zawada: *Wellbeing of the middle class in local community – the case of Strugu Valley.*

Out of several types of conference results – in terms of both substantive and methodological achievements - one that seems to be worthwhile mentioning is the confirmation of the underlying assumption that spatial aspects of quality of life and local development provide a convenient platform for multi- and interdisciplinary policy research, which is especially promising from the evidence-based program design and evaluation policy-making perspective.

Prepared by
Włodzimierz Okrasa

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 793–796

ABOUT THE AUTHORS

Bhattacharjee Atanu was born in Durgapur, India, to Swapan Bhattacharjee and Sandhya Bhattacharjee. He attended A-Zone MP School and T.D.B. College before joining the Institute of Medical Science, Banaras Hindu University, where he obtained master's degrees in health statistics in 2008. He obtained his PhD in statistics in 2012 at the Gauahati University under Dilip C Nath (statistics). His dissertation was titled "Bayesian Analysis for Longitudinal Data on Type 2 Diabetes Patients". After that, he joined the Ward of Clinical Research and Biostatistics at Malabar Cancer Centre (MCC) as a Lecturer. Over the last few years, A. Bhattacharjee has published more than 35 articles, in various peer-reviewed reputed journals.

Dehnel Grażyna is an Assistant Professor at the Department of Statistics, Poznań University of Economics. Her main research domain is short-term and structural business statistics, small area estimation. She is also interested in outlier robust regression applied on business data, data matching and business demography.

Domański Czesław is a Professor the Department of Statistical Methods in University of Lodz, Poland. His research interests are multivariate statistical analysis, construction of tests based on the runs theory and order statistics construction of statistical tables for selected non-parametric statistics based on exact distributions and recursive formulas, non-classical methods of statistical inference including the Bayes and bootstrap analysis and non-parametric inference, analysis of properties of multivariate normality tests, small area statistics, medical statistics, statistical methods on capital and insurance market, tests based on stochastic processes. Professor Domański has published more than 220 research papers in international/national journals and conferences. An author or co-author of 22 books including 15 monographs. He is an active member of many scientific professional bodies.

Dziechciarz Józef (Professor Dr) works at Wroclaw University of Economics, Poland, as director of the Institute for Application of Mathematics. His major areas of research interest include multivariate statistical analysis and econometric modelling of socio-economic data. His applied research concentrates on areas of measurement and modelling of education quality and effectiveness along with market data analysis. The labour market problems including job-related satisfaction measurement is natural extension. He uses the socio-economic data for modelling the behaviour of market agents. The list of multivariate statistical analysis and econometric techniques include those for metric, non-metric and mixed data.

Dziechciarz-Duda Marta (PhD) is an Assistant Professor in the Department of Econometrics and Computer Science at Wroclaw University of Economics, Poland. Her major areas of research interest include multivariate statistical analysis and econometric modelling of socio economic data. Her research concentrates on areas of consumer durable goods market, its structure analysis, segmentation and sales forecasting. She uses information on households' endowment with durable consumer goods as a proxy for the measurement and assessment of households' material situation. Her second field of research interest is problems of measurement and modelling of education quality and effectiveness.

Gołata Elżbieta is an Associate Professor at the Department of Statistics, Poznan University of Economics & Business. Her main research domain is demography, labour market and social statistics, especially territorial differentiation in demographic processes and labour market situation. She is also interested in data quality assessment, particularly in population census and other demographic estimates. Regional statistics with special emphasis on survey sampling and small area estimation applied to data of economic activity of population is also one of her fields of study. Professor Golata is an active member of many scientific professional bodies.

Grzenda Wioletta is an Assistant Professor at the Institute of Statistics and Demography at Warsaw School of Economics. Doctor in mathematics, 2006. Her main areas of interest are: advanced statistical methods, Bayesian statistics, Markov Chain Monte Carlo methods, survival analysis, generalized linear models, data mining and SAS programming. An author of papers on the applications of Bayesian and classical statistical methods in the analysis of unemployment and fertility, and an author and co-author of papers on theoretical mathematics. She has published books on Bayesian statistics, advanced methods of statistical analysis and SAS programming.

Gurgul Henryk (Professor) graduated from the Faculty of Management, University of Economics in 1977 and the Faculty of Applied Mathematics, AGH University of Science and Technology in 1980, both in Cracow. Currently, he is the head of the Department of Applications of Mathematics in Economics at AGH University of Science and Technology. His research is focused on financial econometrics, economic growth, macroeconomics and multisectoral input-output models. Awarded in 2007 with the most prestigious economic prize in Poland - "Bank Handlowy w Warszawie S.A. Award" (City Bank). In 2015 prof. Gurgul received the Medal of Honor of the University of Graz, Austria.

Korter Grace Oluwatoyin (PhD) is a graduate of the Department of Statistics, University of Ibadan and a lecturer at the Department of Mathematics/Statistics, Federal Polytechnic, Offa, Nigeria. Her research interests focus on building econometric and spatial econometric models for policy analysis and designing, organising, conducting and analysing survey data. She received a number of distinctions such as the 2010 Education Trust Fund Academic Staff Training and

Development Scholarship; 2011 Tertiary Education Trust Fund Conference Attendance Intervention Grant; 2013 World Social Science Forum Conference Grant; 2014 EPFL UNESCO Grant and 2014 African Development Bank/African Development Institute Grant. She is a member of the Royal Statistical Society (RSS), London and the Nigerian Statistical Association.

Noronha Vanita works as a Professor in the department of Medical Oncology in Tata Memorial Hospital, Mumbai, India. She specializes in management of malignancies arising in the thoracic region (lung and esophagus), head and neck and genitourinary system. She did basic medical training (MBBS) at Seth G.S. Medical College and KEM Hospital, Mumbai, India, then did residency training in Internal Medicine at State University of New York in Buffalo, NY (USA), followed by postdoctoral fellowship in hematology and medical oncology at Yale University School of Medicine, CT (USA). Her research interests include primarily clinical research, asking questions that arise in the OPD every day, and drug repurposing.

Olubusoye Olusanya Elisa (PhD) is a Lecturer of Statistics and a Senior Research Fellow with the Centre for Econometric and Allied Research University of Ibadan, Nigeria. He has served as a consultant to the African Union on Education Management Information System and development of indicators for monitoring the Plan of Action for the second decade of education for Africa. He has worked in research teams as principal researcher for several organizations, including the African Econometric Research Consortium based in Nairobi, Kenya, and MacArthur Foundation and World Bank funded projects. He is a member of the Royal Statistical Society, London; African Econometric Society; International Biometric Society Group Nigeria; and The Nigerian Statistical Association. He is a Chartered Statistician of the RSS and the 2nd Vice President of the Nigerian Statistical Association.

Pal Surya K. is a Research Scholar, Pursuing Ph.D. (Statistics) School of Studies in Statistics, Vikram University Ujjain, M. P., India. His research interests are in the areas of Statistics, Sampling Theory. He received his M .Phil. (Statistics) from School of Studied in Statistics, Vikram University Ujjain, M.P., India and M.SC. (Mathematics & Statistics) from Department of Mathematics and Statistics (Centre of Excellence Department by U.P. Government). The Department established in 1984 offers postgraduate program in Mathematics with specialization in statistics.

Pekasiewicz Dorota is an Assistant Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. Her major scientific research focuses on order statistics and their applications to estimation procedures and testing hypothesis. Moreover, she is interested in sequential, bootstrap and Bayesian statistical methods.

Salisu Afees Adebare (PhD) is a Senior Research Fellow with the Centre for Econometric and Allied Research (CEAR), University of Ibadan. He lectures applied and theoretical economics courses including Econometrics, Macroeconomics, Economic Modeling and Development Economics. He served as a consultant/facilitator to Alexander Brookes Training, UK, Centre for Management and Development (CMD), Lagos and West African Institute for Financial and Economic Management (WAIFEM), Lagos. He is a recipient of several fellowships and distinctions, including the 2010 Visiting Research Fellowship, Department of Economics, University College, London, and the 2010 Young African Scholar Award American University Cairo (AUC), Egypt. He is a member of the Royal Economic Society (RES), UK; the Nigerian Economic Society (NES); the African Econometric Society (AES); and Research Network, African Economic Research Consortium (AERC).

Singh Housila P. is a Professor at the School of Studies in Statistics, Vikram University Ujjain, M.P., India. His research interests are in the areas of Statistics, Sampling Theory and Statistical Inference. He has published more than four hundred research papers in reputed national/international journals. The School of Studies in Statistics, as premier department, was established in the year 1960 with the objective of developing competent and responsible statistician for the futuristic needs of India. The School has been relentlessly working in the areas of Agriculture, Population Studies, Industry, Pharmacy, Banking, Management, Information Technology, Remote-Sensing, etc. The P.G. programs offered here are truly of an international perspective.

Subramani Jambulingam graduated from University of Madras, Chennai, India. His contributions in research are: the modified MINQUE; non-iterative least squares method for estimation of several missing values from standard experimental designs; several new sampling schemes namely determinant sampling, diagonal systematic sampling, generalized diagonal systematic sampling, modified systematic sampling, optimal circular systematic sampling and recreational mathematics. He has received prestigious National and International Young Statistician awards instituted respectively by ISPS, India and ISI, Netherlands. He has published several research papers in reputed journals; more than two decades of professional experience both in industries and academic institutions; organized several Mathematical Exhibitions to promote interests in Mathematics among the students and youngsters and to overcome math phobia. Jambulingam Subramani is currently serving as an Associate Professor of Statistics, Pondicherry University.

Suder Marcin is working as an assistant at the Faculty of Management at the AGH University of Science and Technology in Krakow. In 2015 he received his PhD degree at the same department for the work titled "Forecasting withdrawals from ATMs as part of the ATM network management." Currently, he continues to work in the field of network management of ATMs.

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 797–798

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served from 31 December 2015 to 31 December 2016 as peer-reviewers of manuscripts for the ***Statistics in Transition new series – Volume 17, Numbers 1–4***; the authors` work has benefited from their feedback.

Bajpai Ram C., Nanyang Technological University, Singapore

Bąk Andrzej, Wroclaw University of Economics, Poland

Bourguignon Marcelo, UERN International Universidade do Estado do Rio Grande do Norte, Brazil

Breidt F. Jay, Department of Statistics, Colorado State University, United States

Bwanakare Second, University of Information Technology and Management in Rzeszow, Poland

Chambers Ray, University of Wollongong, Australia

Christofides Tasos, University of Cyprus, Cyprus

Dall'aglio Vincenzo, University of Parma, Italy

de Santis Gustavo, University of Florence, Italy

Diallo Mamadou, Westat, United States

Domański Czesław, University of Lodz, and Polish Statistical Association, Poland

Getka-Wilczyńska Elżbieta, Warsaw School of Economics, Poland

Gerstenkorn Tadeusz, University of Lodz, Poland

Ghosh Malay, University of Florida, USA

Giordani Paolo, LUISS „Guido Carli" University, Italy

Gogoi Bipin, Dibrugarh University, India

Górecki Tomasz, Adam Mickiewicz University in Poznan, Poland

Gulati Chandra, University of Wollongong, Australia

Hidiroglou Michael, Statistics Canada, Canada

Jajuga Krzysztof, Wroclaw University of Economics, Poland

Janbandu Ranjana, Dr. Ram Manohar Lohia Avadh University, India

Kalton Graham, Westat, United States

Kolonko Józef, University of Economics in Katowice, Poland

Kot Stanisław Maciej, Gdansk University of Technology, Poland

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

Korzeniewski Jerzy, University of Lodz, Poland

Kośny Marek, Wroclaw University of Economics, Poland

Kowalski Arkadiusz, Warsaw School of Economics, Poland

Krzyśko Mirosław, Adam Mickiewicz University in Poznan, Poland

Lapiņš Jānis, Bank of Latvia, Riga, Latvia

Lehtonen Risto, University of Helsinki, Finland

Lemmi Achille, University of Siena, Italy

Maggino Filomena, University of Florence, Italy

Marczewski Krzysztof, Warsaw School of Economics, Poland

Mazurek Edyta, Wroclaw University of Economics, Poland

Młodak Andrzej, Statistical Office Poznan, Poland

Molina Isabel, Carlos III University of Madrid, Spain

Morales Domingo, Miguel Hernández University of Elche, Spain

Morrone Adolfo, Italian National Institute of Statistics (Istat), Italy

Okrasa Włodzimierz, University of Cardinal Stefan Wyszynski, and Central Statistical Office, Poland

Panek Tomasz, Warsaw School of Economics, Poland

Pawlak Mirosław, University of Manitoba, Canada

Pietrzak Michał Bernard, Nicolaus Copernicus University in Torun, Poland

Pollastri Angiola, University of Milano-Bicocca, Italy

Rao Jon, School of Mathematics and Statistics, Carleton University, Canada

Rodrigues Paulo M. M., Bank of Portugal, Portugal

Singh Sarjinder, Texas A&M University–Kingsville, USA

Suchecka Jadwiga, University of Lodz, Poland

Szreder Mirosław, University of Gdansk, Poland

Veijanen Ari, Statistics Finland, Finland

Wallgren Anders, BA Statistiksystem AB, Sweden

Wallgren Britt, BA Statistiksystem AB, Sweden

Wołyński Waldemar, Adam Mickiewicz University of Poznan, Poland

Yannis George, National Technical University of Athens, Greece

Zieliński Wojciech, Warsaw University of Life Sciences, Poland

Žylius Gediminas, Kaunas University of Technology, Lithuania

STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 799–802

INDEX OF AUTHORS, VOLUME 17, 2016

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold italics) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).