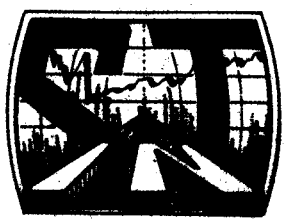


STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 11, Number 2, October 2010

ISSN 1234-7655

Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tele/fax: 00 48 22 – 825 03 95

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 211—213

FROM THE EDITOR

The current issue of the *Statistics in Transition new series* is, in a sense, a special one – namely, it marks some era in the history of the Journal as regards its layout being traditionally composed of two major sections: 'Sampling and estimation methods' and 'Other articles' which, since the year 2008, are complemented by the third section – "Current Issues in Public Statistics". From the next volume on, a new section entitled "Comparative surveys' will appear on a regular basis. Although this change has already been announced (in the previous issue of the Journal), I am mentioning it now in order to turn attention of our readers and collaborators, including potential Journal's contributors, to this initiative and to stress the fact that growing demand for a more systematic presentation of huge works being done in the area of multi-population surveys might be seen as a new hallmark for the contents of the term 'statistics in transition'. Together with Professor Vijaj Verma (University of Siena), who agreed to serve as a honorary co-editor of the new section, we look forward to obtaining papers devoted to any issue related to that area of statistical research.

Regarding this volume's contents, four papers included in the first section address various aspects of estimation. In paper *Estimating Potential Output Using Business Survey Data in a Svar Framework* **Tatiana Cesaroni** analyzes the concept of potential output and output gap (defined as the difference between actual and potential output) that play a central role in the macroeconomic policy interventions and evaluations due to conveying useful information on the cyclical position of a given economy. Estimation of the Italian potential GDP is proposed, based on structural VAR models - the estimates obtained through the SVAR methodology are free from end-of-sample problems, proving to be particularly useful for short-term analysis (especially as compared to such other techniques like the univariate filters – i.e. the Hodrick-Prescott filter). **Yunusa Olufadi's** paper *On The Estimation of Ratio-Cum-Product Estimators Using Two Stage Sampling* examines two-stage ratio-cum-product estimators with unequal sub-sampling fractions and obtains their MSE equations. Also, the optimum sampling and sub-sampling fractions were derived for these estimators and it was shown that, under certain conditions, these two-stage estimators will be more efficient than the Singh estimators (1965, 1967); the application of this technique, along with a numerical illustration and discussions, is included. Highly theoretical but practically oriented paper on important problem of sample rotation patterns by **Jacek Wesołowski**, *Recursive Optimal Estimation in the Szarkowski Rotation Scheme,* is devoted to the question of the recurrence form of optimal linear estimators of mean on every occasion. The presented solution is based on

a general approach devised in Kowalski and Wesołowski (2010). An explicit formulas for the coefficients of the recursion are derived. While the results conform Szarkowski's *three steps* conjecture, the Szarkowski *seven steps* conjecture remains open (even for the case of one singleton hole). In the next paper *Development of Standard Error Estimation Methods in Complex Household Sample Surveys in Poland* **Jan Kordos and Agnieszka Zięba-Pietrzak** provide a general description of estimation methods of standard error and confidence interval from complex household sample surveys in Poland. Presented are methods of estimation that have been applied in recent years: (i) the interpenetrating sub-samples, (ii) the Taylor series linearization, (iii) the jackknife, (iv) the balanced repeated replication, and (v) the bootstrap methods. A short development of each method, along with application in the Polish household sample surveys, are discussed.

A common feature of approaches employed in the next five papers that are gathered in section 'Other articles' and cover jointly wide spectrum of both statistical and econometric issues is extensive use of data for either hypothesis testing or measurement-related purposes. **Henryk Gurgul and Łukasz Lach** in paper *Causality Analysis between Public Expenditure and Economic Growth of Polish Economy in Last Decade* investigate the causal links between different kinds of budgetary expenditure and the economic growth of Poland employing both the linear and nonlinear Granger causality tests in order to evaluate the applicability of Wagner's Law vis-à-vis theory formulated by Keynes. Using aggregate and disaggregate quarterly data on public expenditure on human resources (HR), physical resources (PR), net interest payment (NIP) and other remaining budgetary expenditure (OTHER) they showed that relation between total budgetary expenditure and economic growth is, in general, consistent with Keynesian theory. In paper *Modeling of the internet sample* **Elżbieta Getka-Wilczyńska** discusses some stochastic models of Internet mediated (survey) research in which respondents' arrival process is being conceived as a stream of events and as a pure birth process in two different versions – the Poisson process and the cut models for the population of finite size. **Tomasz Panek** in paper *Multidimensional Approach to Poverty Measurement: Fuzzy Measures of the Incidence and the Depth of Poverty* presents multidimensional approach to analyzing poverty using the fuzzy set theory in order to assess the degree of household poverty threat instead of just dichotomy poor-nonpoor. Based on Polish dat from the EU-SILC survey the degree of monetary poverty and deprivation has been assessed showing in conclusions that indeed poverty in Poland has many dimensions. **Emilia Tomczyk** and **Barbara Kowalczyk** in paper *Influence of Non-response in Business Tendency Surveys on Properties of Expectations* discuss the question of whether expectations of industrial enterprises are formed rationally using data from business tendency surveys and accounting for non-response through weighting schemes. Two basic properties of expectation rational as introduced by J. F. Muth – that is, unbiasedness and orthogonality – are being analyzed and some ample balance statistics to correct for non-response

are proposed. In conclusion it is shown that rationality of expectations of Polish industrial firms is not sensitive to these factors, remaining basically independent from non-response and weighting system. Finally, in paper *Using ROC Curves to Find the Cut-Off Point in Logistic Regression with Unbalanced Samples* **Janusz Śmigielski**, **Anna Majdzińska** and **Witold Śmigielski** consider the problem of finding optimal cut-off value for the estimated probability $P(Y_i=1)$ (in order to transform this probability into values 0 or 1 of the endogenous variable) using the concept of the receiver operator characteristic (ROC) curves in logit models based on unbalanced samples. The proposed method is compared with some other popular methods discussed in the literature showing that the method can produce results relatively better than most of methods offered in the literature.

Włodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 215

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

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

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 217—230

ESTIMATING POTENTIAL OUTPUT USING BUSINESS SURVEY DATA IN A SVAR FRAMEWORK

Tatiana Cesaroni[1]

ABSTRACT

Potential output and the related concept of output gap play a central role in the macroeconomic policy interventions and evaluations. In particular, the output gap, defined as the difference between actual and potential output, conveys useful information on the cyclical position of a given economy. The aim of this paper is to propose estimates of the Italian potential GDP based on structural VAR models. With respect to other techniques, like the univariate filters (i.e. the Hodrick-Prescott filter), the estimates obtained through the SVAR methodology are free from end-of-sample problems, thus resulting particularly useful for short-term analysis. In order to provide information on the economic fluctuations, data coming from business surveys are considered in the SVAR model. This kind of data, given their cyclical profile, are particularly useful for detrending purposes, as they allow including information concerning the business cycle activity. To assess the estimates reliability, an end-of-sample revisions evaluation is performed. The ability of the cyclical GDP component obtained with the SVAR decomposition to detect business cycle turning points, over the expansion and recession phases of the Italian business cycle chronology is then performed.

Key words: potential output, business survey data, structural VAR models, end-of-sample revisions.

1. Introduction

Potential output and output gap are considered important indicators of the economic activity evolution. More in detail, the output gap, i.e. the difference between the actual output level and its potential, provides information concerning the cyclical position of the economy. In this sense it represents a benchmark to achieve non inflationary growth since if the output gap is positive (negative) the inflationary pressures rise (fall) and the policy makers are expected to tighten (ease) monetary policies. This indicator is also used by central banks to fix interest rates according to the so-called Taylor rules (Taylor, 1993).

[1] Italian Ministry of Economy and Finance.

However, in spite of the attention received, the estimates of those aggregates are still surrounded by a huge amount of uncertainty (cfr. Orphanides and van Norden, 1999 and 2001). This is mainly due to the fact that the output decomposition into its trend and cyclical components are not unique depending on the method used.

In the literature different methods have been used to estimate potential GDP. The most known univariate statistical techniques are based on the use of univariate filters (i.e. Hodrick and Prescott, 1997 and Baxter and King, 1995). Other univariate approaches include unobserved components models (see for details, Harvey, 1985 and Clark, 1987) and the Beveridge and Nelson (1981) decomposition. In addition, multivariate decompositions based on those techniques (i.e. multivariate filters or multivariate unobserved components models) have also been developed. Recently, considerable attention has been focused on the use of VAR models. To this end St-Amant and van Norden (1997) use a VAR model with long run restrictions including output, inflation, unemployment and real interest rate to estimate the Canadian output gap. Similarly Claus (2003) employs a SVAR model with long run restrictions to estimate New Zealand output gap for the period 1970-99.

The aim of this paper is to estimate Italian potential output using a multivariate decomposition based on the use of structural VAR models. Compared to other standard techniques, this kind of models show several advantages. Firstly, the estimates are free from end-of-sample problems, thus proving particularly useful for short-term analysis. In fact, compared to other methods using both past and future information to estimate the current data (i.e. moving averages), the end-of-sample VAR estimates are obtained by using only backward information. Secondly, the use of a multivariate decomposition model allows including information coming from more then one variable. In this sense, if compared to univariate decomposition methods, which only incorporate information coming from the decomposed variable, the multivariate method takes into account the external dynamics coming from other data. Moreover, as against other decomposition methods based on univariate filtering, the detrended series obtained with the SVAR methodology satisfies the Cogley and Nason (1995) critique, inasmuch the decomposition introduces no spurious cyclicality in the data.

Thirdly, compared to other multivariate techniques (i.e. multivariate filters) the framework allows for an economic interpretation of each variable's shocks. Fourthly, given its ability to act as a prediction model, the SVAR can be applied for forecast purposes.

Furthermore, to incorporate information on the economic fluctuations, data coming from business tendency surveys are considered in the model. Such data, given their cyclical behaviours are particularly useful for detrending purposes, since allow incorporating information on the cyclical economic activity. To assess the estimate reliability, an end-of-sample revision evaluation is performed. The results show that, compared with others standard methods, the output gap

estimates obtained through the SVAR model seems to have a negligible impact on the end-of-sample data revisions. This result makes this methodology particularly suitable for short-term analysis.

Finally, the ability of the output gap indicators (obtained through different methods) to detect the business cycle turning points is performed by comparing their peaks and troughs over expansion and recession periods of the Italian business cycle chronology.

The paper is organized as follows. Section 2 introduces the SVAR model and the identifying restrictions. Section 3 reports the empirical output gap estimates for Italy. Section 4 contains an evaluation of the impact of data revisions on SVAR estimates and a comparison with other univariate detrending methods. Section 5 includes an assessment of the ability of the estimated GDP cyclical components to detect turning points of the Italian official chronology. Section 6 concludes the work.

2. The model

To provide output gap estimates for Italy, we apply a SVAR model based on Blanchard and Quah (1989) identifying restrictions. The MA representation of the bivariate structural VAR model is given by:

$$\begin{bmatrix} \Delta y_t \\ bs_t \end{bmatrix} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} + \begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} v_{st} \\ v_{dt} \end{bmatrix} \tag{1}$$

where Δy_t is the growth rate of output, bs_t is a cyclical stationary variable coming from business tendency surveys, v_{st} and v_{dt} represent structural incorrelated supply and demand shocks and $A(L)$ is a 2x2 dimension polinomial matrix in the lag operator L. Alternatively, the model can be written in a compact form:

$$x_t = k + A(L)v_t \tag{2}$$

where $x_t = [\Delta y_t \quad bs_t]$ represents the vector of endogenous variables and $v_t = [v_{st} \quad v_{dt}]$ is the vector of aggregate shocks. Moreover, the shocks are normalized in order to have unit variance ($E(v_t v'_t) = I$).

The identifying restrictions are provided by assuming that demand-side shocks (i.e. to the cyclical indicator) only have a short-run impact on output, whereas supply-side shocks (i.e. productivity shocks) can produce long-run effects on output. More in detail, the identification is ruled out, imposing long-run restrictions on the coefficients of the MA representation of the structural VAR model.

Since the structural shocks are not observed, to evaluate the effects on the economy we need to derive them from the estimated residuals of the reduced-form model. The standard matrix representation of the bivariate reduced VAR form is given by:

$$\begin{bmatrix} \Delta y_t \\ bs_t \end{bmatrix} = \begin{bmatrix} \Phi_{10} \\ \Phi_{20} \end{bmatrix} + \begin{bmatrix} \Phi_{11}(L) & \Phi_{12}(L) \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ bs_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{st} \\ \varepsilon_{dt} \end{bmatrix} \tag{3}$$

or in a more compact formula:

$$x_t = \Phi_0 + \Phi_1(L)x_{t-1} + \varepsilon_t \tag{4}$$

where $\varepsilon_t = [\varepsilon_{st}, \varepsilon_{dt}]$ indicates the residual vector of the estimated model and $\Sigma_\varepsilon = E(\varepsilon_t \varepsilon_t')$ indicates the variance and covariance residual matrix, which generally is not diagonal. If the process is invertible (the polinomial matrix $\Phi(L)$ has unit root out of the unit circle), its moving average representation is given by:

$$x_t = K + C(L)\varepsilon_t \tag{5}$$

where $K = (I - \Phi_1)^{-1}\Phi_0$ e $C(L) = (I - \Phi_1(L)L)^{-1}$

Under the hypothesis that innovations are a linear combination of structural shocks, by equating (2) and (5) we obtain:

$$K + A(L)v_t = K + C(L)\varepsilon_t \tag{6}$$

For *L=0*, since $C(0) = I$ we have:

$$A(0)v_t = \varepsilon_t \tag{7}$$

where $A(0)$ is a 2x2 dimensions matrix. The variance and covariance matrix of the reduced form innovations is given by:

$$E(\varepsilon_t \varepsilon_t') = \Sigma_\varepsilon = A(0)E(v_t v_t')A(0)' \tag{8}$$

or in a matrix form:

$$\Sigma_\varepsilon = \begin{bmatrix} A_{11}(0)^2 + A_{12}(0)^2 & A_{11}(0)A_{21}(0) + A_{12}(0)A_{22}(0) \\ A_{11}(0)A_{21}(0) + A_{12}(0)A_{22}(0) & A_{21}(0)^2 + A_{22}(0)^2 \end{bmatrix}$$

Structural shocks v_t are determined from equation (7) as follows:

$$v_t = A(0)^{-1}\varepsilon_t \tag{9}$$

or in a matrix form:

$$\begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} = \begin{bmatrix} A_{11}(0) & A_{12}(0) \\ A_{21}(0) & A_{22}(0) \end{bmatrix}^{-1} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{gt} \end{bmatrix} \tag{10}$$

To recover the structural form shocks, it is necessary to know the coefficients of the $A(0)$ matrix. This latter expresses the contemporary effects of structural shock*s* on the variables considered. To identify the four coefficients of matrix *A(0),* the following restrictions are applied:

$$Var(\varepsilon_{yt}) = A_{11}(0)^2 + A_{12}(0)^2 \tag{11}$$

$$Var(\varepsilon_{gt}) = A_{21}(0)^2 + A_{22}(0)^2 \tag{12}$$

$$Cov(\varepsilon_{yt}\varepsilon_{gt}) = A_{11}(0)A_{21}(0) + A_{12}(0)A_{22}(0) \tag{13}$$

$$C_{11}(L)A_{12}(0) + C_{12}(L)A_{22}(0) = 0 \tag{14}$$

The first three restrictions stem directly from (8), the last restriction is obtained combining (6) and (7) with equation (2) and by assuming that cumulated demand shocks have no permanent effects on output[1] in line with the Blanchard and Quah framework.

For the GDP to be decomposed into cycle/trend components, the output gap Δy_t^{gap} is obtained by cumulating the demand shocks to output. Similarly, the potential output component Δy_t^{p} is determined by cumulating supply-side shocks. Starting from (2) and given that $C(L)A(0) = A(L)$, we have:

$$x_t = K + A(L)v_t = K + C(L)A(0)v_t = K + \sum_{i=0}^{\infty} \Phi_1^i L^i A(0)v_t$$

$$= K + \sum_{i=0}^{\infty} \Phi_1^i A(0)v_{t-i} \tag{15}$$

Considering only the first variable, we obtain:

$$\Delta y_t = K_1 + A_{11}(L)v_{st} + A_{12}(L)v_{dt}$$
$$= K_1 + A_{11}(0)v_{st} + A_{12}(0)v_{dt} + A_{11}(1)v_{st} + A_{12}(1)v_{dt} + A_{11}(2)v_{st}$$
$$+ A_{12}(2)v_{dt} + A_{11}(3)v_{st} + A_{12}(3)v_{dt} + ...$$

The potential GDP growth rate is given by:

[1] From equation (2) this implies that the coefficient
$A_{12}(L) = C_{11}(L)A_{12}(0) + C_{12}(L)A_{22}(0) = 0$

$$\Delta y_t^{pot} = K_1 + A_{11}(L)v_{st} = K_1 + A_{11}(0)v_{st} + A_{11}(1)v_{st} + A_{11}(2)v_{st} + A_{11}(3)v_{st} + ...$$

$$= K_1 + \sum_{i=0}^{\infty} \Phi^i{}_{11} L^i A_{11}(0)v_{st} = K_1 + A_{11}(0)\sum_{i=0}^{\infty} \Phi^i{}_{11} v_{st-i} \qquad (16)$$

the output gap is given by:

$$\Delta y_t^{gap} = A_{12}(L)v_{dt} = A_{12}(0)v_{dt} + A_{12}(1)v_{dt} + A_{12}(2)v_{dt} + A_{12}(3)v_{dt} + ...$$
$$= \sum_{i=0}^{\infty} \Phi^i{}_{12} L^i A_{12}(0)v_{dt} = A_{12}(0)\sum_{i=0}^{\infty} \Phi^i{}_{12} v_{dt-i} \qquad (17)$$

By using this kind of decomposition is thus possible to obtain an estimate of potential growth and cyclical output component based on economic hypothesis of the structural shocks effects.

3. Empirical results

In this Section, the results of the SVAR model specification are showed. As a preliminary analysis, we estimated different bivariate models by using output and various survey data indicators. Output is defined as the Italian Gross Domestic Product (expressed in euros at constant 1995 prices, seasonally adjusted source ISTAT). The business survey data come from Italian Manufacturing Business Surveys carried out by ISAE. In particular, we used data on the degree of plant utilization, on inventories, on the production level and on the confidence climate index[1] etc. These data, (except the degree of plant utilization) are qualitative data and are quantified through the balances[2]. The selection of business survey data to be included in the model was based on their degree of contemporary correlation with the GDP cyclical component obtained with an Hodrick-Prescott filter and on the basis of their stationarity in the sample.

Although we tried different specifications, in what follows we show the results of the bivariate model including the degree of plant utilization. This variable is able capture the whole economy cyclical dynamics[3] with great precision and to match business cycle evolution without introducing phase shifts.

The structural model specification, called SVAR, thus includes GDP in log differences and the degree of plant utilization. The lag structure of the reduced form was selected by using the Schwartz and Akaike criteria. The results of the Portmanteau test for the residual autocorrelation do not allow rejecting the null

[1] The confidence climate index is obtained combing data on orders level, inventories and production expectations.

[2] Balances are built as the difference between positive and negative answers provided by firms.

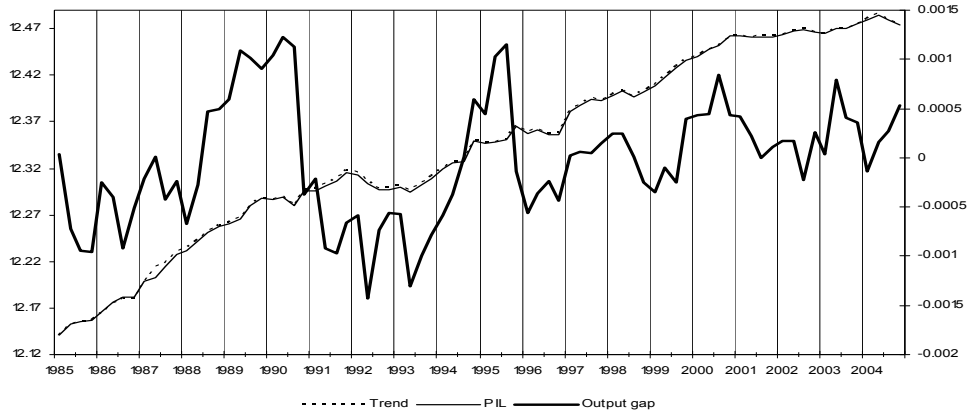
[3] Although the survey data refer to the manufacturing sector, they are able to thoroughly capture the whole economy dynamics (on this point see Hearn and Woitek, 2001 and Cesaroni, 2007).

hypothesis of autocorrelation absence. The usual heteroscedasticity test indicates omoscedastic residuals.

Figure 1 shows the estimated cyclical and trend components alongside with the actual GDP series. The output gap determined through the SVAR specification is positive from the second half of the Eighties till the Nineties and from 1994 to 1996.

The end-of-sample cycle becomes more erratic. These findings reflect the stagnation experienced by the Italian manufacturing sector in the last five years.

Figure 1. Trend/Cycle decomposition SVAR Model



4. Data revisions impact

A major aspect in the evaluation of a decomposition method performance is the data revision impact on the reliability of the end-of-sample estimates of the trend/cycle components.

Indeed, the end-of-sample estimates are subject to revision when new data become available. This updating process generates uncertainty on the real-time estimates that are of the utmost importance for policy-makers' decisions (van Norden, 1995). To this end, in what follows we assess the stability of the output gap estimates with respect to data revisions.

In our analysis, only the revisions due to new data availability are taken into account, while the impact on official data of the uncertainty estimates due to *ex post* revisions is not considered. This allows evaluating the effect of the end-of-sample revisions due to new data availability. However, on the basis of the evidence provided by Orphanides and van Norden (1999), the effect of National Accounts revisions on the output gap estimates should not be significant.

The reliability of real-time estimates is evaluated by quantifying the impact of 9-step-ahead data revisions on the output gap estimates referred to 2002 Q4. The

revisions are computed with respect to 9 quarters starting from 2003:1 to 2005:1 using the following formula:

$$(\left|y_{t/t+T} - y_{t/t+i}\right| \cdot 100) \tag{18}$$

where $y_{t/t+T}$ indicates the estimates at time *t,* including only the information available at time $t+T$ and $y_{t/t+i}$ indicates the estimates in *t,* obtained through the information available at $t+i$ with $i<T$.

In our case, the 1-step-ahead revisions *(t+1)*, as against to the estimates of 2002 Q3, are obtained as the difference between the estimates referring to 2002 Q4, made using all the information available at 2005 Q1 ($y_{2002q4/2005q1}$), and the estimates of 2002 Q4, based on the information available at 2003 Q1 ($y_{2002q4/2003q1}$).

Table 1 provides the data revisions of the output gap indicators obtained using linear and quadratic trend, the Hodrick-Prescott (1997) filter and the SVAR model. The impact evaluation of data revisions on the output gap real-time estimates shows that the estimates based on Linear trend and on the Hodrick-Prescott filter experienced the highest revisions. The revision amplitude at the end of period for those methods is equal to +1.035 and 0.84 respectively. Quite the reverse, the SVAR model revisions indicate a marginal impact on the estimates. Indeed, the amplitude of the highest revision equals 0.015.

Table 1. Data Revisions referring to 2002 Q4 estimates (% variations as against 2002 Q4)

t=2002:4	Pt/t+9-Pt/t+1	Pt/t+9-Pt/t+2	Pt/t+9-Pt/t+3	Pt/t+9-Pt/t+4	Pt/t+9-Pt/t+5	Pt/t+9-Pt/t+6	Pt/t+9-Pt/t+7	Pt/t+9-Pt/t+8	Pt/t+9-Pt/t+9
Sample	1980:1 2003:1	1980:1 2003:2	1980:1 2003:3	1980:1 2003:4	1980:1 2004:1	1980:1 2004:2	1980:1 2004:3	1980:1 2004:4	1980:1 2005:1
Linear trend	1.035	0.895	0.776	0.646	0.520	0.401	0.296	0.163	0.000
Quadratic trend	0.542	0.446	0.384	0.296	0.245	0.210	0.186	0.108	0.000
H-P filter	0.840	0.552	0.390	0.230	0.155	0.116	0.108	0.061	0.000
SVAR	0.015	0.016	0.010	0.008	0.008	0.010	0.008	0.007	0.000

These results corroborate the view whereby the output gap estimates obtained using VAR models are more reliable at the end of sample. The accurateness and reliability of SVAR estimates compared to univariate detrending methods make these models particularly suitable for short-term analysis purposes.

5. Business cycle chronology

To evaluate whether the estimated GDP cyclical components accurately indicate business cycle turning points, we make a comparison between the peaks and troughs identified through different output gap estimates and the turning points obtained through official cyclical Italian chronology. In particular, the output gap estimates obtained using a quadratic trend, the Hodrick-Prescott (1997) filter with a lambda parameter set to 1600, and the SVAR specification are compared. The sample period is 1985-05.
The Italian cyclical chronology used here comes from Altissimo, Marchetti and Oneto (1999). This methodology detects turning points and cyclical phases on the basis of the coincident indicator absolute variation level[1] and it is based on the classical cycle definition by Burns and Mitchell (1946).

Table 2. Maximum and minimum turning points. Italian cyclical chronology

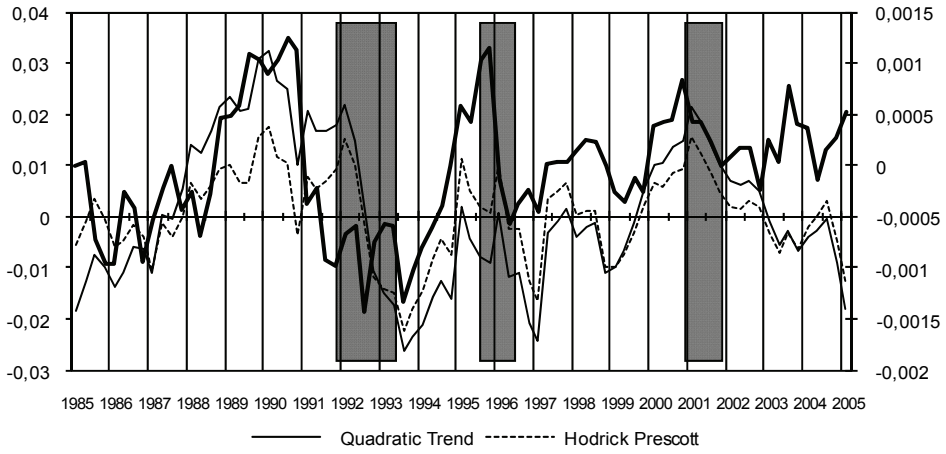
	Initial Minimum	Maximum	PHASES (in months)			Total Cycle
			Final Minimum	Expansion	Recession	
VIII	dic-77	mar-80	mar-83	27	36	63
IX	mar-83	mar-92	july-93	108	16	124
X	july-93	nov-95	nov-96	28	12	40
XI	nov-96	gen-01		49		

Source: ISAE.

Table 2 reports maximum and minimum turning points of the official Italian cyclical chronology, together with the length (in months) of the expansion and recession periods. The output gaps obtained with different detrending methods are evaluated so as to compare the different cyclical GDP components and turning points.

[1] Variables included in the coincident indicator are GDP, the industrial production index, imports of investment goods, the share of overtime hours, railway transport, machinery and equipment investments and the market services' value added.

Figure 3. Cyclical chronology (recession periods: grey area/expansion periods: white area)



Looking at the graph (Fig. 3), we notice that all the output gap estimates are able to indicate quite precisely the business cycle turning points, even though each estimate differs from the other in the dynamics displayed into the expansion and recessions zones. Moreover, the results show that, although the quadratic trend and the Hodrick-Prescott evolutions are relatively similar, the SVAR model estimates differ from those methods, particularly starting from 2001. The output gap, which is negative from 2001 to 2005 when using univariate estimates, seems positive in the same period when adopting the VAR model estimates. The difference in the two output gap indicator dynamics of SVAR as against the univariate methods stems from the use of an external signal (i.e. coming from business survey data).

6. Conclusions

This paper investigates the effects of a decomposition of real GDP into its trend and cyclical components by using a multivariate decomposition. In particular, we focused on the possibility to obtain reliable estimates of potential output and output gap using structural VAR models including data from business surveys.

From an economic point of view, those models provide an economic interpretation to the structural shocks. Furthermore, given that restrictions to shape the structure of each component are not required, the methodology does not impose an *a priori* limitation to modelling trend and cycle dynamics in the data. In this sense, while most detrending methods assume a random walk process for the trend component, the VAR decomposition does not involve a similar

assumption. Since the cyclical position can be identified more precisely when new data are available, a sensitiveness evaluation of the different output gap estimates with respect to data revisions is performed.

In our findings, the estimated output gap indicator is able to indicate quite precisely the turning points over the expansions and recessions periods of the Italian official chronology. The results show that, compared to other standard detrending methods, the output gap estimates based on SVAR model seem to have a negligible impact on data revisions at the end of sample. The results confirm the strength of this decomposition technique used in short-term analysis.

Appendix

Table 3. Portmanteau Test VAR model

Lags	Q-Stat	Prob.	Adj Q-Stat	Prob.	df
1	1.950138	NA*	1.974824	NA*	NA*
2	2.486994	0.6470	2.525445	0.6401	4
3	12.95609	0.1134	13.40243	0.0987	8
4	15.75192	0.2029	16.34541	0.1759	12
5	20.81236	0.1858	21.74321	0.1517	16
6	22.26165	0.3265	23.31001	0.2738	20
7	22.55443	0.5462	23.63086	0.4829	24
8	27.61519	0.4850	29.25393	0.3997	28
9	30.22760	0.5564	32.19749	0.4570	32
10	32.21071	0.6495	34.46391	0.5417	36
11	33.46755	0.7576	35.92111	0.6544	40
12	37.12075	0.7591	40.21899	0.6344	44
13	40.26798	0.7784	43.97687	0.6384	48
14	44.28770	0.7676	48.84927	0.5986	52
15	50.78118	0.6721	56.84124	0.4435	56
16	60.08622	0.4726	68.47254	0.2119	60
17	63.52040	0.4934	72.83340	0.2102	64
18	65.47009	0.5645	75.34913	0.2529	68

H0: no residual autocorrelations up to lag h.
Sample: 1985q1 2005q1.
Included observations: 80.
**The test is valid only for lags larger than the VAR lag order.*
df is degrees of freedom for (approximate) chi-square distribution.

Table 4. Lag selection criteria-VAR model

Endogenous variables: delta y and degree of plants utilization.
Exogenous variables: C.
Sample: 1985q1 2005q1.
Number of observations included:74.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	110.6097	NA	0.000182	-2.935397	-2.873125	-2.910556
1	162.0656	98.73970	5.05e-05*	-4.217989*	-4.031173*	4.143465*
2	163.2966	2.295634	5.44e-05	-4.143151	-3.831790	-4.018945
3	169.6471	11.49955*	5.11e-05	-4.206678	-3.770773	-4.032790
4	170.3205	1.182988	5.60e-05	-4.116769	-3.556321	-3.893200
5	172.9084	4.406566	5.83e-05	-4.078607	-3.393614	-3.805355
6	174.2266	2.173153	6.29e-05	-4.006124	-3.196588	-3.683190
7	174.6483	0.672520	6.95e-05	-3.909414	-2.975334	-3.536798

** lag order selection criterion.*
LR: sequential modified LR test statistic (each test at 5% level).
FPE: Final prediction error.
AIC: Criterio di informazione di Akaike.
SC: Criterio di informazione di Schwartz.
HQ: Criterio di informazione di Hannan-Quinn.

REFERENCES

ALTISSIMO F. MARCHETTI D.J. ONETO G.P. (1999) "The Italian Business Cycle: New Coincident and Leading Indicators and Some Stylized Facts", ISAE W.P n° 8.

BAXTER A. KING R. (1995) "Measuring business cycles approximate band-pass filters for economic time series" NBER W.P. No. 5022.

BEVERIDGE S. NELSON C.R. (1981) "A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle*" Journal of Monetary Economics* **7**, 151–174.

BLANCHARD O.J. QUAH D. (1989) "The Dynamics Effects of Aggregate Demand and Supply Disturbances" *The American Economic Review,* **79** 655–673.

BURNS E MITCHELL (1946) *Measuring Business Cycles.* New York, NBER.

CLARK P. K. (1987) "The cyclical component of US economic activity" *The Quarterly Journal of Economics* **102**, 794–814.

CESARONI T. (2007) "Inspecting the cyclical properties of Italian Manufacturing Business Survey" data W.P. ISAE n. 83.

CLAUS I. (2003) "Estimating potential output for New Zealand" *Applied Economics* **35**, 751–760.

COGLEY, TIMOTHY & NASON, JAMES M., (1995) "Effects of the Hodrick-Prescott filter on trend and difference stationary time series Implications for business cycle research", *Journal of Economic Dynamics and Control,* Elsevier, vol. 19(1–2), pages 253–278.

HARVEY A. C. (1985) "Trend and Cycles in U.S. macroeconomic time series" *Journal of Business and Economic Statistics* **3**, 216–27.

HEARN B.A and WOITEK U. (2001) "More international evidence on the historical properties of business cycles*" Journal of Monetary Economics,* 4, 289–319.

HODRICK R.J. and PRESCOTT E.C. (1997) "Postwar U.S. Business Cycles: An Empirical Investigation*" Journal of Money, Credit, and Banking* **29** (1):1–16.

ORPHANIDES, VAN NORDEN S. (2001) "The unreliability of output gap estimates in real time" *The review of economics and statistics* vol. LXXXIV n°4.

ORPHANIDES, VAN NORDEN S. (1999) *"*The Reliability of Output Gap Estimates in Real Time" Federal Reserve Board Finance and Economics Discussion series n° 1999–38.

ST-AMANT P. and S. VAN NORDEN (1997) "Measurement of the Output Gap: A Discussion of Recent Research at the Bank of Canada" Technical Report No. 79. Ottawa: Bank of Canada.

VAN NORDEN S. (1995) "Why is so hard to measure current output gap" Banque di Canada, mimeo.

TAYLOR J. (1993) "Discretion versus Policy Rules in Practice" Carnegie-Rocester Conference series on Public Policy vol. 39 pag 195–214.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 231—252

DEVELOPMENT OF STANDARD ERROR ESTIMATION METHODS IN COMPLEX HOUSEHOLD SAMPLE SURVEYS IN POLAND

Jan Kordos[1], Agnieszka Zięba[2]

ABSTRACT

The authors begin with a general description of estimation methods of standard error and confidence interval from complex household sample surveys in Poland. The following estimation methods have been applied in resent years: (i) the interpenetrating sub-samples, (ii) the Taylor series linearization, (iii) the jackknife, (iv) the balanced repeated replication, and (v) the bootstrap methods. A short development of each method is presented and its application in the Polish household sample surveys described. At the end some concluding remarks are given.

Key words: interpenetrating samples, Taylor series linearization, random groups, balanced repeated replication, jackknife, bootstrap, complex sample surveys.

1. Introduction

The Polish household sample surveys, such as the Household Budget Survey (HBS), the Labour Force Survey (LFS) and the EU Statistics of Income and Living Conditions Survey (EU-SILC) are complex sample surveys. These surveys typically use some of the following sampling techniques: sampling without replacement from a finite population, systematic sampling, stratification, clustering, unequal probabilities of selection, multi-stage or multi-phase sampling. As a consequence, the values of the variables of interest in a complex survey sample are neither independent nor identically distributed. In addition, survey processing, aimed at improving the quality and usability of survey data, reducing the bias of the estimates, etc. further increase the complexity of the survey data. Weighted analyses are necessary for unbiased (or nearly unbiased) estimates of population parameters. Standard error or confidence interval estimation depends upon the sampling plan specifics and requires approximate methods.

[1] Warsaw School of Economics.
[2] Warsaw School of Economics.

It is well known that sampling variance is an important measure of the quality of estimates of finite population parameters (totals, means, quantiles, ratios, etc.). It measures the amount of *sampling error* in the estimate due to observing a sample instead of the whole population. Estimates of sampling variance are needed to produce the *coefficients of variation* (cv) that are disseminated along with the survey estimates and to construct confidence intervals for finite population parameters of interest. Sampling variance is also used as the variability measure for inferences about super-population models when the *design-based approach* is recommended, that is when the sample design is *informative* (Binder and Roberts 2001, 2003, Rao, 2003).

Estimation of the sampling variance can become very complicated due to the complex sample design, use of non-linear estimators, impact of survey processing etc. Standard error estimation for estimators based on complex sample survey data must recognize the following factors:

a. most estimators are non-linear (a ratio of linear estimators is common);
b. estimators are weighted;
c. the sampling plan will generally have used stratification prior to first-stage sampling (and perhaps also at subsequent sampling stages); and
d. elements in the sample will generally not be statistically independent owing to multistage cluster sampling.

In almost all cases, it is not possible to obtain a closed-form algebraic expression for the estimated standard error or confidence intervals. Thus, the research literature on standard error estimation for complex sample survey data contains several approximate methods from which sample survey data analysts can choose [Brick et al, 2000; Eurostat, 2009; Kish and Frankel, 1974; Kovar et al, 1988; McCarthy and Snowden, 1985; Osier, 2006; Shao, 2003; Shao and Tu, 1995; Sitter, 1992; Tibshirani, 1985; Verma, 2005, 2010; Verma and Betti, 2005, Wolter, 1985].

In Poland the following estimation methods for standard error estimators from complex sample surveys have been used:

(i) the interpenetrating sub-samples (random groups),
(ii) the Taylor series linearization,
(iii) the jackknife replication techniques,
(iv) the balanced repeated replication (BRR),
(v) the bootstrap methods.

As stressed above, most household surveys are based on complex sample designs applied to very large populations. The primary sampling units (PSUs) are generally selected using probability proportional to size (PPS) without replacement sampling, making the concept of "sampling fraction" more complex.

However, the number of PSUs is often large and the PSU sampling fraction in each stratum is fairly small, giving a value close to 1.0 for all first-stage *fpc* terms. Thus, a common approximation in the analysis of complex sample survey data is one where the PSUs have been sampled with replacement in each stratum. If this approximation is made in the presence of some strata with large first-stage

sampling fractions, the variance will be overestimated to some extent. Such overestimation is often accepted in view of the complexity of standard error estimation without the approximation.

There are two basic approaches to variance estimation: i) an analytical approach using the *linearization* method, and ii) *resampling* and *replication* methods (jackknifing, balanced repeated replication, bootstrapping). The description of these methods can be found in many books on survey sampling, see for example (Osier, 2006; Shao, 2003; Shao and Tu, 1995; Tibshirani, 1985; Wolter, 1985; Lohr ,1999). The increase in computing power has made the use of the resampling techniques feasible for large survey samples. These methods are also relatively easy to implement because, regardless of the point estimator, a resampling method always uses the same procedure, replicated many times, while the linearization approach requires a development of a new formula for every estimator and weight adjustment and usually still requires additional simplifying assumptions (Binder, Kovacevic, and Roberts, 2004).

In Poland, we started applying different methods of standard error estimation in complex household sample surveys nearly forty years ago. Some of these methods were generally described in different papers mainly in Polish (GUS (1986 ,2009, 2010ab; Kordos, 1985, 1996, Kordos and Szarkowski, 1974; Kordos et al., 2002; Lednicki, 1982, 2011).

Short descriptions of standard error estimation methods for complex household sample surveys in Poland are given below. For each estimation method of standard error for complex sample surveys in Poland, a general development of the method is presented and application of the method for particular survey in Poland described.

2. Standard error estimation methods for complex sample surveys in Poland

Household sample surveys in Poland have a long tradition; however we focus our attention first on random household budget surveys which were started over fifty years ago (GUS, 1986; Kordos, 1996, 2001; Kordos et al., 2002; Lednicki 1982). A lot of attention was devoted to these surveys due to their special role in the analysis of the living conditions of the population. Various methods of surveys were experimented with and attempts were made to improve their methodology and organization. Here attention is paid to different methods of standard error estimation in complex household sample surveys.

At the beginning of 1970s we applied the interpenetrating sub-samples method for standard error estimation in the household budget surveys (Kordos and Szarkowski, 1974). Later our statisticians undertook some research in studying other methods for standard error estimation in complex household sample surveys, such as the Taylor linearisation, jackknife, balanced repeated replication, and bootstrapping.

2.1. The interpenetrating sub-samples

When two or more sub-samples are taken from the same population by the same sampling plan so that each sub-sample covers the population and provides estimators of the population parameters on application of the same sampling procedures, the sub-samples are known as *interpenetrating sub-samples* (Mahalanobis, 1946; Sukhatme & Sukhatme, 1970; Som, 1996). The sub-samples may or may not be drawn independently and there may be different levels of interpenetration corresponding to different stages in a multi-stage sample scheme.

This technique may be used to (Murthy, 1967; Som, 1996; Sukhatme, Sukhatme, 1970; Zasepa, 1972):

1) compute the sampling error from the first-stage units if these comprise one level of interpenetration (both by the standard method and also by a nonparametric test);
2) provide control in data collection and processing;
3) examine the factors causing variation, e.g. enumerators, field schedules, different methods of data collection and processing;
4) supply advanced estimates on the basis of one or more sub-samples when the total sample cannot be covered due to some emergency;
5) provide a basis of analytical studies by the method of fractile graphical analysis.

This technique has been used in a number of sample surveys in India, including the Indian National Sample Survey, in Peru, Zimbabwe, the Philippines and the USA [Som, 1996].

In Poland, the interpenetrating sub-samples were used in the HBS and the LFS for :

a) standard error estimation,
b) providing control in data processing,
c) supplying advanced estimates on the basis of one or more sub-samples when the total sample cannot be covered for different reasons, and
d) modular topics.

We used this method for standard error estimation in the HBS from 1970 till 1999, and the LFS from 1992 till 1999 (Kordos & Szarkowski, 1974; Szarkowski and Witkowski, 1994; Zasępa, 1972, 1993). . However, we realized some shortcomings of this method and some critical remarks were presented in the paper by Kordos and Szarkowski (1974).

The general idea of interpenetrating sub-samples is given below. For estimating parameter θ of population, we select from the population s sub-samples by the same sampling plan so that each sub-sample covers the population and provides estimators of the population parameters on application of the same sampling procedures. For each sub-sample the parameter θ is estimated, then we

get $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$, where $\hat{\theta}_h$ stands for an estimate of parameter θ from the sub-sample h (h = 1, 2, …, s). Then, the parameter θ is estimated from the formula:

$$\hat{\theta} = \frac{1}{s}\sum_{h=1}^{s}\hat{\theta}_i \tag{1}$$

It is possible to prove that if estimators $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$ are independent with the same expected values, then

$$s^2(\hat{\theta}) = \frac{1}{s(s-1)}\sum_{h=1}^{s}(\hat{\theta}_h - \hat{\theta})^2 \tag{2}$$

is unbiased estimator of variance of statistics given by (1).

As mentioned in the paper (Kordos, Szarkowski, 1974) there is also another advantage of using interpenetrating sub-samples. Let $\hat{\theta}_{\min,s_i}$ stands for the smallest, and $\hat{\theta}_{\max,s_i}$ for the largest of observed values of $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$ i.e.

$$\hat{\theta}_{\min,s,...,} = \min\{\hat{\theta}_1,...,\hat{\theta}_s\}$$
$$\hat{\theta}_{\max,s} = \max\{\hat{\theta}_1,...,\hat{\theta}_s\}.$$

Then, it is possible to show, that if estimator (1) is unbiased and symmetrical (what is fulfilled when sizes of samples are sufficiently large), than probability that the interval $\{\hat{\theta}_{\min,s_i}, \hat{\theta}_{\max,s_i}\}$ covers estimated parameter θ is equal:

$$P = 1-(\frac{1}{2})^{s-1} \tag{3}$$

It means that the interval $\{\hat{\theta}_{\min,s_i}, \hat{\theta}_{\max,s_i}\}$ may be interpreted as some kind of confidence interval for parameter θ, and level of confidence is given by (3). It is worth to add that for s = 4, P = 0.875, for s= 5, P = 0.94, for s = 8, P = 0.992, and for s = 10, P= 0.998.

The sub-samples may also be distinguished by differences in the survey procedures or processing features. These are sometimes known as *replicated samples***.** We also called them later *random groups*.

As stressed above, some of our statisticians from the Central Statistical Office of Poland (GUS) tried to apply other methods of standard error estimation, such as the Taylor linearization technique, the jackknife replication technique, the balanced repeated replication or some bootstrap methods.

2.2. The linearization method for variance estimation

The Taylor linearization technique for standard error estimation was introduced, after some experiments, in the 4th quarter of 1999 in the Polish Labour Force Surveys. At that time we had no access to ready-computer programme, and our sampling statistician, Andrzej Szarkowski, prepared such a computer-programme himself which was used on a larger scale in GUS for the Labour Force Survey from 1999 until to 2002. Since 2003 the bootstrap method has been used till present time (Popiński, 2006).

Basic idea of linearization method consists of deriving from a complex non-linear statistic a linear statistic which has the same asymptotic variance (Osier, 2006).

$$V(\hat{\theta}) \approx V(\hat{T}) = V(\sum_{i=1}^{n} w_i T_i) \tag{4}$$

where:

$\hat{\theta}$ – complex non-linear statistic

n – the sample size

w_i – sample weight of item i

Ti – linearized variable (variable whose expression depends on $\hat{\theta}$).

We would like to mention here some of our Polish experiments in case of the complex structure of the *Laeken* indicators (Zięba, Kordos, 2010). The asymptotic variance of the estimator is the variance of its linearization (Niemiro and Wieczorkowski, , 2005):

$$V(\hat{T}) = \sum_{h=1}^{L} \frac{a_h}{a_h - 1} \sum_{c=1}^{a_h} \left(\sum_{i=1}^{n_{hc}} w_{hci} T_{hci} - \frac{\sum_{c=1}^{a_h} \sum_{i=1}^{n_{hc}} w_{hci} T_{hci}}{a_h} \right)^2 \tag{5}$$

where:

w_{hci} – the survey weight attached to i^{th} sample unit (individual) in the c^{th} cluster (PSU) of the h^{th} stratum

T_{hci} – corresponding linearized variable

$\hat{T} = \sum w_{hci} T_{hci}$ – the estimator of target parameter θ

h – number of strata *(h=1, 2,...,L)*

a_h – number of clusters (PSUs) in stratum h

n_{hc} – the sample size (number of individuals) in the c^{th} cluster (PSU) of the h^{th} stratum.

In our experiments different linearization framework was applied. Estimation of the *At-risk-of-poverty rate after social transfers* and *Relative median poverty risk gap* makes use of nonparametric techniques of quantile estimation. (Zięba, Kordos, 2010). The asymptotic variances of these estimators involve the density (*f*) of the underlying probability distribution and standard kernel estimators were applied. Estimation of *S80/S20 income quintile share ratio* is closely related to the theory of M-estimators and GINI can be estimated by a U-statistic (Niemiro, Wieczorkowki, 2005).

Summing up our experience with the Taylor linearization estimation, we would like to stress that the method can be applied in general sampling designs, theory is well developed, and now software is available. However, there are also some cons: (i) finding partial derivatives may be difficult; (ii) different method is needed for each statistic; (iii) the function of interest may not be expressed a smooth function of population totals or means; (iv) accuracy of the linearization approximation.

2.3. The Jackknife Replication Technique

The jackknife replication technique was used in Poland experimentally with other estimation methods of standard errors for complex sample surveys, such as the Taylor linearisation, balanced repeated replication, and some bootstrap methods. From our experiments and from international literature, we have drawn conclusions, that the jackknife method is not as efficient as the bootstrap method used by us (Zięba, Kordos, 2010).

The jackknife, originally introduced as a method of bias estimation (Quenouille, 1949) and subsequently proposed for variance estimation (Tukey, 958), involves the systematic deletion of groups of units at a time, the re-computation of the statistic with each group deleted in turn, and then the combination of all these recalculated statistics. The simplest jackknife entails the deletion of single observations, but this delete-one jackknife is inconsistent for non-smooth estimators, such as the median and other estimators based on quantiles (Efron, 1979). Shao and Wu (1989) and Shao and Tu (1995) have shown that the inconsistency can be repaired by deleting groups of observations. Rao and Shao (1992) describe a consistent version of the delete-one jackknife variance estimator using a particular hot deck imputation mechanism to account for non-response.

Jackknife, and bootstrap are very similar, but bootstrap overshadows the others for it is a more thorough procedure in the sense that it draws many more sub-samples than the others. Through simulations, it is found that the bootstrap technique provides less biased and more consistent results than the Jackknife method does.

The jackknife is a less general technique than the bootstrap, and explores the sample variation differently. However the jackknife is easier to apply to complex

sampling schemes, such as multi-stage sampling with varying sampling weights, than the bootstrap.

The jackknife and bootstrap may in many situations yield similar results. But when used to estimate the standard error of a statistic, bootstrap gives slightly different results when repeated on the same data, whereas the jackknife gives exactly the same result each time (assuming the subsets to be removed are the same).

The jackknife works well only for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases.

Thus improvements to this technique were developed.

More information on comparisons of jackknife, linearization and bootstrap is in our last publication (Zięba, Kordos, 2010).

2.4. The balanced repeated replication (BRR)

Balanced half-sampling (McCarthy, 1969) is the simplest form of balanced repeated replication. It was originally developed for stratified multistage designs with two primary sampling units drawn with replacement in the first stage. Two main generalizations to surveys with more than $n_h = 2$ observations per stratum have been proposed. The first, investigated by Gurney and Jewett (1975), Gupta and Nigam (1987), Wu (1991) and Sitter (1993), uses orthogonal arrays, but requires a large number of replicates, making it impractical for many applications. The second generalization, a simpler more pragmatic approach, is to group the primary sampling units in each stratum into two groups, and to apply balanced repeated replication using the groups rather than individual units (Rao and Shao, 1996; Wolter, 1985). The balanced repeated replication variance estimator can be highly variable, and a solution to this suggested by Robert Fay of the US Bureau of the Census (Fay, 1989) is to use a milder reweighting scheme. Another solution (Rao and Shao, 1996, 1999) is to repeat the method over differently randomly selected groups to provide several estimates of variance, averaging of which will provide a more stable overall variance estimate.

With BRR, a half-sample replicate is formed by selecting one unit from each pair of PSUs and weighting the selected unit by 2 (so that it represents both units). Consequently, estimates from every PSU are in each replicate although half are weighted by zero. Though the number of half-samples can be quite large (2L, where L = the number of strata), the BRR method requires that only some of the possible half-sample replicates be created to obtain a variance estimator that reduces to the textbook (approximate sampling formula) variance estimator, which is an unbiased estimator of the true population variance. It is used a Hadamard matrix (a $k \times k$ orthogonal matrix consisting of 1's and -1's, where k is a multiple of 4) to specify the replicates, choosing a value of k that is greater than L. To minimize the number of replicates, usually the smallest value of k possible is chosen. With a fully balanced design, each pair of sample units is assigned to a

unique row in the Hadamard matrix. Within pairs, each PSU is assigned to one panel, and this panel assignment is used in conjunction with each column value to assign replicate factors (Wolter, 1985, pp. 111-115). To reduce the number of replicates, surveys may assign more than one strata to the same row in the Hadamard matrix. This is known as *partial* balancing (Wolter, 1985, pp. 125-131) or as the *grouped* balanced half-sample method (Rao and Shao (1996) . With grouped half-sample replication, units in the L strata are divided into G groups with approximately L/G strata in each group. Units within groups are split into two panels, and half-samples are formed for each group. With grouped balanced half-samples, the replicate variance estimator is not equivalent to the textbook estimator. Although it is still an unbiased estimator for the true variance, its variance is larger than the corresponding fully balanced method because cross-product terms between replicates no longer cancel (Wolter, 1985, pp. 127). Rao and Shao (1996) prove that the grouped half-sample replicate estimator is inconsistent.. Consequently, replicate weighting cells are often collapsed (more than in the full survey data procedure), which can induce a positive bias in the variance estimates (Rao and Shao,1999)). The Modified Half Sample (MHS) replication method developed by Robert Fay (1989) addresses this problem of overly perturbed replicate weights and drastically reduced replicate sample sizes.

The method of balanced half-samples was originally developed for the case of a large number of strata and a sample composed of only two elements (or two PSUs) per stratum. The idea of using half-samples for variance estimation was introduced at the US Bureau of the Census around 1960 (Särndal et al, 1992; Walter, 1985).

The BRR procedure consists of three steps:

1. Forming balanced half-samples from the full sample;
2. Constructing the replicate weights to be used in calculating the estimate of the parameter of interest for the subsamples; and
3. Computing the estimates of variance for the parameter of interest.

Forming Random Groups

The random groups comprise all possible balanced half-samples of PSUs. The sample design must include two PSUs per stratum (although many actual samples using this method form psuedo strata including psuedo-PSUs that meet this requirement). Each half sample is formed by deleting one PSU from each stratum. All possible combinations of such half-samples comprise the set of replicates.

Constructing Replicate Weights

Typically, BRR is implemented using replicate weights. Weights are set to zero for elements in excluded PSUs, and corresponding adjustments made to the weights of elements in the remaining PSUs. One set of such weights is constructed for each replicate.

Fay's method uses a milder adjustment to each weight. In some instances this approach improves the estimates of statistics such as medians and percentiles.

Computing Estimates of Variance

Assume that A such groups have been constructed. Then, for each group (a = 1,... ,A), the target statistics $\hat{o}_{(a)}$ is calculated based on data from half-sample a. The point estimate for the statistic may be estimated as the average of these half-sample estimates, or based on the single overall sample. The variance estimate is given by

$$V(\hat{\theta}) = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_a - \bar{\hat{\theta}})^2 \tag{6}$$

When using Fay's method the formula becomes

$$V(\hat{\theta}) = \frac{1}{A(1-k)^2}\sum_{a=1}^{A}(\hat{\theta}_a - \bar{\hat{\theta}})^2 \tag{7}$$

The BRR method for the Polish Household Budget Survey

The balanced repeated replication technique has been used in Poland for the Household Budget Survey since 2000 till now (Kordos et al., 2002; Lednicki, 2010). The general idea is presented below.

The basic estimated parameters for HBS are monthly expenditure (income or consumption) per head or equivalent unit (Lednicki, 2010), i.e.

$$R = \frac{X}{Y} \tag{8}$$

Estimate of parameter R is

$$\hat{r} = \frac{\sum_h \hat{x}_h}{\sum \hat{y}_h} \tag{9}$$

In turn, the values in nominator and denominator of (9) are estimated as follows:

$$\begin{aligned} \hat{x}_h &= \sum_i \sum_j w_{hij} \cdot x_{hij} \\ \hat{y}_h &= \sum_i \sum_j w_{hij} \cdot y_{hij} \end{aligned} \tag{10}$$

where: x_{hij} is value of variable X in the j-th household and the i-th PSU in stratum h, y_{hij} is value of variable Y in the j-th household and the i-th PSU in stratum h, w_{hij} – weight of the j-th household and the i-th PSU in stratum h.

Weights w_{hij} are calculated taking into account selection probability of the household and next it is adjusted to ex post stratification. For that purpose data from population census are used on size of household, separately for urban and rural areas. For each of 12 categories of households the following coefficients are calculated:

$$M_k = \frac{G_K}{\hat{g}_k}, k = 1, 2, ..., 12) \tag{11}$$

where: G_k – the number of households of category k, $\hat{g}_k$ – the number of households of category k estimated from the sample.

Multiplier M_k is used for correction of primarily selection weights.

The method of balanced half-samples is used to estimate standard error of parameter ***r*** (Särndal et al., 1992) . Using this method, first calculated are values $x_{1h}, x_{2h}; y_{1h}, y_{2h}$, which are estimates of sum of values variables X and Y in stratum h for the first and second sample respectively. Next for *r* value calculated are balanced replicates r_α (α = 1, 2, …, A; A =128), according to the following formulas:

$$x_\alpha = 2 \cdot \sum_{h=1}^{L} [P_{\alpha h} \cdot x_{1h} + (1 - P_{\alpha h}) \cdot x_{2h}] \tag{12}$$

$$y_\alpha = 2 \cdot \sum_{h-1}^{L} [P_{\alpha h} \cdot y_{1h} + (1 - P_{\alpha h}) \cdot y_{2h}] \tag{13}$$

where: $P_{\alpha h}$ – element of **P** Hadamard matrix, in which elements on values –1 were replaced zeros. Elements of matrix **P** are 1 or 0 (are orthogonal); L – the number of strata (L = 96).

Using (12) and (13) r_a is calculated:

$$r_\alpha = \frac{x_\alpha}{y_\alpha} \tag{14}$$

Next variance V(r) and coefficient of variation CV(r) are estimated:

$$V(r) = \frac{1}{A} \sum_{a=1}^{A} (r_a - \hat{r})^2 \tag{15}$$

$$CV(r) = \frac{\sqrt{V(r)}}{r} \tag{16}$$

Each year for a large number of items are published their estimates, standard errors and relative standard errors (e.g. GUS, 2010b).

2.5. The bootstrap methods

The bootstrap is a method which has been increasingly used to estimate the standard error of estimates obtained from complex survey designs since the publication of the first research article (Efron, 1979). This method has been shown to work well for a wide range of estimators, including medians and quantiles, as well as smooth functions based on totals. In addition, the bootstrap can be less computer intensive than the jackknife method for surveys with a very large number of primary sampling units (PSUs), (Faucher, et al., 2003).

Bootstrap Method for Standard Error Estimation

Sampling statisticians started to study the use of bootstrapping for variance estimation in the mid eighties. A direct extension to surveys samples of the standard bootstrap method developed for i.i.d. samples is to apply the standard bootstrap independently in each stratum. This methodology is often referred to as the ***naïve bootstrap***. Because the naïve bootstrap variance estimator is inconsistent in the case of bounded stratum sample sizes, several ***modified bootstrap*** methods were proposed. The following bootstrap methods that were modified for survey samples are discussed in Shao and Tu (1996):

(i) The with-replacement bootstrap (McCarthy and Snowden, 1985),
(ii) the rescaling bootstrap (Rao and Wu, 1988, Rao, Wu and Yue, 1992),
(iii) the mirror-match bootstrap (Sitter, 1992), and
(iv) the without-replacement bootstrap (Gross, 1980, Chao and Lo, 1985, Bickel and Freedman, 1984, Sitter, 1992).

Modified Bootstrap

Rao and Wu (1988) proposed a bootstrap method for stratified multi-stage designs with WR sampling of PSUs that applied a scale adjustment directly to the survey data values. Rao, Wu and Yue (1992) presented a modification of the 1988 method where the scale adjustment is applied to the survey weights rather than to the data values. This modification increases the applicability of the method, from variance estimation for smooth statistics to the inclusion of non-smooth statistics as well. Here we describe the modified rescaling bootstrap method proposed by Rao, Wu and Yue (1992), and its connection with the McCarthy and Snowden's :

To estimate the variance of the estimator $\hat{\theta}$, the following steps (i) to (iv) are independently replicated *B* times, where B is quite large (typically, *B*=500).

(i) Independently in each stratum *h*, select a bootstrap sample by drawing a simple random sample of $n_h^{(b)}$ primary sampling units (PSUs) with replacement

from the n_h sample PSUs. Let $t_{hi}^{(b)}$ be the number of times that PSU *hi* is selected in the bootstrap sample *b*, *b*=1,2, …, *B*.

(ii) For each secondary sampling unit (SSU) *k* in PSU *hi*, calculate the initial bootstrap weight by rescaling its initial sampling weight:

$$w_{hik}^{(b)} = w_{hik}\{(1-\sqrt{\frac{n_h^{(b)}}{n_h-1}}) + \sqrt{\frac{n_h^{(b)}}{n_h-1}} \cdot \frac{n_h}{n_h^{(b)}} \cdot t_{hi}^{(b)}\}\ , \tag{17}$$

where w_{hik} is the initial sampling weight of the SSU hik, equal to the inverse of its selection probability, i.e. $w_{hik} = \frac{1}{\pi_{hik}}$.

(iii) To obtain the final bootstrap weight $fw_{hik}^{(b)}$, adjust the initial bootstrap weight $w_{hik}^{(b)}$ by *f* using all the same weight adjustments (e.g. non-response and calibration) that were applied to the initial sampling weight w_{hik} to produce the final survey weight fw_{hik}.

(iv) Calculate $\hat{\theta}^{(b)}$, the bootstrap replicate of estimator $\hat{\theta}$ by replacing the final survey weights fw_{hik} in the formula for $\hat{\theta}$.

The bootstrap variance estimator of $\hat{\theta}$ is then given by

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta})^2, \text{ or} \tag{18a}$$

$$v^*_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS})^2, \text{ with } \hat{\theta}^*_{\mathrm{BS}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}\ . \tag{18b}$$

The estimators (18a) and (18b) are Monte Carlo approximations of the bootstrap estimator of $V(\hat{\theta})$ given by $\hat{V}_{BS}(\hat{\theta}) = E_{BS}[\hat{\theta}^{(b)} - E_{BS}(\hat{\theta}^{(b)})]^2$, where E_{BS} denotes the expectation with respect to bootstrap sampling.

Both (18a) and (18b) are used in practice and they usually produce very similar values. However, the variance estimate (18a) is always larger than (18b), i.e. $v_{BS}(\hat{\theta}) \geq v^*_{BS}(\hat{\theta})$:

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS} + \hat{\theta}^*_{BS} - \hat{\theta})^2 = \frac{1}{B},\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS})^2 + (\hat{\theta}^*_{BS} - \hat{\theta})^2$$

$$= v^*_{BS}(\hat{\theta}) + (\hat{\theta}^*_{BS} - \hat{\theta})^2 \tag{19}$$

We see that the estimator (18a) includes a positive quantity $(\hat{\theta}^*_{BS} - \hat{\theta})^2$ which converges to zero for all consistent estimators $\hat{\theta}$.

Size of the bootstrap sample

If $n_h^{(b)} \leq n_h - 1$, then the bootstrap weights are never negative. When $n^{(b)}{}_h = n_h - 1$, the rescaling bootstrap reduces to McCarthy and Snowden's . Usually surveys use $n_h^{(b)} = n_h - 1$, mainly because it greatly simplifies the calculation of the bootstrap weights; the rescaling formula given in (17) becomes:

$$w_{hik}^{(b)} = w_{hik}\{\cdot\frac{n_h}{n_h - 1}\cdot t_{hi}^{(b)}\} \tag{20}$$

Note that if $t_{hi}^{(b)} = 0$, i.e. PSU hi is not selected to the bootstrap replicate b, its bootstrap weight (20) is zero. This formula is used for the McCarthy and Snowden's estimator.

At the Central Statistical Office of Poland (GUS), we use the McCarthy-Snowden bootstrap in the Labour Force Survey since 2003 (Popiński, 2006), and in the EU- SILC since 2005 till present time (GUS, 2008).

Bootstrap percentile confidence intervals

The percentile method does not assume that the sampling distribution is normal. Instead, the percentile method assumes that the distribution of the bootstrapped statistics approximates the true sampling distribution.

To estimate confidence intervals, we simply generate a large number of bootstrapped statistics and sort them in ascending order. The 95% confidence interval then can be estimated simply by selecting the bootstrapped statistics at the 2.5-th and 97.5-th percentiles.

1. Compute the statistic θ from your sample of values X.
2. Generate B bootstrapped samples, X_b^*, by randomly sampling X with replacement. Typically, B is a large number (e.g., > 1000).
3. For each bootstrapped sample X_b^*, compute θ_b^*.
4. Sort θ_b^* from the smallest to the largest value.
5. Letting $[\theta_1^* \leq \theta_2^* \leq \theta_3^* \leq,..., \theta_B^*]$, represent the ordered θ_b^* values, then the (100 ×α)% confidence interval for θ is $(\theta^*_{(l_o+1)}, \theta^*_{hi})$.
6. In a two-tailed test, H_0 (i.e., that $\hat{\theta} = \mu$) is rejected if $\mu \leq \theta^*_{(l_o+1)}$ or $\mu \geq \theta^*_{hi}$.

It means that one way to estimate confidence intervals from bootstrap samples is to take the α and 1 − α quantiles of the estimated values. These are called *bootstrap percentile intervals*. These can be contrasted with the asymptotic intervals derived from the maximum likelihood estimates plus or minus 1.96 standard errors. The intervals from the asymptotic theory are apparently too narrow (as well as being symmetric).

This simple bootstrap method is not the only way of making improved inferences over the asymptotic approach. Other bootstrap schemes are available, as are approaches based on likelihood or Bayesian considerations.

The interval between the 2.5% and 97.5% percentiles of the bootstrap distribution of a statistic is a 95% bootstrap percentile confidence interval for the corresponding parameter.

More accurate bootstrap confidence intervals: BCa and tilting

Any method for obtaining confidence intervals requires some conditions in order to produce exactly the intended confidence level. These conditions (for example, normality) are never exactly met in practice. So a 95% confidence Bootstrap Confidence Intervals in practice will not capture the true parameter value exactly 95% of the time. In addition to "hitting" the parameter 95% of the time, a good confidence interval should divide its 5% of "misses" equally between high misses and low misses. We will say that a method for obtaining 95% confidence intervals is accurate in a particular setting if 95% of the time it produces intervals that accurate capture the parameter and if the 5% misses are equally shared between high and low misses. Perfect accuracy isn't available in practice, but some methods are more accurate than others.

One advantage of the bootstrap is that we can to some extent check the accuracy of the bootstrap *t* and percentile confidence intervals by examining the bootstrap distribution for bias and skewness and by comparing the two intervals with each other. In general, the t and percentile intervals may not be sufficiently accurate when

- the statistic is strongly biased, as indicated by the bootstrap estimate of bias;
- the sampling distribution of the statistic is clearly skewed, as indicated by the bootstrap distribution and by comparing the t and percentile intervals; or
- we require high accuracy because the stakes are high

Percentile-T Method

Another method for generating confidence intervals that has been useful in some situations is the *percentile-t method*. The t statistic is the basis of the well-known t test, which assumes that the sample statistic of interest (typically the mean) is distributed normally. If this assumption is not valid, then the *t* calculated will not follow the theoretical t distribution, and statistical inferences based on that distribution will be incorrect. The *percentile-t method* addresses this issue by

using the bootstrap to estimate the true distribution of the *t* statistic, and then use the estimated distribution to create confidence intervals and make statistical inferences. The *percentile-t method* is similar to the percentile method, except the bootstrap is used to estimate the distribution of *t*, rather than $\hat{\theta}$. Confidence intervals calculated using the *percentile-t method* sometimes are referred to as "*studentized" i*ntervals.

Here is the algorithm:

1. Compute the statistic θ and the standard deviation of that statistic, $\hat{s}$, from the sample of values X. Then, compute T using the formula T = $\frac{\hat{\theta}-\mu}{\hat{s}/\sqrt{n}}$, where n is the sample size.
2. Generate B bootstrapped samples, X_b^* by randomly sampling X with replacement. Typically, B is a large number (e.g., > 1000).
3. For each X_b^*, compute θ_b^*, $\hat{s}_b^*$, and $T_b^* = \frac{\theta_b^* - \hat{\theta}}{s_b^{*/}/\sqrt{n}}$, where $\hat{s}$ is the standard deviation and n is sample size.
4. Sort T_b^* from smallest to the largest value.
5. Define l_o= ROUND(αB/2)and hi = B − lo.
6. Letting $[T_1^* \le T_2^* \le T_3^* \le ...T_B^*]$ represent the ordered T_b^*values, then the (100 ×α)% confidence interval for T is $((T_{lo}^*, T_{hi}^*)$.
7. In a two-tailed test, Ho, (that $\hat{\theta}$ = μ) is rejected if T«T_{lo} or »T_{hi} and the (100 ×α)% confidence interval for μ is $[\hat{\theta} - T_{hi}^* \frac{\hat{s}}{\sqrt{n}}, \hat{\theta} - T_{lo}^* \frac{\hat{s}}{\sqrt{n}}]$.

Notice that the percentile-t method requires an estimate of standard deviation of the sampling distribution, $s\sqrt{n}$, for each bootstrapped sample. In cases where there is no analytical formula for s, then it can be calculated by performing a bootstrap on the bootstrapped sample. In other words, the calculation of s may require a bootstrap within a bootstrap, or a double bootstrap. Obviously, such a procedure can be costly in terms of computing power.

We have presented these two bootstrap methods of precision estimation for complex sample surveys, and some research in this field will be conducted.

3. The practice of presentation and use of sampling errors in Polish household surveys

Sampling error is often the only error source presented in Polish publications when reporting survey estimates. Sampling errors are communicated in a number

of different forms, e.g., standard errors, relative standard error (coefficient of variation), and confidence intervals.

Reports from the Polish household sample surveys, such as the Household Budget Survey (HBS), the Labour Force Surve,0y (LFS) and EU-Statistics of Income and Living Conditions survey (EU_SILC), contain presentation of sampling errors, estimated according to the methods described above. Results of the HBS are published yearly in a special publication "*Household Budget Survey in .....".* Such publication contains selected items of estimated value with standard error and relative standard error (e.g. GUS, 2010a). Results of the LFS are published quarterly by the Central Statistical Office in a special publication "*Labour Force Survey in Poland*" in the Information and Statistical Papers series. The mentioned publication issues consist of methodological section, analytic chapter and statistical tables. Statistical tables are composed of review tables with selected results supplied *with precision indexes*, review tables from the years 1992-2004 and detailed tables presenting results of survey conducted in the most recently surveyed quarter. The LFS reports include the assessment of the precision of the most important items expressed as relative of a half of 95% confidence interval (e.g. GUS, 2010b). Reports of the EU-SILC results present in the annexes sampling error and relative sampling error for different income items by size and type of households (absolute error of estimates and relative error of estimates as they are called) (e.g. GUS, 2009). These households sample survey results are available also via the CSO internet site http://www.stat.gov.pl (link English/basic data/.

However, we have noticed that some of users of complex household survey results treated them as simple random samples for statistical analysis. For these reasons, several publications were devoted to study the *design effects* for several characteristics (e.g. Kordos et al., 2002).

In our opinion, presentation of sampling errors in reports from the complex household sample surveys in Poland is quite comprehensive, but formal. Users of these results should be informed how to interpret sampling errors for estimated parameters, and that for statistical analysis it is necessary to take into account that these results were obtained by complex sample surveys.

4. Concluding remarks

We presented here only general development of standard error estimation methods in complex household sample surveys in Poland during last forty years. Some of our experiments in this field have been only mentioned. Based on empirical data, we have compared three methods of sampling errors estimation: linearization, jackknife (JRR) and bootstrap (McCarthy & Snowden) for five income poverty indicators using data from European Statistics on Income and Living Conditions (EU-SILC) carried out by Central Statistical Office of Poland in year 2007 (Zięba & Kordos, 2010). We have studied how many bootstrap

replicates are needed for efficient standard error estimation. Our special attention has been paid to bootstrap estimation methods, and we have also tried to compare two bootstrap methods for precision estimation, i.e. the percentile method, and the McCarthy & Snowden method. Our study in this field is still continued to find efficient methods of precision estimation in complex household sample surveys.

Although the bootstrap may seem to perform better than other methods of standard error estimation of complex household sample surveys, in the sense that it permits statistical inference in very general circumstances, it is not substitute for good quality data. The performance of the bootstrap depends of the sample size. It is not possible to recommend minimum sample sizes, because each problem is different. However, increasing the number of bootstrap replicates or using a more sophisticated bootstrap procedure does not compensate for insufficient data. All the bootstrap can do is quantifying the uncertainty in the conclusion. It means that for social data which are usually biased for different reasons, bootstrap may only estimate precision of the estimates.

5. Acknowledgment

The authors would like to thank the Central Statistical Office of Poland for the production and provision of the survey data used, Prof. Vijey Verma and Dr. Waldemar Popinski, the referees, for their very constructive comments, Dr. Robert Wieczorkowski for consultation and computation, and Mr. Bronisław Lednicki for sampling consultation.

REFERENCES

BICKEL, P.J., and FREEDMANN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist., 12,* pp. 470–482.

BINDER, D.A., KOVACEVIC, M.S., and ROBERTS, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.* 3301–3312

BRICK, M.J., MORGANSTEIN, D., VALLIANT, R. (2000), Analysis of Complex Sample Data Using Replication. *WESTAT*, 1–10.

CHAN, M.T., and LO, S.H. (1985). A bootstrap method for final populations.. *Sankhya. A,* 47, 399–405.

EFRON, B., (1979) Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, 1–26

EUROSTAT, (2009) *Methodological studies and quality assessment of EU-SILC*, Report SILC.04 15 April 2009, SAS programs for variance estimation of the measures required for Intermediate Quality Report

FAUCHER D., LANGLET E. R., LESAGE E., (2003), An application of the bootstrap variance estimation method to the participation and activity limitation survey, *Assemblée annuelle de la SSC*, *Recueil de la Section des méthodes d'enquête.*

FAY, R.E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 212–217.

GROSS, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*., pp.181–184.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed arthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika,* 74, 735–742.

GURNEY, M. nad JEWETT, R.S. (1975). Constructing orthogonal replications for standard errors. *J. Amer .Statist. Assoc.*70, 819–821.

GUS (1986), *Methods and Organisation of Household Budget Survey*. Methodological Publication, No. 62, Warszawa (in Polish)..

GUS (2009), *Incomes and Living Conditions of the Population in Poland* (report from the EU-SILC survey of 2007 and 2008), Statistical Information and Elaborations, Warsaw.

GUS (2010a), *Household Budget Surveys in 2009*, Statistical Information and Elaborations, Warsaw.

GUS (2010b), *Labour Force Survey in Poland – I Quarter 2010*), Statistical Information and Elaborations, Warsaw..

KISH L., (1965) *Survey Sampling*, New York, Wiley

KISH L., FRANKEL M. R., (1974) *Inference from Complex Samples*, "Journal of the Royal Statistical Society", Series B. (Methodological), Vol. 36, No. 1. 1–37

KORDOS, J. (1985), Towards an Integrated System of Household Surveys in Poland, *Bulletin of the International Statistical Institute,* Vol. 51, Amsterdam, Book 2 , pp. 13–18.

KORDOS, J. (1996), Forty Years of the Household Budget Surveys in Poland. *Statistics in Transition*, Vol. 2, No. 7, pp.1119–1138.

KORDOS, J. (2001), Some Data Quality Issues in Statistical Publications in Poland, *Statistics in Transition,* vol. 5, No. 3, December 2001, pp. 475–488.

KORDOS ,J., LEDNICKI, B. and ZYRA, M.(2002), The Household Sample Surveys in Poland, *Statistics in Transition,* Vol. 5, No.4, pp. 555–589.

KORDOS ,J., SZARKOWSKI, A. (1974)**,** Method of Precision Estimation for Basic Parameters in Household Budget Surveys .*Wiadomości Statystyczne***,** 1974, nr 5. (in Polish).

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, **16***, 25–46.

LEDNICKI, B. (1982), Sampling Design and Estimation Method in Household Budget Rotation Survey. *Wiadomosci Statystyczne*, No. 9. (in Polish).

LEDNICKI, B. (2011), Application of the balanced repeated replication method in Polish Household Budget Survey. In: *Methods and Organisation of Household Budget Survey*. Methodological Publication, Warszawa (in Polish) (in preparation).

LOHR, S. (1999), *Sampling: Design and Analysis,* Duxbury Press.

MAHALANOBIS, P.C.(1946), Recent experience in statistical sampling in the Indian Statistical Institute*, JRSS(A),* 108, 326–378.

MCCARTHY P. J., SNOWDEN C. B., (1985), The Bootstrap and Finite Population Sampling, *Vital and Health Statistic*s, Series 2, no. 95, Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington DC

NIEMIRO W., WIECZORKOWSKI R., (2005) Approximating Variance for Measures of Income Inequality: An Approach Based on the Theory of M-Estimators and U-Stat*i*stic*s, Eurostat, 2005*

OSIER G., (2006) *Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation*, Technical report, Eurostat and Statistics Finland Methodological Workshop on EU-SILC, Helsinki, 7–8 November 2006 (http://www.stat.fi/eusilc/workshop_en.html)

QUENOUILLE, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics* 20, 255–375.

POPIŃSKI W., (2006), Development of the Polish Labour Force Survey, *Statistics in Transition,* Vol. 7, No. 5, pp.. 1009–1030.

RAO, J.N.K., and WU, C.F.J., (1988). Resampling Inferences with complex survey data. JASA, 83, 321–241.

RAO, J.N.K., and WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology,* 18, 209–217.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometricka,* 79, 811–822.

RAO, J.N.K. and SHAO, J. (1996) On Balanced Half-Sample Variance Estimation in Stratified Random Sampling. *Journal of the American Statistical Association,* 91, pp. 343 –348.

RAO, J.N.K. and SHAO, J. (1999). Modified Balanced Repeated Replication for Complex Survey Data. *Biometrika*, 86, pp. 403–415.

SÄRNDAL, C.E., SWENSSON, B., WRETMAN, J.H., (1992) *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SHAO J., (2003), Impact of the Bootstrap on Sample Surveys, "*Statististical Science*" Vol. 18, No 2, 191–198.

SHAO J. and TU, D., (1996) *The Jackknife and Bootstrap*, Springer-Verlag, New York.

SHAO, J. and WU, C. F.(1989). A general theory for jackknife variance estimation. *Ann. Statist., 15,*1563–1579.

SITTER R. R., (1992) Comparing Three Bootstrap Methods for Survey Data, *The Canadian Journal of Statistic*s, Vol. 20, No. 2, 135–154.

SITTER R. R., (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika,* 80, 211–221.

SOME, R. K. (1996), *Practical Sampling Techniques*, Marcel Dekker, Inc., New York.

SUKHATME, P.V. and SUKHATME, B.V. (1970), *Sampling Theory of Surveys with Applications*, FAO, Rome.

SZARKOWSKI, A., WITKOWSKI, J. (1994), The Polish Labour Force Survey*, Statistics in Transition*, Vol. 1, No. 4, pp. 467–483.

TIBSHIRANI, R., (1985) *How many bootstraps?*, Technical report no 362, Department of Statistics, Stanford University

TUKEY, J.W.(1958). Bias and confidence in not-quite large samples. . *Annals of Mathematical Statistics.* 29:614.

VERMA V. (1993). *Sampling Errors in Household Surveys: A Technical Study*. New York: United Nations Department for Economic and Social Information and Analysis. (Statistical Division INT-92-P80-15E).

VERMA V., BETTI G. (2005), Sampling errors and design effects, *Working Papers*, no. 53, Dipartimento di Metodi Quantitativi, Università di Siena.

VERMA, V., BETTI, G. (2010): Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics*; iFirst.

WOLTER, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

YUNG, W. and RAO, J.N.K. (1996). Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling. *Survey Methodology,* **22**, pp. 23–31.

ZASĘPA, R. (1972), *Sampling Methods*, PWE, Warsaw (in Polish).

ZASEPA, R. (1993). Precision of family budget survey results*, Wiadomosci Statystyczne,* No. 3. (in Polish).

ZIĘBA, A. and KORDOS, J. (2010), Comparing Three Methods of Standard Error Estimation for Poverty Measures. In: J. Wywiał, W. Gamrot (Eds), *Research in Social and Economic Surveys,* Katowice, University of Economics.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 253—265

ON THE ESTIMATION OF RATIO-CUM-PRODUCT ESTIMATORS USING TWO-STAGE SAMPLING

Olufadi, Yunusa[1]

ABSTRACT

In this paper, we apply two-stage sampling which is not only commonly used in survey sampling but also has great advantages over element and cluster sampling. These advantages are comprehensively described in some sampling literature. Thus, we examine two-stage ratio-cum-product estimators with unequal sub-sampling fractions and obtain their MSE equations. The optimum sampling and sub-sampling fractions were also derived for these estimators. It is shown theoretically that these two-stage estimators will be more efficient than Singh (1965, 1967) estimators if certain conditions are satisfied. Finally, a numerical illustration with discussions is carried out to show the application of this technique.

Key words: auxiliary information, efficiency, mean square error, optimum sampling fraction, ratio-cum-product estimator, two-stage sampling.

1. Introduction

The literature on survey sampling (Cochran (1977), Sukhatme et al. (1984) and references cited therein) describes a great number of techniques for using auxiliary information by means of ratio and product estimators. It is well known that Ratio estimators are efficient when there is a strong positive correlation (and preferably also a strong ratio relationship) between the variable of interest and the supplementary variable. Product estimators are only efficient when the variable of interest is negatively correlated with the supplementary variable, and preferably when it has a strong inverse relationship with it, so that the products are very similar.

However, it has been established theoretically that, in general, the linear regression estimator is more efficient than ratio and product estimators except when the regression line of y on x passes through the neighbourhood of the origin, in which case the efficiencies of these estimators are almost equal. Various

[1] Department of Statistics, University of Ilorin, Ilorin, Nigeria.

improvements of these estimators have been considered by many authors particularly in the presence of more than one auxiliary variable.

Olkin (1958) was the first author to deal with the problem of estimating the mean of a survey variable when auxiliary variables are made available. He suggested the use of more than one auxiliary variable, considering a linear combination of ratio estimators based on each auxiliary variable separately. The coefficients of the linear combination were determined so as to minimize the variance of the estimator. Analogously to Olkin, Singh (1967b) gave a multivariate expression of Murthy's (1964) product estimator, while Raj (1965) suggested a method of using multi-auxiliary variables through a linear combination of single difference estimators. Moreover, Rao and Mudholkar (1967) proposed a multivariate estimator based on a weighted sum of single ratio and product estimators.

Many other contributions are present in sample survey literature and recently some new estimators have emerged. Tracy *et al.* (1996) and Perri (2005), when two auxiliary variables are available, proposed an alternative to Singh's (1965, 1967) ratio-cum-product estimators while Kadilar and Cingi (2004, 2005) examined the combination of regression-type estimators in the presence of two auxiliary variables. Other references are: Muhammad *et al.* (2010), Abdul and Javid (2010), Tailor and Sharma (2009), Diana and Perri (2007), Samiuddin and Hanif (2007), Abu-Dayyeh *et al.* (2003), Upadhyaya and Singh (1999, 2003) and many others.

In most practical situations, the sampling frame may be unavailable and when it does, it may be expensive or not feasible to obtain the sampling units directly from a population of interest. Thus, in this case, we can make use of two-stage sampling whose efficiency could be improved by the use of auxiliary information. The use of two stage sampling in ratio-cum-product estimators when there is more than one auxiliary variable is scarcely considered and to the best of our knowledge, we can only mention Sukhatme *et al.* (1984), Sahoo and Panda (1997, 1999) and Hossain and Ahmed (2001) in the case of single auxiliary variable. Thus, in this paper, an attempt is made to apply two-stage sampling for estimating the population mean in Singh's (1965, 1967a) estimators which involve two auxiliary variables.

2. Two-stage sampling set-up

Let U be a finite population partitioned into N First Stage Units (FSU) denoted by $U_1, U_2, ..., U_I, ...U_N$ such that the number of Second Stage Units (SSU) in U_i is M_i. Let y and (x, z) be the study variable and auxiliary variates taking the values y_{ij} and (x_{ij}, z_{ij}) respectively, for the j^{th} SSU on the unit U_i $(i = 1,2,..., N; j = 1,2,..., M_i)$. Where x is assumed to be positively correlated

with y and z is negatively correlated with y. Suppose it is of interest to estimate the population mean $\overline{Y}$ of y and assume that the population means $(\overline{X}, \overline{Y})$ of (x, z) are known.

Define $\overline{P}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} P_{ij}$; $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} \overline{P}_i$; $P = X$, Y , Z . Assume that a sample s of n FSU's is drawn from U and then a sample s_i of m_i SSU's from the i^{th} selected FSU from U_i using simple random sampling without replacement.

Define also $\overline{p}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} p_{ij}$; $\overline{p} = \frac{1}{n}\sum_{i=1}^{n} \overline{p}_i$; $p = x$, y , z .

3. Ratio-cum-product estimators

The product-ratio estimators using one dual auxiliary variable were introduced by Badhopadyaya (1980) and Srivenkataramana (1980). The estimators are defined for simple random sampling and can be effective when one auxiliary variable is positively correlated with the study variable and the other variable is negatively correlated with study variate. Recently, Singh *et al.* (2005), Singh and Espejo (2007), Singh and Upadhyaya (1995) and a host of others have used the combination of ratio and product estimators to estimate the unknown population mean $\overline{Y}$, when $\overline{X}$ and $\overline{Z}$ are known but our interest in this paper is to apply two-stage sampling with unequal FSU to the ratio-cum-product estimators developed by Singh (1965, 1967a). These estimators are as given below:

$$t_1 = \overline{y}\frac{\overline{X}}{\overline{x}}\frac{\overline{z}}{\overline{Z}}; \quad t_2 = \overline{y}\frac{\overline{X}}{\overline{x}}\frac{\overline{Z}}{\overline{z}}; \quad t_3 = \overline{y}\frac{\overline{x}}{\overline{X}}\frac{\overline{z}}{\overline{Z}}; \quad t_4 = \overline{y}\frac{\overline{x}}{\overline{X}}\frac{\overline{Z}}{\overline{z}}.$$

The estimators are biased even if the bias may be considered negligible for large samples. Under suitable conditions, involving the correlation coefficients between the variables, it is easy to prove (Singh, 1965) that they are more efficient than ratio and product estimators which make use of a single auxiliary variable. Their corresponding Mean Square Errors (MSE) to the first degree of approximation is presented below:

$$MSE(t_1) = \lambda\left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} - R_2 S_{yz} + R_1 R_2 S_{xz}\right)\right] \qquad (1)$$

$$MSE(t_2) = \lambda\left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} + R_2 S_{yz} - R_1 R_2 S_{xz}\right)\right] \qquad (2)$$

$$MSE(t_3) = \lambda\left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 + 2\left(R_1 S_{xy} + R_2 S_{yz} + R_1 R_2 S_{xz}\right)\right] \quad (3)$$

$$MSE(t_4) = \lambda\left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} - R_2 S_{yz} - R_1 R_2 S_{xz}\right)\right] \quad (4)$$

where, $R_1 = \frac{\overline{Y}}{\overline{X}}$; $R_2 = \frac{\overline{Y}}{\overline{Z}}$; $\lambda = \frac{1-f}{n}$; $f = \frac{n}{N}$; $S_p^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2$; $S_{pq} = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\left(Q_i - \overline{Q}\right)$ $P,Q = X,Y,Z$; $p,q = x,y,z$; $P \neq Q$ and $p \neq q$.

4. Suggested method of estimation

The suggested two-stage ratio-cum-product estimators are as defined below:

$t_1^* = \overline{y}^* \frac{\overline{X}}{\overline{x}^*}\frac{\overline{z}^*}{\overline{Z}}$; $t_2^* = \overline{y}^* \frac{\overline{X}}{\overline{x}^*}\frac{\overline{Z}}{\overline{z}^*}$; $t_3^* = \overline{y}^* \frac{\overline{x}^*}{\overline{X}}\frac{\overline{z}^*}{\overline{Z}}$; $t_4^* = \overline{y}^* \frac{\overline{x}^*}{\overline{X}}\frac{\overline{Z}}{\overline{z}^*}$.

Define $\overline{y}^* = \overline{Y}(1+e_0)$, $\overline{x}^* = \overline{X}(1+e_1)$, $\overline{z}^* = \overline{Z}(1+e_2)$ such that $E(e_h) = 0$; $E(e_h^2) = \frac{V(\overline{P})}{\overline{P}^2}$ for $h = 0,1,2$; $E(e_h e_k) = \frac{Cov(\overline{P},\overline{Q})}{\overline{P}\overline{Q}}$ for $h,k = 0,1,2$ and $h \neq k$.

Under two-stage sampling with unequal FSU,

$V(\overline{P}) = \lambda S_{1p}^2 + \sum_{i=1}^{N}\lambda_{2i} S_{2pi}^2$; $Cov(\overline{P},\overline{Q}) = \lambda S_{1pq} + \sum_{i=1}^{N}\lambda_{2i} S_{2pqi}$ where

$\lambda_{2i} = \frac{M_i^2(1-f_{2i})}{nNm_i}$; $f_{2i} = \frac{m_i}{M_i}$; $S_{1p}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{P}_i - \overline{P}\right)^2$;

$S_{2pi}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_{ij} - \overline{P}_i\right)^2$; $S_{1pq} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{P}_i - \overline{P}\right)\left(\overline{Q}_i - \overline{Q}\right)$;

$S_{2pqi} = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_{ij} - \overline{P}_i\right)\left(Q_{ij} - \overline{Q}_i\right)$. S_{1p}^2 and S_{2pi}^2 are the variance among FSU means and variance among subunits for the i^{th} FSU while S_{1pq} and S_{2pqi} are their corresponding covariances.

In order to evaluate the precision of these estimators with respect to Singh (1965, 1967a) estimators, their MSEs need to be obtained; and this can be found to the first degree of approximation using Taylor's linearization method described by Wolter (1985).

If we express t_1^* in terms of $e's$, we have

$$t_1^* = \bar{Y}(1+e_0)(1+e_2)(1+e_1)^{-1} \tag{5}$$

Expanding the right hand side of (5) and neglecting terms involving powers of $e's$ greater than two, $t_1^* - \bar{Y}$ yields

$$t_1^* - \bar{Y} = \bar{Y}\left(e_1^2 + e_0 + e_2 - e_1 + e_0e_2 - e_0e_1 - e_1e_2\right) \tag{6}$$

Squaring both sides of (6) and neglecting terms of order greater than two, we have

$$\left(t_1^* - \bar{Y}\right)^2 = \bar{Y}^2\left(e_1^2 + e_0^2 + e_2^2 + 2e_0e_2 - 2e_0e_1 - 2e_1e_2\right) \tag{7}$$

Taking expectation on both sides of (7) and simplifying results in

$$MSE\left(t_1^*\right) = \lambda S^2_{1R_0^*} + \sum_{i=1}^{N} \lambda_{2i} S^2_{2R_{0i}^*} \tag{8}$$

where

$$S^2_{1R_0^*} = \left[S^2_{1y} + R_1^2 S^2_{1x} + R_2^2 S^2_{1z} - 2\left(R_1 S_{1xy} - R_2 S_{1yz} + R_1 R_2 S_{1xz}\right)\right]$$

$$S^2_{2R_{0i}^*} = \left[S^2_{2yi} + R_1^2 S^2_{2xi} + R_2^2 S^2_{2zi} - 2\left(R_1 S_{2xyi} - R_2 S_{2yzi} + R_1 R_2 S_{xzi}\right)\right].$$

Following the procedure above, it is easy to verify that

$$MSE\left(t_2^*\right) = \lambda S^2_{1R_{00}^*} + \sum_{i=1}^{N} \lambda_{2i} S^2_{2R_{00i}^*} \tag{9}$$

$$MSE\left(t_3^*\right) = \lambda S^2_{1P_0^*} + \sum_{i=1}^{N} \lambda_{2i} S^2_{2P_{0i}^*} \tag{10}$$

$$MSE\left(t_4^*\right) = \lambda S^2_{1p_{00}^*} + \sum_{i=1}^{N} \lambda_{2i} S^2_{2P_{00i}^*} \tag{11}$$

where

$$S^2_{1R_{00}^*} = \left[S^2_{1y} + R_1^2 S^2_{1x} + R_2^2 S^2_{1z} - 2\left(R_1 S_{1xy} + R_2 S_{1yz} - R_1 R_2 S_{1xz}\right)\right]$$

$$S^2_{2R_{00i}^*} = \left[S^2_{2yi} + R_1^2 S^2_{2xi} + R_2^2 S^2_{2zi} - 2\left(R_1 S_{2xyi} + R_2 S_{2yzi} - R_1 R_2 S_{xzi}\right)\right]$$

$$S^2_{1P_0^*} = \left[S^2_{1y} + R_1^2 S^2_{1x} + R_2^2 S^2_{1z} + 2\left(R_1 S_{1xy} + R_2 S_{1yz} + R_1 R_2 S_{1xz}\right)\right]$$

$$S^2_{2P_{0i}^*} = \left[S^2_{2yi} + R_1^2 S^2_{2xi} + R_2^2 S^2_{2zi} + 2\left(R_1 S_{2xyi} + R_2 S_{2yzi} + R_1 R_2 S_{xzi}\right)\right]$$

$$S^2_{1P_{00}^*} = \left[S^2_{1y} + R_1^2 S^2_{1x} + R_2^2 S^2_{1z} + 2\left(R_1 S_{1xy} - R_2 S_{1yz} - R_1 R_2 S_{1xz}\right)\right]$$

$$S^2_{2P_{00i}^*} = \left[S^2_{2yi} + R_1^2 S^2_{2xi} + R_2^2 S^2_{2zi} + 2\left(R_1 S_{2xyi} - R_2 S_{2yzi} - R_1 R_2 S_{xzi}\right)\right]$$

REMARK 1: The MSE of t_1^* depends on the unknown population variances $S^2_{1R_0^*}$ and $S^2_{2R_{0i}^*}$ whose values may be obtained from the data from a past survey. When this is not feasible, the data from a pilot survey could be used to estimate these parameters. However, the values of $S^2_{1R_0^*}$ and $S^2_{2R_{0i}^*}$ from these estimates will depart from the actual population variances depending on how close our estimates is to the true values of $S^2_{1R_0^*}$ and $S^2_{2R_{0i}^*}$. Similar explanation goes for (9), (10) and (11).

5. Efficiency comparison

We compare the MSE of the suggested method of estimation given in (8), (9), (10) and (11) with that of Singh's (1965, 1967) estimators. We will have the conditions as follows:

(i) $$MSE\left(t_1^*\right) - MSE\left(t_1\right) \le 0$$
$$\lambda\left(S^2_{1R_0^*} - S^2_{1R}\right) + \sum_{i=1}^{N} \lambda_{2i} S^2_{2R_{0i}^*} \le 0$$

(ii) $$MSE\left(t_2^*\right) - MSE\left(t_2\right) \le 0$$
$$\lambda\left(S^2_{1R_{00}^*} - S^2_{2R}\right) + \sum_{i=1}^{N} \lambda_{2i} S^2_{2R_{00i}^*} \le 0$$

(iii) $$MSE\left(t_3^*\right) - MSE\left(t_3\right) \le 0$$
$$\lambda\left(S^2_{1P_0^*} - S^2_{1P}\right) + \sum_{i=1}^{N} \lambda_{2i} S^2_{2P_{0i}^*} \le 0$$

(iv) $$MSE\left(t_4^*\right) - MSE\left(t_4\right) \le 0$$
$$\lambda\left(S^2_{1P_{00}^*} - S^2_{2P}\right) + \sum_{i=1}^{N} \lambda_{2i} S^2_{2P_{00i}^*} \le 0$$

where

$$S^2_{1R} = \left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} - R_2 S_{yz} + R_1 R_2 S_{xz}\right)\right]$$
$$S^2_{2R} = \left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} + R_2 S_{yz} - R_1 R_2 S_{xz}\right)\right]$$
$$S^2_{1P} = \left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 + 2\left(R_1 S_{xy} + R_2 S_{yz} + R_1 R_2 S_{xz}\right)\right]$$
$$S^2_{2P} = \left[S_y^2 + R_1^2 S_x^2 + R_2^2 S_z^2 - 2\left(R_1 S_{xy} - R_2 S_{yz} - R_1 R_2 S_{xz}\right)\right]$$

When these conditions are satisfied, the suggested method of estimation will be more efficient than Singh (1965, 1967) estimators.

6. Optimum sampling and sub-sampling fractions

In this section, we deduce the optimum sampling and sub-sampling fractions that will minimize the MSEs given by equations (8), (9), (10) and (11) for a specified cost and variance.

Consider the cost function

$$C = nc_1 + c_2 \sum_{i=1}^{n} m_i + c_3 \sum_{i=1}^{n} M_i$$

Where C is the total cost; c_1 and c_2 are the fixed cost per FSU and SSU respectively while c_3 is the cost of listing per SSU in a selected FSU.

However, the total cost C depends on the selected units which always vary from sample to sample, hence, we shall use the average cost denoted C'

$$\begin{aligned} C' &= nc_1 + c_2 \frac{n}{N} \sum_{i=1}^{N} m_i + c_3 \frac{n}{N} \sum_{i=1}^{N} M_i \\ &= n\left(c_1 + c_3 \overline{M}\right) + nc_2 \overline{m} \end{aligned} \tag{12}$$

We shall now proceed to determine the best n and $\overline{m}$ which will minimize (8) subject to the cost constraints. To do this, MSE of t_1^* is rearranged as follows:

Let $f_{2i} = \dfrac{m_i}{M_i} = f_2$ such that $f_2 = \dfrac{\overline{m}}{\overline{M}}$ and then $m_i = \dfrac{M_i}{\overline{M}} \overline{m}$.

If we substitute the values of m_i and f_2 in (8) and rearranging, we have

$$MSE\left(t_1^*\right) = \frac{\lambda_1 S_{1R_0^*}^2 - S_{2R_0^*}^2}{n} + \frac{\overline{M} S_{2R_0^*}^2}{n\overline{m}} \tag{13}$$

Where $\lambda_1 = 1 - f$ and $S_{2R_0^*}^2 = \dfrac{1}{N} \sum_{i=1}^{N} M_i S_{2R_{0i}^*}^2$.

Using Lagrange multiplier we construct the function

$$F = \frac{\lambda_1 S_{1R_0^*}^2 - S_{2R_0^*}^2}{n} + \frac{\overline{M} S_{2R_0^*}^2}{n\overline{m}} + \theta\left[n\left(c_1 + c_3 \overline{M}\right) + c_2 n\overline{m} - C'\right]$$

Differentiating F partially with respect to n and $\overline{m}$, equating the derivatives to zero, and simplifying, we obtain the two equations below

$$\lambda_1 \overline{m} S^2_{1R_0^*} - \overline{m} S^2_{2R_0^*} + \overline{M} S^2_{2R_0^*} = \theta\left(\overline{m} c_1 n^2 + \overline{m} c_3 n^2 \overline{M} + c_2 \overline{m}^2 n^2\right) \tag{14}$$

$$\overline{M} S^2_{2R_0^*} = \theta c_2 \overline{m}^2 n^2 \tag{15}$$

Subtracting (15) from (14) and solving for θ, we obtain

$$\theta = \frac{\lambda_1 S^2_{1R_0^*} - S^2_{2R_0^*}}{c_1 n^2 + c_3 n^2 \overline{M}}$$

Substituting the value of θ in (15) and solving for $\overline{m}$, we obtain the optimum $\overline{m}$ as

$$\overline{m}_{opt} = \sqrt{\frac{S^2_{2R_0^*}\left(c_1 + c_3 \overline{M}\right)}{c_2\left(\lambda_1 S^2_{1R_0^*} - S^2_{2R_0^*}\right)}}.$$

To obtain the optimum first stage sample size, we substitute $\overline{m}_{opt}$ into (12) and solve for n, this results in

$$n_{opt} = \frac{C'}{c_1 + c_3 \overline{M} + c_2 \sqrt{\frac{S^2_{2R_0^*}\left(c_1 + c_3 \overline{M}\right)}{c_2\left(\lambda_1 S^2_{1R_0^2} - S^2_{2R_0^*}\right)}}}$$

If, however, the MSE is specified, say M_0, then $\overline{m}_{opt}$ is substituted in (13). By equating (13) to M_0 and solving for n, we get

$$n_{opt} = \frac{\left(\lambda_1 S^2_{1R_0^*} - S^2_{2R_0^*}\right)\left[c_1 + (c_2 + c_3)\overline{M}\right]}{M_0\left(c_1 + c_3 \overline{M}\right)}.$$

REMARK 2: following the approach above, it is easy to show that the optimum n and $\overline{m}$ for the estimators t_2^*, t_3^* and t_4^* would yield similar results but with $S^2_{1R_0^*}$ and $S^2_{2R_0^*}$ now being replaced by $S^2_{1R_{00}^*}$ and $S^2_{2R_{00}^*}$; $S^2_{1P_0^*}$ and $S^2_{2P_0^*}$; $S^2_{1P_{00}^*}$ and $S^2_{2P_{00}^*}$ respectively.

7. Numerical illustration and discussion

To illustrate the properties of the suggested estimators for estimating the population mean $\bar{Y}$, we consider a real data set whose detail description is given below.

One of the States in Nigeria (Kwara State) is divided into 16 Local Government Areas (LGA) and a random sample of four LGAs is selected. Within each selected LGA, a random sample of m_i (for $m_i = 3$, 5, 10 and 20; $i = 1,2,...n$) primary schools was selected from M_i. Information on the enrolment of pupils for the sessions 2008/2009 (y), 2007/2008 (x) and 2006/2007 (z) was obtained among other characteristics of interest. The summary of the data is presented below. $N = 16$, $n = 4$, $S_y^2 = 7658.25$, $S_x^2 = 4273$, $S_z^2 = 2048.667$, $S_{xy} = 5692.833$, $S_{yz} = 3941.333$, $S_{xz} = 2958$. Other details are provided in table 1.

The MSEs of the estimators t_1 through t_4 are 1408.83, 1208.42, 12087.91 and 1464.53 respectively, and that of t_1^*, t_2^*, t_3^* and t_4^* are given in Table 2.

Table 1. Summary of the data

LG	M_i	m_i	$\bar{x}_i$	$\bar{y}_i$	$\bar{z}_i$	S_{2xi}^2	S_{2yi}^2	S_{2zi}^2	S_{2xyi}	S_{2yzi}	S_{2xzi}
	K=20										
1	78	4	58.25	68.750	45.25	400.25	464.92	250.25	411.08	327.75	314.92
2	139	7	85.86	101.00	70.57	1046.48	1037.33	797.29	1033.5	884.17	894.93
3	95	5	32.00	40.600	24.00	332.50	473.30	249.50	396.50	343.25	287.25
4	55	3	95.33	109.67	83.33	424.33	364.33	241.33	393.17	295.67	319.33
						K=10					
1	78	8	53.25	62.880	42.38	297.64	311.55	235.13	301.04	268.34	261.32
2	139	14	101.5	118.14	84.93	617.35	740.75	376.23	668.69	511.55	470.5
3	95	10	45.50	54.200	35.50	311.83	394.62	223.83	347.33	289.33	261.17
4	55	6	94.50	112.83	79.83	492.30	667.77	418.57	561.50	510.97	452.10
						K=5					
1	78	16	49.060	60.130	39.44	169.66	222.52	168.66	188.53	183.94	165.84
2	139	28	97.430	113.32	81.71	384.58	553.63	384.58	493.08	438.98	413.83
3	95	20	42.900	53.850	33.00	227.26	306.56	227.26	283.51	250.79	240.89
4	55	12	106.33	122.58	92.17	413.61	563.17	413.61	519.88	455.76	454.76

Table 2. MSE for the different estimators

Estimators	MSE			
	$k=20$	$k=10$	$k=5$	$k=3$
t_1^*	86726.540	28348.6100	11569.740	1272.51
t_2^*	118966.59	38066.3400	11699.410	1096.73
t_3^*	753023.59	2702242.91	100808.02	9768.25
t_4^*	62875.180	27144.8300	8512.7200	1215.96

From table 2, we see that the MSEs of the estimators t_1^*, t_2^*, t_3^* and t_4^* is larger than that of t_1 through t_4 when $k=20$, 10 and 5; therefore, we can say that t_1^*, t_2^*, t_3^* and t_4^* are less efficient than t_1 through t_4. However, for $k=3$ the suggested estimator is more efficient than Singh (1965, 1967) estimators since they have smaller MSE.

This is not a surprising result, since potential sample units do tend to be more similar (show a high degree of inner homogeneity) if they are formed by putting elements, which are physically (or geographically) close to each other than if they are far apart. This argument is "almost always" true. The word "almost always" is a technical term which means that the event in question is not impossible theoretically, but that the probability of it actually happening in practice is close to zero.

Furthermore, in most practical situations, a complete list of sampling units (frame) from which drawing our sample may be impossible; even if such a frame exists, it will be uneconomical to obtain information from a sample of elements of the population scattered all over the region. Thus, to overcome these problems (absence or incomplete list; frame construction; cost consideration; organization of the survey and so on) our suggested method of estimation will be found to be more useful and rewarding than Singh's (1965, 1967) estimators.

Moreover, it could also be observed from table 2 that as k decreases i.e. increasing the sub-sampling size, the MSEs of the suggested estimators increases substantially at the rate of about 32% and 45%. Thus, we may say that the effect of increasing the sub-sampling size is justified through precision.

Finally, we can conclusively say that our suggested method of estimation is not only found to be more applicable in many situations but also more efficient than Singh's (1965, 1967) estimators when the sub-sampling size is large. It should be noted that the conclusion above on efficiency is based on the data set used for the numerical illustration.

Acknowledgement

The author is grateful to the referees for their comments and suggestions which have greatly helped in improving the manuscript.

REFERENCES

ABDUL, H. and JAVID, S. (2010): A family of ratio estimators for population mean in extreme ranked set sampling using two auxiliary variables. SORT 34 (1) January-June 2010, 45–64

ABU-DAYYEH, W.A., AHMED, M.S., AHMED, R.A., and MUTTLAK, H.A. (2003): Some estimators of a finite population mean using auxiliary information, Applied Mathematics and Computation 139: 287–298.

BANDYOPADHYAY, S. (1980): Improved ratio and product estimators. Sankhya Series C, 42(2), 45-49.

Cochran, W.G. (1963): Sampling techniques. John Wiley & Sons, New York.

DES, R. (1965): On a method of using multi-auxiliary information in sample surveys. J. Amer. Statist. Assoc. 60, 270–277

DIANA, G. and PERRI, P.F. (2007): Estimation of finite population mean using multi-auxiliary information. METRON - International Journal of Statistics, vol. LXV, n. 1, pp. 99–112.

HOSSAIN, M.I., and AHMED, M. S. (2001): A Class of Predictive Estimators in Two-Stage Sampling Using Auxiliary Information. Information and Management Sciences, Volume 12, Number 1, pp.49–55

KADILAR, C. and CINGI, H. (2004): Estimator of a population mean using two auxiliary variables in simple random sampling, International Mathematical Journal, 5, 357–367.

KADILAR, C. and CINGI, H. (2005): A new estimator using two auxiliary variables, Applied Mathematics and Computation, 162, 901–908.

KIREGYERA, B. (1980): A Chain Ratio-Type Estimator in Finite Population Double Sampling using two Auxiliary Variables. Metrika, 27: 217–223.

KIREGYERA, B. (1984): A Regression-Type Estimator using two Auxiliary Variables and Model of Double sampling from Finite Populations. Metrika, 31: 215–226.

MISHRA, G. and ROUT, K. (1997): A regression estimator in two-phase sampling in presence of two auxiliary variables, Metron 55(1–2), 177–186.

MUHAMMAD, H., NAQVI H., and MUHAMMAD, Q. S. (2010): Some New Regression Types Estimators in Two Phase Sampling. World Applied Sciences Journal 8 (7): 799–803.

MURTHY, M. N. (1964): Product method of estimation. Sankhya, A, 26, 69–74.

OLKIN, I. (1958): Multivariate ratio estimation for finite populations, Biometrika 45, 154–165.

PERRI, P. F. (2005): Combining two auxiliary variables in ratio-cum-product type estimators, Proceedings of Italian Statistical Society, Intermediate Meeting on Statistics and Environment, Messina, 21–23, 193–196.

RAO, P. S. R. S and MUDHOLKAR, G.S. (1967): Generalized multivariate estimator for the mean of finite populations, J. Amer. Statist. Assoc. 62, 1009–1012.

SAHOO, L.N. and PANDA, P. (1997): A class of estimators in two-stage samplings with varying probabilities. South African Statistics, J. 31, 151–160.

SAHOO, L.N. and PANDA, P. (1999): A class of estimators using auxiliary information in two-stage sampling. Australian and New Zealand Journal of Statistics, 41(4), 405–410

SAMIUDDIN, M. and HANIF, M. (2007): Estimation of population mean in single phase and two phase sampling with or with out additional information. Pak. J. Statist., 23 (2): 99–118.

SINGH, G.N. and UPADHYAYA, L.N (1995): A Class of Modified Chain-Type Estimators using Two Auxiliary Variables in Two-Phase Sampling. Metron, Vol. 53, N 3–4, 117–125.

SINGH, H.P. and ESPEJO, M.R. (2007): Double sampling ratio-product estimator of a finite population mean in sample surveys. J. Appl. Statist., 34: 71–85.

SINGH, H. P., SINGH, R., ESPEJO, M. R., PINEDA, M. D., and NADARAJAN, S. (2005): On the efficiency of the dual to ratio-cum-product estimator. Mathematical Proceedings of the Royal Irish Academy, 105A (2), 51–56.

SINGH, M. P. (1965): On the estimation of ratio and product of the population parameters. Sankhya, B, 27, 321–328.

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

SINGH, M. P. (1967b): Multivariate product method of estimation for finite populations, Journal of the Indian Society of Agricultural Statistics, 31, 375–378.

SINGH, R., CHAUHAN, P and SAWAN, N. (2007): A family of estimators for estimating population mean using known correlation coefficient in two phase sampling. Statistics in Transition, 8 (1): 89–96.

SRIVENKATARAMANA, T. (1980): A dual to ratio estimator in sample surveys. Biometrika, 67, 199–204.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASHOK, C. (1984): Sampling theory of surveys with applications. Iowa State University Press, USA.

TAILOR R. and 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, vol. 10, pp. 15–24.

TRACY, D. S., SINGH, H. P., and SINGH, R. (1996): An alternative to the ratio-cum-product estimator in sample surveys, Journal of Statistical Planning and Inference, 53, 375–387.

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

UPADHYAYA, L.N. and SINGH, H.P. (2003): A Note on the Estimation of Mean Using Auxiliary Information. Statistics in Transition, Vol. 6, No. 4, 571–575.

WOLTER, K. M. (1985): Introduction to Variance Estimation, Springer-Verlag.

STATISTICS IN TRANSITION new series, October 2010
Vol.11, No. 2, pp. 267-285

RECURSIVE OPTIMAL ESTIMATION IN SZARKOWSKI ROTATION SCHEME

Jacek Wesołowski[1]

ABSTRACT

In late 90ties Szarkowski observed that under the rotation pattern typical for the Labour Force Survey the recursion for the optimal estimator of the mean on a given occasion has to use estimators and observations only from three last occasions. Since the fundamental work of Patterson (1950) it had been known that for rotation patterns with "holes" it is a difficult problem to determine the depth of such recursion formulas. Under special assumptions the problem has been settled only recently in Kowalski and Wesołowski (2010). In the present paper it is shown that these assumptions are always satisfied in the case of the Szarkowski rotation pattern 110011. Moreover, explicit formulas for the coefficients of recursion are derived.

1. Introduction

Andrzej Szarkowski passed away in June 2003. He was a creative and passionate statistician with considerable mathematical background. He devoted his talents to the Labour Force Survey (LFS) conducted by the Central Statistical Office in Poland taking care about mathematical methodology of this survey for more than 10 years, just from its beginning in 1992. For details on development of methodology of this survey in Poland, see Szarkowski and Witkowski (1994) and Popiński (2006). In particular, in 1993 Szarkowski introduced in the LFS a rotation pattern 110011. One of the issues related to this approach he was very concerned about was the recurrence form of optimal linear estimators of mean on every occasion under this pattern. In late nineties he studied Patterson (1955) paper, where the rotation pattern with no holes had been thoroughly treated. However, it appeared to be of not much help since a real challenge is posed by the HOLES! On the basis of intensive numerical experiments Szarkowski conjectured that the pattern

[1] Główny Urząd Statystyczny and Politechnika Warszawska, Warszawa, POLAND
e-mail: wesolo@mini.pw.edu.pl

110011 forces the recurrence to move THREE STEPS back when the correlation is exponential in the occasions span and SEVEN STEPS back when there are no restrictions on correlations. For a while we had sought together a mathematical explanation of this phenomena with no luck.

It took more than ten years to answer in affirmative Szarkowski's THREE STEPS conjecture. The explanation is given in the present paper. It is based on a general approach described in Kowalski and Wesołowski (2010) (KW in the sequel). Earlier the problem for rotation schemes with singleton holes was solved in Kowalski (2009) (particular cases of 1011 and 1101 rotation patterns were covered even earlier, in Ciepiela (2004)). In Section 2 the general approach from KW, which is based on TWO ASSUMPTIONS, is adjusted to a setup with a single hole of any size h, that is for the rotation pattern 11...110...011...11. In Section 3 we prove that for the Szarkowski scheme these TWO ASSUMPTIONS **are necessarily satisfied** and thus the general procedure works. Moreover, explicit formulas for the coefficients of the recursion are derived. In Section 4 we give proofs of lemmas which are used in Section 3 to derive the main result.

Szarkowski's SEVEN STEPS conjecture remains open. Even in the case of one singleton hole it is not known how far back one has to go in the recursion formula.

2. General method

Consider a doubly-infinite matrix of random variables (X_{ij}), $i, j \in \mathbb{Z}$. Index i identifies a unit and index j is an occasion number (time). We assume that for any $j \in \mathbb{Z}$ we have

$$\mathbb{E}\, X_{i,j} = \mu_j, \quad \text{for all } i \in \mathbb{Z},$$

and, without loss of generality we assume that $\mathbb{V}\mathrm{ar}(X_{i,j}) = 1$ for all $i, j \in \mathbb{Z}$. The correlation structure is described as follows

$$\mathbb{C}\mathrm{orr}(X_{i,j}, X_{k,l}) = I(k = i)\rho^{|j-l|}.$$

Fix natural numbers n and h and consider a sequence of random vectors $\underline{X}_j = (X_{j,j}, \ldots, X_{j,j+n+h-1})$, $j \in \mathbb{Z}$. Note that $C = \mathbb{C}\mathrm{ov}\, \underline{X}_j$ is an $(n+h) \times (n+h)$ matrix with all entries equal zero except the entries just above the diagonal which are all equal ρ. Moreover

$$\mathbb{C}\mathrm{ov}(\underline{X}_j, \underline{X}_k) = C^{|k-j|}$$

and C^j is a matrix with all entries equal zero except the jth over diagonal with all entries equal ρ^j when $j \leq n + h - 1$ and it is a zero matrix when $j > n + h - 1$.

A rotation pattern is any vector $(\epsilon_1, \ldots, \epsilon_{n+h})$ with 0-1 entries such that $\epsilon_1 = \epsilon_{n+h} = 1$ and there are exactly h zeros among the entries. Let $p-1$ denotes the dimension of the largest zero subvector of subsequent entries in the rotation pattern.

We modify vectors $\underline{X}_j$ into

$$\underline{Y}_j = (X_{j,k}\epsilon_{k-j+1},\ k = j, \ldots, j+n+h-1), \quad j \in \mathbb{Z}.$$

For a given $j \in \mathbb{Z}$ let $\widehat{\mu}_j$ denote the BLUE of μ_j based on $\underline{Y}_l$, $l \leq j$.

We study the recurrence formula for the BLUE estimators of the following form

$$\widehat{\mu}_t = a_1\,\widehat{\mu}_{t-1} + \ldots + a_p\,\widehat{\mu}_{t-p} + \langle \underline{r}_0, \underline{Y}_t\rangle + \langle \underline{r}_1, \underline{Y}_{t-1}\rangle + \ldots + \langle \underline{r}_p, \underline{Y}_{t-p}\rangle,$$

for any $t \in \mathbb{Z}$, where the parameters $a_1, \ldots, a_p \in \mathbb{R}$ and $\underline{r}_0, \underline{r}_1, \ldots, \underline{r}_p \in \mathbb{R}^{n+h}$ are to be identified. Here we use the symbol $\langle \underline{a}, \underline{b}\rangle$ to denote the scalar product of vectors $\underline{a} = (a_1, \ldots, a_d)$ and $\underline{b} = (b_1, \ldots, b_d)$, that is $\langle \underline{a}, \underline{b}\rangle = \sum_{i=1}^{d} a_i b_i$. Note that the parameters are assumed to be constant, i.e. they do not depend on t.

Note that, alternatively, $\widehat{\mu}_t$ can be defined as optimal unbiased linear estimator $\sum_{s\leq t} \langle \underline{w}_s, \underline{X}_s\rangle$, with additional constraints

$$\underline{w}_{s,j}(1-\epsilon_j) = 0, \quad j = 1, \ldots, n+h, \quad s \leq t, \tag{1}$$

imposed by the holes in the rotation pattern. Therefore the above recursion can be written in the form

$$\widehat{\mu}_t = a_1\,\widehat{\mu}_{t-1} + \ldots a_p\,\widehat{\mu}_{t-p} + \langle \underline{r}_0, \underline{X}_t\rangle + \langle \underline{r}_1, \underline{X}_{t-1}\rangle + \ldots + \langle \underline{r}_p, \underline{X}_{t-p}\rangle. \tag{2}$$

Note that (1) forces respective entries of vectors $\underline{r}_j$, $j = 0, \ldots, h+1$, to be equal zero.

Under certain assumptions (see ASSUMPTION 1 and 2, below) there exists a general algorithm, described in KW (see also Kowalski (2010)), which completely solves the problem. It is rather complicated. Here we describe it in the case of a single hole of any size h in the rotation pattern (thus $p = h+1$). More precisely we assume that the rotation patterns have a form $[1, 1, \ldots, 1, 0, 0, \ldots, 0, 0, 1, 1, \ldots, 1]$ where the zeros occur at places $s+1, s+2, \ldots, s+h$, for arbitrary s satisfying $1 \leq s < n$.

Recall that the Chebyshev polynomials of the first kind (T_n) *are defined through a three step recurrence*

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \ldots$$

and $T_0 \equiv 1$, $T_1(x) = x$.

Consider a polynomial P *of degree* p *defined by*

$$P(x) = 1 - \rho^2 + (n + h - 1)(1 + \rho^2 - 2\rho x) + \\ -(1 + \rho^2 - 2\rho x)^2 \text{ tr } \left(\mathbf{T}_p(x)\mathbf{R}_p^{-1}(\rho)\right), \tag{3}$$

where $\mathbf{T}_p(x)$ *is a* $h \times h$ *symmetric matrix polynomial*

$$\mathbf{T}_p(x) = \begin{bmatrix} T_0(x) & T_1(x) & T_2(x) & \dots & T_{p-3}(x) & T_{p-2}(x) \\ T_1(x) & T_0(x) & T_1(x) & \dots & T_{p-4}(x) & T_{p-3}(x) \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ T_{p-3}(x) & T_{p-4}(x) & T_{p-5}(x) & \dots & T_0(x) & T_1(x) \\ T_{p-2}(x) & T_{p-3}(x) & T_{p-4}(x) & \dots & T_1(x) & T_0(x) \end{bmatrix} \tag{4}$$

and $\mathbf{R}_p$ *is a* $h \times h$ *invertible constant tridiagonal matrix*

$$\mathbf{R}_p = \begin{bmatrix} 1+\rho^2 & \rho & 0 & \dots & 0 & 0 \\ \rho & 1+\rho^2 & \rho & \dots & 0 & 0 \\ 0 & \rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ 0 & 0 & 0 & \dots & 1+\rho^2 & \rho \\ 0 & 0 & 0 & \dots & \rho & 1+\rho^2 \end{bmatrix} \tag{5}$$

ASSUMPTION 1: All the roots $x_1, \dots, x_p$ **(real or complex) of the polynomial** P **defined through** (3)**,** (4)**,** (5) **are distinct and do not belong to the interval** $[-1, 1]$**.**

Under ASSUMPTION 1 *equation*

$$d + \frac{1}{d} = 2x_i,$$

has exactly one solution d_i *such that* $|d_i| < 1$*,* $i = 1, \dots, p$*.*

Let $\underline{d} = [d_1, \dots, d_p]^T$*. Consider linear system*

$$S(\underline{d})\underline{c} = (1 - \rho^2)\underline{e}, \tag{6}$$

where $S(\underline{d})$ *is a* $(p+1)p \times p^2$ *matrix of the form*

$$S(\underline{d}) = \begin{bmatrix} G(d_1) & G(d_2) & G(d_3) & \dots & G(d_p) \\ H(d_1) & 0 & 0 & \dots & 0 \\ 0 & H(d_2) & 0 & \dots & 0 \\ 0 & 0 & H(d_3) & \dots & 0 \\ 0 & 0 & 0 & \dots & H(d_p) \end{bmatrix}$$

with $p \times p$ *blocks* $G(d_i)$*,* $H(d_i)$*,* $i = 1, \dots, p$ *defined as*

$$G(v) = \begin{bmatrix} g_0(v) & g_1(v) & g_1(v) & g_1(v) & \dots & g_1(v) & g_1(v) \\ g_1(v) & 1 & -v\rho & 0 & \dots & 0 & 0 \\ g_1(v) & 0 & 1 & -v\rho & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ g_1(v) & 0 & 0 & 0 & \dots & 1 & -v\rho \\ g_1(v) & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

$$H(v) = \begin{bmatrix} h_0(v) & h_1(v) & h_1(v) & \dots & h_1(v) & h_1(v) & h_1(v) \\ h_1(v) & v(1+\rho^2) & -v^2\rho & \dots & 0 & 0 & 0 \\ h_1(v) & -\rho & v(1+\rho^2) & \dots & 0 & 0 & 0 \\ h_1(v) & 0 & -\rho & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ h_1(v) & 0 & 0 & \dots & v(1+\rho^2) & -v^2\rho & 0 \\ h_1(v) & 0 & 0 & \dots & -\rho & v(1+\rho^2) & -v^2\rho \\ h_1(v) & 0 & 0 & \dots & 0 & -\rho & v(1+\rho^2) \end{bmatrix},$$

and

$$g_1(v) = 1 - v\rho, \quad g_0(v) = 1 - \rho^2 + (n+h-1)g_1(v),$$

$$h_1(v) = (1-v\rho)(v-\rho), \quad h_0(v) = v(1-\rho^2) + (n+h-1)h_1(v).$$

The unknown vector $\underline{c}$ *has the following structure*

$$\underline{c} = \left[\underline{c}_1,\ \underline{c}_2,\ \dots,\ \underline{c}_p\right]^T,$$

where $\underline{c}_i = [c_{0,i}, c_{1,i}, \dots, c_{h,i}]^T$, $i = 0, 1, \dots$. *Finally,* $\underline{e}$ *is the* $(p+1)p$*-dimensional unit vector* $\underline{e} = [1, 0, \dots, 0]^T$.

ASSUMPTION 2. Linear system (6) **has a unique solution.**

Under ASSUMPTION*s* 1 *and* 2 *the recurrence* (2) *holds with parameters* $a_1, \dots, a_p$ *and* $\underline{r}_0, \dots, \underline{r}_p$ *defined as follows:*

- *The linear system*

$$x_1 d_i^{p-1} + x_2 d_i^{p-2} + \dots + x_{p-1} d_i + x_p = d_i^p, \quad i = 1, \dots, p$$

has a unique solution, which equals $\underline{a} = [a_1, \dots, a_p]^T$, *that is*

$$\underline{a} = \begin{bmatrix} d_1^{p-1} & d_1^{p-2} & \dots & d_1 & 1 \\ d_2^{p-1} & d_2^{p-2} & \dots & d_2 & 1 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ d_p^{p-1} & d_p^{p-2} & \dots & d_p & 1 \end{bmatrix}^{-1} \begin{bmatrix} d_1^p \\ d_2^p \\ \vdots \\ d_p^p \end{bmatrix}.$$

• *For any* $i = 1, \ldots, p$ *let* D_i *be an* $(n+h) \times (p+1)p$ *matrix defined as*

$$D_i = \begin{bmatrix}
d_1^i & 0 & 0 & \ldots & 0 & d_2^i & 0 & 0 & \ldots & 0 & \ldots & d_p^i & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\
d_1^i & 0 & 0 & \ldots & 0 & d_2^i & 0 & 0 & \ldots & 0 & \ldots & d_p^i & 0 & 0 & \ldots & 0 \\
d_1^i & d_1^i & 0 & \ldots & 0 & d_2^i & d_2^i & 0 & \ldots & 0 & \ldots & d_p^i & d_p^i & 0 & \ldots & 0 \\
d_1^i & 0 & d_1^i & \ldots & 0 & d_2^i & 0 & d_2^i & \ldots & 0 & \ldots & d_p^i & 0 & d_p^i & \ldots & 0 \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\
d_1^i & 0 & 0 & \ldots & d_1^i & d_2^i & 0 & 0 & \ldots & d_2^i & \ldots & d_p^i & 0 & 0 & \ldots & d_p^i \\
d_1^i & 0 & 0 & \ldots & 0 & d_2^i & 0 & 0 & \ldots & 0 & \ldots & d_p^i & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\
d_1^i & 0 & 0 & \ldots & 0 & d_2^i & 0 & 0 & \ldots & 0 & \ldots & d_p^i & 0 & 0 & \ldots & 0
\end{bmatrix}$$

where the first s *rows are identical, then the rows with numbers* $s+1$, $s+2$, …, $s+h$, *associated to the hole in the rotation pattern, are perturbed in a regular manner, and the last rows with numbers* $s+h+1$, $s+h+2$, …, $n+h$ *are again identical and the same as the first* s *rows.*

Let Δ *be a* $(n+h) \times (n+h)$ *diagonal matrix defined as*

$$\Delta = (\mathrm{Id} - CC^T)^{-1}.$$

All the elements of the diagonal of Δ *are equal* $(1-\rho^2)^{-1}$ *except the last one which equals 1.*

Let

$$V_i = \Delta(D_i - CD_{i+1}), \quad i = 0, 1, \ldots$$

Let $\underline{c} = \underline{c}(\underline{d})$ *be the unique solution of* (6), *which by ASSUMPTION 2 is guaranteed to exist. Then, denoting additionally* $a_0 = -1$, *we have*

$$\underline{r}_0 = V_0\, \underline{c}(\underline{d}), \quad \underline{r}_j = \left(V_j + \sum_{i=0}^{j-1} (a_i C^T - a_{i+1}\mathrm{Id}) V_{j-1-i} \right) \underline{c}(\underline{d}) \tag{7}$$

for $j = 1, \ldots, p$.

Thus the problem has a solution provided the ASSUMPTIONs 1 and 2 are satisfied. It was proved in KW that the ASSUMPTIONs are always satisfied when $p = 0$ (no holes) or $p = 1$ (singleton hole; actually, any number of singleton holes has been allowed). Intensive numerical experiments provided strong motivation to conjecture that ASSUMPTIONs 1 and 2 are satisfied for any $p \geq 0$. However, proving this conjecture seems to be rather difficult even in the case of a single hole of any size.

3. Szarkowski scheme

Here we concentrate on a special rotation pattern 110011 with a single hole of size 2, called the Szarkowski scheme. As already mentioned, this scheme has been adopted for the Labuor Force Survey in the CSO in Poland. We will prove that under this rotation pattern ASSUMPTIONS 1 and 2 are satisfied. Moreover, we will derive explicit analytic formulas for the parameters of the recurrence a_i, $i = 1, 2, 3$, and $\underline{r}_j$, $j = 0, 1, 2, 3$.

Note that under this pattern $n = 4$, $h = 2$, $p = 3$ and the missing elements are at positions defined by the vectors $\underline{e}_3$ and $\underline{e}_4$ in six-dimensional space $\mathbb{R}^6$. We seek a representation

$$\widehat{\mu}_t = a_1 \widehat{\mu}_{t-1} + a_2 \widehat{\mu}_{t-2} + a_3 \widehat{\mu}_{t-3}$$
$$+\langle \underline{r}_0, \underline{X}_t \rangle + \langle \underline{r}_1, \underline{X}_{t-1} \rangle + \langle \underline{r}_2, \underline{X}_{t-2} \rangle + \langle \underline{r}_3, X_{t-3} \rangle \tag{8}$$

for the BLUE of the mean μ_t on the t-th occasion.

The main result will be preceded by three auxiliary lemmas, proofs of which are postponed to Section 4.

Lemma 1. *Let $\rho \in (-1, 1) \setminus \{0\}$. Then the polynomial*

$$W_3(x) = x^3 + (2 - \rho^2 + 2\rho^4)x + 2(2 + 2\rho^2 + 2\rho^4 + \rho^6)$$

has one real root $x_1 < -2|\rho|$ and two conjugate complex roots x_2 and x_3.

Lemma 2. *For any $\rho \in (-1, 1) \setminus \{0\}$ let*

$$Q(d) = \begin{bmatrix} 5(1 - d\rho)(d - \rho) + d(1 - \rho^2) & (1 - d\rho)(d - \rho) & (1 - d\rho)(d - \rho) \\ (1 - d\rho)(d - \rho) & d(1 + \rho^2) & -d^2\rho \\ (1 - d\rho)(d - \rho) & -\rho & d(1 + \rho^2) \end{bmatrix}.$$

The equation

$$\det Q(d) = 0 \tag{9}$$

has exactly three distinct roots $d_1 = d_1(\rho)$, $d_2 = d_2(\rho)$ and $d_3 = d_3(\rho)$ such that $|d_i| < 1$, $i = 1, 2, 3$. The number d_i is the unique solution of equation

$$-\rho \left(d + \frac{1}{d} \right) = x_i(\rho),$$

where $x_i(\rho)$ is the root of the polynomial W_3, satisfying $|d_i| < 1$, $i = 1, 2, 3$.

The root d_1 is real and the roots d_2, and d_3 are conjugate complex. Moreover,

$$d_1 d_2 d_3 (d_1 + d_2 + d_3) = -d_1 d_2 - d_2 d_3 - d_3 d_1. \tag{10}$$

Lemma 3. *Let $\rho \in (-1,1) \setminus \{0\}$. Let $\alpha = d_1$, $\beta = d_2$ and $\gamma = d_3$ be the roots of* (9) *defined in Lemma* 2. *Let $\underline{e}_1 = (1,0,\ldots,0) \in \mathbb{R}^{12}$. Consider the following 12×9 system of linear equations in $\underline{c}_j = [c_{0,j},\, c_{1,j},\, c_{2,j}]^T$, $j = 1,2,3$,*

$$\mathbb{Q}\,\underline{c} = \begin{bmatrix} Q_0(\alpha) & Q_0(\beta) & Q_0(\gamma) \\ Q(\alpha) & 0 & 0 \\ 0 & Q(\beta) & 0 \\ 0 & 0 & Q(\gamma) \end{bmatrix} \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \underline{c}_3 \end{bmatrix} = (1-\rho^2)\underline{e}_1, \tag{11}$$

where $Q(d)$ is defined in Lemma 2 *and*

$$Q_0(d) = \begin{bmatrix} 5(1-d\rho)+1-\rho^2 & 1-d\rho & 1-d\rho \\ 1-d\rho & 1 & -d\rho \\ 1-d\rho & 0 & 1 \end{bmatrix}.$$

The linear system (11) *has the unique solution*

$$\begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \underline{c}_3 \end{bmatrix} = (1-\rho^2)\widetilde{\mathbb{Q}}^{-1}\widetilde{\underline{e}}_1,$$

where $\widetilde{\underline{e}}_1 = (1,0,\ldots,0) \in \mathbb{R}^9$ and $\widetilde{\mathbb{Q}}$ is 9×9 invertible matrix defined as

$$\widetilde{\mathbb{Q}} = \begin{bmatrix} Q_0(\alpha) & Q_0(\beta) & Q_0(\gamma) \\ \widetilde{Q}(\alpha) & 0 & 0 \\ 0 & \widetilde{Q}(\beta) & 0 \\ 0 & 0 & \widetilde{Q}(\gamma) \end{bmatrix} \tag{12}$$

with

$$\widetilde{Q} = \begin{bmatrix} (1-d\rho)(d-\rho) & d(1+\rho^2) & -d^2\rho \\ (1-d\rho)(d-\rho) & -\rho & d(1+\rho^2) \end{bmatrix}. \tag{13}$$

Now we are ready to formulate and prove the main result of the paper, which completely covers the problem of recursive optimal estimation under the pattern 110011.

Theorem 1. *Let $\rho \in (-1,1) \setminus \{0\}$. Let $\alpha = d_1$, $\beta = d_2$ and $\gamma = d_3$ be the roots of* (9) *defined in Lemma* 2. *Let $\underline{c} = \underline{c}(d)$ be defined as in Lemma* 3.

Under the Szarkowski rotation pattern recurrence (8) *always holds with*

$$a_1 = \alpha+\beta+\gamma, \quad a_2 = -(\alpha\beta+\beta\gamma+\gamma\alpha), \quad a_3 = \alpha\beta\gamma. \tag{14}$$

Denote

$$V = \begin{bmatrix} 1-\rho\alpha & 0 & 0 & 1-\rho\beta & 0 & 0 & 1-\rho\gamma & 0 & 0 \\ 1-\rho\alpha & 1 & 1 & 1-\rho\beta & 1 & 1 & 1-\rho\gamma & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1-\rho\alpha & 0 & 0 & 1-\rho\beta & 0 & 0 & 1-\rho\gamma & 0 & 0 \\ 1-\rho^2 & 0 & 1 & 1-\rho^2 & 0 & 1 & 1-\rho^2 & 0 & 1 \end{bmatrix}$$

Then

$$\underline{r}_i = \frac{1}{1-\rho^2} V D_i \underline{c}, \quad i = 0, 1, 2, \tag{15}$$

with

$$D_0 = \mathrm{Diag}[1, -\alpha\rho, 0, 1, -\beta\rho, 0, 1, -\gamma\rho, 0],$$

$$D_1 = \mathrm{Diag}[\beta+\gamma, -\alpha(\beta+\gamma)\rho, -\rho, \gamma+\alpha, -\beta(\gamma+\alpha)\rho, -\rho, \alpha+\beta, -\gamma(\alpha+\beta)\rho, -\rho],$$

$$D_2 = \mathrm{Diag}[\beta\gamma, -\alpha\beta\gamma\rho, -(\beta+\gamma)\rho, \gamma\alpha, -\alpha\beta\gamma\rho, -(\gamma+\alpha)\rho, \alpha\beta, -\alpha\beta\gamma\rho, -(\alpha+\beta)\rho].$$

and

$$\underline{r}_3 = \frac{1}{1-\rho^2} \widetilde{I} V D_3 \underline{c} \tag{16}$$

with

$$D_3 = -\rho \mathrm{Diag}[\beta\gamma, 0, \beta\gamma, \gamma\alpha, 0, \gamma\alpha, \alpha\beta, 0, \alpha\beta].$$

and

$$\widetilde{I} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Proof of Theorem 1. First, note that $p = 3$ and the matrices $\mathbf{T}_3(x)$ and $\mathbf{R}_3(\rho)$ have the forms

$$\mathbf{T}_3(x) = \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_3(\rho) = \begin{bmatrix} 1+\rho^2 & -\rho \\ -\rho & 1+\rho^2 \end{bmatrix}.$$

Therefore $\det \mathbf{R}_3(\rho) = 1 + \rho^2 + \rho^4$,

$$\mathbf{R}_3^{-1}(\rho) = \frac{1}{1+\rho^2+\rho^4} \begin{bmatrix} 1+\rho^2 & \rho \\ \rho & 1+\rho^2 \end{bmatrix}$$

and

$$\operatorname{tr}\,(\mathbf{T}_3(x)\,\mathbf{R}_3(\rho)) = \frac{2(1+\rho^2+x\rho)}{1+\rho^2+\rho^4}.$$

Consequently the polynomial P has the form

$$P(x) = \frac{-8x^3\rho^3 - 2x(2\rho+\rho^3-2\rho^5) + 2(2+2\rho^2+2\rho^4+\rho^6)}{1+\rho^2+\rho^4}$$

Observe that

$$P\left(-\frac{x}{2\rho}\right) = \frac{W_3(x)}{1+\rho^2+\rho^4},$$

where polynomial W_3 is defined in Lemma 1. By Lemma 1 polynomial W_3 has one real root less than -2ρ and two complex roots. Therefore polynomial P has one real root outside interval $[-1,1]$ and two complex roots. Hence ASSUMPTION 1 is satisfied.

To show that ASSUMPTION 2 also holds we note that the matrix $S = S(d)$ in (6) has dimensions 12×9. Moreover, (6) is identical to (11). Now, from Lemmas 2 and 3 it follows that ASSUMPTION 2 is also satisfied.

The coefficients a_1, a_2 and a_3 solve the Vandermonde linear system

$$a_1 d_i^2 + a_2 d_i + a_3 = d_i^3, \quad i = 1,2,3.$$

Therefore

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} d_1^2 & d_1 & 1 \\ d_2^2 & d_2 & 1 \\ d_3^2 & d_3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} d_1^3 \\ d_2^3 \\ d_3^3 \end{bmatrix}$$

$$= -\frac{\begin{bmatrix} d_2-d_3 & d_3-d_1 & d_1-d_2 \\ d_3^2-d_2^2 & d_1^2-d_3^2 & d_2^2-d_1^2 \\ d_2(d_2-d_3)d_3 & d_1(d_3-d_1)d_3 & d_1(d_1-d_2)d_2 \end{bmatrix} \begin{bmatrix} d_1^3 \\ d_2^3 \\ d_3^3 \end{bmatrix}}{(d_1-d_2)(d_2-d_3)(d_3-d_1)}.$$

Thus

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} d_1+d_2+d_3 \\ -d_1d_2-d_2d_3-d_3d_1 \\ d_1d_2d_3 \end{bmatrix}.$$

Denote

$$\underline{u}_1 = (1-d_1\rho,\ 1,\ -d_1\rho,\ 1-d_2\rho,\ 1,\ -d_2\rho,\ 1-d_3\rho,\ 1,\ -d_3\rho)^T,$$

$$\underline{u}_2 = (1-d_1\rho,\ 0,\ 1,\ 1-d_2\rho,\ 0,\ 1-d_3\rho,\ 0,\ 1)^T$$

and

$$\underline{v} = [\underline{v}_1^T, \underline{v}_2^T, v_3^T], \qquad \underline{w} = [\underline{w}_1^T, \underline{w}_2^T, \underline{w}_3^T],$$

where for $i = 1, 2, 3$

$$\underline{v}_i = ((1 - d_i\rho)(d_i - \rho),\, d_i(1 + \rho^2),\, -d_i^2\rho)^T,$$

$$\underline{w}_i = ((1 - d_i\rho)(d_i - \rho),\, -\rho,\, d_i(1 + \rho^2))^T.$$

Note that from (11) (its second and third row) it follows that

$$\underline{u}_i^T \underline{c} = 0, \quad i = 1, 2. \tag{17}$$

Similarly, the fourth, sixth and eighth row of (11) imply

$$\underline{v}_i^T \underline{c}_i = 0, \quad i = 1, 2, 3 \tag{18}$$

and the fifth, seventh and ninth row of (11) imply

$$\underline{w}_i^T \underline{c}_i = 0, \quad i = 1, 2, 3. \tag{19}$$

Let $\alpha = d_1$, $\beta = d_2$ and $\gamma = d_3$. Denote also $\widetilde{\alpha} = 1 - \alpha\rho$, $\widetilde{\beta} = 1 - \beta\rho$ and $\widetilde{\gamma} = 1 - \gamma\rho$. From (7) we get

$$(1 - \rho^2)V_0 = (1 - \rho^2)\Delta(D_0 - CD_1)$$

$$= \begin{bmatrix} \widetilde{\alpha} & 0 & 0 & \widetilde{\beta} & 0 & 0 & \widetilde{\gamma} & 0 & 0 \\ \widetilde{\alpha} & -\alpha\rho & 0 & \widetilde{\beta} & -\beta\rho & 0 & \widetilde{\gamma} & -\gamma\rho & 0 \\ \widetilde{\alpha} & 1 & -\alpha\rho & \widetilde{\beta} & 1 & -\beta\rho & \widetilde{\gamma} & 1 & -\gamma\rho \\ \widetilde{\alpha} & 0 & 1 & \widetilde{\beta} & 0 & 1 & \widetilde{\gamma} & 0 & 1 \\ \widetilde{\alpha} & 0 & 0 & \widetilde{\beta} & 0 & 0 & \widetilde{\gamma} & 0 & 0 \\ 1 - \rho^2 & 0 & 0 & 1 - \rho^2 & 0 & 0 & 1 - \rho^2 & 0 & 0 \end{bmatrix}.$$

Note that the second and third rows of the above matrix are equal $\underline{u}_1^T$ and $\underline{u}_2^T$, respectively. Thus (17) implies

$$\underline{r}_0 = V_0\,\underline{c} = \frac{1}{1 - \rho^2} \begin{bmatrix} \widetilde{\alpha} & 0 & 0 & \widetilde{\beta} & 0 & 0 & \widetilde{\gamma} & 0 & 0 \\ \widetilde{\alpha} & -\alpha\rho & 0 & \widetilde{\beta} & -\beta\rho & 0 & \widetilde{\gamma} & -\gamma\rho & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \widetilde{\alpha} & 0 & 0 & \widetilde{\beta} & 0 & 0 & \widetilde{\gamma} & 0 & 0 \\ 1 - \rho^2 & 0 & 0 & 1 - \rho^2 & 0 & 0 & 1 - \rho^2 & 0 & 0 \end{bmatrix} \underline{c}$$

and thus (15) for $j = 0$ holds.

To simplify formulas, denote $\widehat{\alpha} = \beta + \gamma$, $\widehat{\beta} = \alpha + \gamma$ and $\widehat{\gamma} = \alpha + \beta$. From (7) we obtain

$$(1-\rho^2)[V_1 - (C^T + a_1\mathrm{Id})V_0]$$

$$= -\left[\begin{array}{cccccc|ccc}
\widetilde{\alpha}\,\widehat{\alpha} & 0 & 0 & \widetilde{\beta}\,\widehat{\beta} & 0 & 0 & \widetilde{\gamma}\,\widehat{\gamma} & 0 & 0 \\
\widetilde{\alpha}(\widehat{\alpha}+\rho) & -\alpha\,\widehat{\alpha}\rho & 0 & \widetilde{\beta}(\widehat{\beta}+\rho) & -\beta\,\widehat{\beta}\rho & 0 & \widetilde{\gamma}(\widehat{\gamma}+\rho) & -\gamma\,\widehat{\gamma}\rho & 0 \\
\widetilde{\alpha}(\widehat{\alpha}+\rho) & \widehat{\alpha}-\alpha\rho^2 & -\alpha\,\widehat{\alpha}\rho & \widetilde{\beta}(\widehat{\beta}+\rho) & \widehat{\beta}-\beta\rho^2 & -\beta\,\widehat{\beta}\rho & \widetilde{\gamma}(\widehat{\gamma}+\rho) & \widehat{\gamma}-\gamma\rho^2 & -\gamma\,\widehat{\gamma}\rho \\
\widetilde{\alpha}(\widehat{\alpha}+\rho) & \rho & \widehat{\alpha}-\alpha\rho^2 & \widetilde{\beta}(\widehat{\beta}+\rho) & \rho & \widehat{\beta}-\beta\rho^2 & \widetilde{\gamma}(\widehat{\gamma}+\rho) & \rho & \widehat{\gamma}-\gamma\rho^2 \\
\widetilde{\alpha}(\widehat{\alpha}+\rho) & 0 & \rho & \widetilde{\beta}(\widehat{\beta}+\rho) & 0 & \rho & \widetilde{\gamma}(\widehat{\gamma}+\rho) & 0 & \rho \\
\widetilde{\alpha}\rho+\widehat{\alpha}(1-\rho^2) & 0 & 0 & \widetilde{\beta}\rho+\widehat{\beta}(1-\rho^2) & 0 & 0 & \widetilde{\gamma}\rho+\widehat{\gamma}(1-\rho^2) & 0 & 0
\end{array}\right]$$

Note that the second row of this matrix can be written in the form

$$(\widetilde{\alpha}\,\widehat{\alpha},\ -\alpha\rho\,\widehat{\alpha},\ -\rho,\ \widetilde{\beta}\,\widehat{\beta},\ -\beta\rho\,\widehat{\beta},\ -\rho,\ \widetilde{\gamma}\,\widehat{\gamma},\ -\gamma\rho\,\widehat{\gamma},\ -\rho) + \rho\underline{u}_2.$$

The third and fourth rows, respectively, can be written as

$$a_1\underline{u}_1 - \underline{v} \quad \text{and} \quad a_1\underline{u}_2 + \underline{w}.$$

The fifth and sixth rows, respectively, can be written as

$$(\widetilde{\alpha}\,\widehat{\alpha},\ 0,\ 0,\ \widetilde{\beta}\,\widehat{\beta},\ 0,\ 0,\ \widetilde{\gamma}\,\widehat{\gamma},\ 0,\ 0) + \rho\underline{u}_2$$

and

$$((1-\rho^2)\widehat{\alpha},\ 0,\ -\rho,\ (1-\rho^2)\widehat{\beta},\ 0,\ -\rho,\ (1-\rho^2)\widehat{\gamma},\ 0,\ -\rho) + \rho\underline{u}_2.$$

Therefore, from (17), (18) and (19) we get

$$-(1-\rho^2)\underline{r}_1$$

$$= \left[\begin{array}{ccccccccc}
\widetilde{\alpha}\,\widehat{\alpha} & 0 & 0 & \widetilde{\beta}\,\widehat{\beta} & 0 & 0 & \widetilde{\gamma}\,\widehat{\gamma} & 0 & 0 \\
\widetilde{\alpha}\,\widehat{\alpha} & -\rho\alpha\,\widehat{\alpha} & -\rho & \widetilde{\beta}\,\widehat{\beta} & -\rho\beta\,\widehat{\beta} & -\rho & \widetilde{\gamma}\,\widehat{\gamma} & -\rho\gamma\,\widehat{\gamma} & -\rho \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\widetilde{\alpha}\,\widehat{\alpha} & 0 & 0 & \widetilde{\beta}\,\widehat{\beta} & 0 & 0 & \widetilde{\gamma}\,\widehat{\gamma} & 0 & 0 \\
(1-\rho^2)\widehat{\alpha} & 0 & -\rho & (1-\rho^2)\widehat{\beta} & 0 & -\rho & (1-\rho^2)\widehat{\gamma} & 0 & -\rho
\end{array}\right] \underline{c}.$$

Consequently, (15) holds for $j = 1$.

Now we consider the case $j = 2$. Denote additionally $\overline{\alpha} = \beta\gamma$, $\overline{\beta} = \gamma\alpha$, $\overline{\gamma} = \alpha\beta$. From (7) we get

$$(1-\rho^2)\left[V_2 - (C^T + a_1\mathrm{Id})V_1 + (a_1C^T - a_2\mathrm{Id})V_0\right]$$

$$= \left[\begin{array}{cccccc|ccc}
\widetilde{\alpha}\,\overline{\alpha} & 0 & 0 & \widetilde{\beta}\,\overline{\beta} & 0 & 0 & \widetilde{\gamma}\,\overline{\gamma} & 0 & 0 \\
\widetilde{\alpha}(\widehat{\alpha}\rho+\overline{\alpha}) & -a_3\rho & 0 & \widetilde{\beta}(\widehat{\beta}\rho+\overline{\beta}) & -a_3\rho & 0 & \widetilde{\gamma}(\widehat{\gamma}\rho+\overline{\gamma}) & -a_3\rho & 0 \\
\widetilde{\alpha}(\widehat{\alpha}\rho+\overline{\alpha}) & \overline{\alpha}-\widehat{\alpha}\alpha\rho^2 & -a_3\rho & \widetilde{\beta}(\widehat{\beta}\rho+\overline{\beta}) & \overline{\beta}-\widehat{\beta}\beta\rho^2 & -a_3\rho & \widetilde{\gamma}(\widehat{\gamma}\rho+\overline{\gamma}) & \overline{\gamma}-\widehat{\gamma}\gamma\rho^2 & -a_3\rho \\
\widetilde{\alpha}(\widehat{\alpha}\rho+\overline{\alpha}) & \widehat{\alpha}\rho & \overline{\alpha}-\widehat{\alpha}\alpha\rho^2 & \widetilde{\beta}(\widehat{\beta}\rho+\overline{\beta}) & \widehat{\beta}\rho & \overline{\beta}-\widehat{\beta}\beta\rho^2 & \widetilde{\gamma}(\widehat{\gamma}\rho+\overline{\gamma}) & \widehat{\gamma}\rho & \overline{\gamma}-\widehat{\gamma}\gamma\rho^2 \\
\widetilde{\alpha}(\widehat{\alpha}\rho+\overline{\alpha}) & 0 & \widehat{\alpha}\rho & \widetilde{\beta}(\widehat{\beta}\rho+\overline{\beta}) & 0 & \widehat{\beta}\rho & \widetilde{\gamma}(\widehat{\gamma}\rho+\overline{\gamma}) & 0 & \widehat{\gamma}\rho \\
\widetilde{\alpha}\,\widehat{\alpha}\rho+\overline{\alpha}(1-\rho^2) & 0 & 0 & \widetilde{\beta}\,\widehat{\beta}\rho+\overline{\beta}(1-\rho^2) & 0 & 0 & \widetilde{\gamma}\,\widehat{\gamma}\rho+\overline{\gamma}(1-\rho^2) & 0 & 0
\end{array}\right]$$

Note that the second row of this matrix can be written as

$$[\widetilde{\alpha}\,\overline{\alpha},\, -\rho a_3,\, -\rho\,\widehat{\alpha},\, \widetilde{\beta}\,\overline{\beta},\, -\rho a_3,\, -\rho\,\widehat{\beta},\, \widetilde{\gamma}\,\overline{\gamma}, -\rho a_3,\, -\rho\,\widehat{\gamma}] + \rho a_1\underline{u}_2 - \rho\underline{w} - \rho^2\underline{u}_1$$

while the third and the fourth rows, respectively, are

$$-a_2\underline{u}_1 - [\widehat{\alpha}\underline{v}_1^T,\, \widehat{\beta}\underline{v}_2^T,\, \widehat{\gamma}\underline{v}_3^T] \quad \text{and} \quad -a_2\underline{u}_2 - [\widehat{\alpha}\underline{w}_1^T,, \widehat{\beta}\underline{w}_2^T,\, \widehat{\gamma}\underline{w}_3^T]$$

and the fifth and sixth row, respectively, are

$$[\widetilde{\alpha}\,\overline{\alpha},\, 0,\, 0, \widetilde{\beta}\,\overline{\beta},\, 0,\, 0, \widetilde{\gamma}\,\overline{\gamma}] + a_1\rho\underline{u}_2 - \rho\underline{w} - \rho^2\underline{u}_1$$

and

$$[(1-\rho^2)\overline{\alpha},\, 0,\, -\rho\,\widehat{\alpha},\, (1-\rho^2)\overline{\beta},\, 0,\, -\rho\,\widehat{\beta},\, (1-\rho^2)\overline{\gamma},\, 0,\, -\rho\,\widehat{\gamma}] + a_1\rho\underline{u}_2 - \rho\underline{w} - \rho^2\underline{u}_1.$$

Therefore, from (17), (18) and (19) we get

$$-(1-\rho^2)\underline{r}_2$$

$$= \left[\begin{array}{ccccccccc}
\widetilde{\alpha}\,\overline{\alpha} & 0 & 0 & \widetilde{\beta}\,\overline{\beta} & 0 & 0 & \widetilde{\gamma}\,\overline{\gamma} & 0 & 0 \\
\widetilde{\alpha}\,\overline{\alpha} & -\rho a_3 & -\rho\,\widehat{\alpha} & \widetilde{\beta}\,\overline{\beta} & -\rho a_3 & -\rho\,\widehat{\beta} & \widetilde{\gamma}\,\overline{\gamma} & -\rho a_3 & -\rho\,\widehat{\gamma} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\widetilde{\alpha}\,\overline{\alpha} & 0 & 0 & \widetilde{\beta}\,\overline{\beta} & 0 & 0 & \widetilde{\gamma}\,\overline{\gamma} & 0 & 0 \\
(1-\rho^2)\overline{\alpha} & 0 & -\rho\,\widehat{\alpha} & (1-\rho^2)\overline{\beta} & 0 & -\rho\,\widehat{\beta} & (1-\rho^2)\overline{\gamma} & 0 & -\rho\,\widehat{\gamma}
\end{array}\right] \underline{c}.$$

Consequently, (15) holds for $j = 2$.

Now we consider the case $j = 3$. From (7) we get

$$(1-\rho^2)\left[V_3 - (C^T + a_1\mathrm{Id})V_2 + (a_1C^T - a_2\mathrm{Id})V_1 + (a_2C^T - a_3\mathrm{Id})V_0\right]$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\widetilde{\alpha}\,\overline{\alpha}\rho & 0 & 0 & -\widetilde{\beta}\,\overline{\beta}\rho & 0 & 0 & -\widetilde{\gamma}\,\overline{\gamma}\rho & 0 & 0 \\ -\widetilde{\alpha}\,\overline{\alpha}\rho & a_3\rho^2 & 0 & -\widetilde{\beta}\,\overline{\beta}\rho & a_3\rho^2 & 0 & -\widetilde{\gamma}\,\overline{\gamma}\rho & a_3\rho^2 & 0 \\ -\widetilde{\alpha}\,\overline{\alpha}\rho & -\overline{\alpha}\rho & a_3\rho^2 & -\widetilde{\beta}\,\overline{\beta}\rho & -\overline{\beta}\rho & a_3\rho^2 & -\widetilde{\gamma}\,\overline{\gamma}\rho & -\overline{\gamma}\rho & a_3\rho^2 \\ -\widetilde{\alpha}\,\overline{\alpha}\rho & 0 & -\rho\,\overline{\alpha} & -\widetilde{\beta}\,\overline{\beta}\rho & 0 & -\rho\,\overline{\beta} & -\widetilde{\gamma}\,\overline{\gamma}\rho & 0 & -\rho\,\overline{\gamma} \\ -\widetilde{\alpha}\,\overline{\alpha}\rho & 0 & 0 & -\widetilde{\beta}\,\overline{\beta}\rho & 0 & 0 & -\widetilde{\gamma}\,\overline{\gamma}\rho & 0 & 0 \end{bmatrix}$$

Note that the third and fourth row of this matrix, respectively, are

$$a_3\underline{u}_1 - [\overline{\alpha}\underline{v}_1^T, \overline{\beta}\underline{v}_2^T, \overline{\gamma}\underline{v}_3^T] \quad \text{and} \quad -a_3\underline{u}_2 + [\overline{\alpha}\underline{w}_1^T, \overline{\beta}\underline{w}_2^T, \overline{\gamma}\underline{w}_3^T]$$

Again, using (17), (18) and (19) we conclude that

$$\underline{r}_3 = -\frac{\rho}{1-\rho^2} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \widetilde{\alpha}\,\overline{\alpha} & 0 & 0 & \widetilde{\beta}\,\overline{\beta} & 0 & 0 & \widetilde{\gamma}\,\overline{\gamma} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \widetilde{\alpha}\,\overline{\alpha} & 0 & \overline{\alpha} & \widetilde{\beta}\,\overline{\beta} & 0 & \overline{\beta} & \widetilde{\gamma}\,\overline{\gamma} & 0 & \overline{\gamma} \\ \widetilde{\alpha}\,\overline{\alpha} & 0 & 0 & \widetilde{\beta}\,\overline{\beta} & 0 & 0 & \widetilde{\gamma}\,\overline{\gamma} & 0 & 0 \end{bmatrix} \underline{c}.$$

Consequently, (16) holds. □

4. Proofs of lemmas

Proof of Lemma 1. The coefficients of the polynomial W_3 are positive. Therefore it is strictly increasing. Thus there is exactly one real root $x_1 = x_1(\rho) < 0$ and two complex conjugate roots $x_2 = x_2(\rho)$ and $x_3 = x_3(\rho)$, $x_2 = \overline{x}_3$. We need to show that the real root x_1 is outside of the interval $[-2|\rho|, 2|\rho|]$. Since W_3 is strictly increasing to show that $x_1 < -2|\rho|$ it suffices to prove that $W_3(-2|\rho|)$ is positive. We note that

$$W_3(-2|\rho|) = -8|\rho|^3 - (2-\rho^2+2\rho^4)|\rho| + 2(2+2\rho^2+2\rho^4+\rho^6)$$

$$= 4 - 2|\rho| + 4\rho^2 - 7|\rho|^3 + 4\rho^4 - 2|\rho|^5 + 2\rho^6$$

$$= (1-|\rho|)^2 + 2(1-|\rho|^3)^2 + (1-|\rho|^5)^2 + 3\rho^2(1-|\rho|) + \rho^4(1-\rho^6) + 3\rho^4 > 0$$

□

Proof of Lemma 2. Expanding the determinant in the equation (9) and introducing the variable

$$x = -\rho\left(d + \frac{1}{d}\right)$$

we arrive at the equivalent equation

$$W_3(x) = 0.$$

By Lemma 1 we conclude that

$$-\frac{x_i}{2\rho} = \frac{1}{2}\left(d + \frac{1}{d}\right) \notin [-1, 1], \quad i = 1, 2, 3. \tag{20}$$

That is exactly one of the two solutions of the above equation (for the variable d) is inside open unit disc and exactly one outside. In particular,

$$d_1 = \frac{-x_1 - \sqrt{x_1^2 - 4\rho^2}}{2\rho} \in \mathbb{R}. \tag{21}$$

Since x_2 and x_3 are complex conjugate, then d_2 and d_3 are also complex conjugate, since they both are in open unit disc.

By the Viete formulas for W_3 we have

$$x_1 + x_2 + x_3 = 0. \tag{22}$$

Note that if d_i is the solution we seek, then the remaining solution of (20) is $1/d_i$. Therefore (22) is equivalent to

$$d_1 + \frac{1}{d_1} + d_2 + \frac{1}{d_2} + d_3 + \frac{1}{d_3} = 0.$$

Multiplying the above identity by $d_1 d_2 d_3$ we arrive at (10). □

Proof of Lemma 3. Due to Lemma 2 the linear system

$$\widetilde{\mathbb{Q}}\,\underline{c} = (1 - \rho^2)\widetilde{e}_1 \tag{23}$$

is equivalent to (11).

We will prove that the matrix $\widetilde{\mathbb{Q}}$ is invertible by showing that its determinant is non-zero. Equivalently we consider determinant of the matrix

$$\widetilde{\mathbb{Q}}_1 =$$

$$\begin{bmatrix}
i_\alpha & i_\beta & i_\gamma & j_\alpha & j_\alpha & j_\beta & j_\beta & j_\gamma & j_\gamma \\
j_\alpha & j_\beta & j_\gamma & 1 & -\rho\alpha & 1 & -\rho\beta & 1 & -\rho\gamma \\
j_\alpha & j_\beta & j_\gamma & 0 & 1 & 0 & 1 & 0 & 1 \\
k_\alpha & 0 & 0 & (1+\rho^2)\alpha & -\rho\alpha^2 & 0 & 0 & 0 & 0 \\
k_\alpha & 0 & 0 & -\rho & (1+\rho^2)\alpha & 0 & 0 & 0 & 0 \\
0 & k_\beta & 0 & 0 & 0 & (1+\rho^2)\beta & -\rho\beta^2 & 0 & 0 \\
0 & k_\beta & 0 & 0 & 0 & -\rho & (1+\rho^2)\beta & 0 & 0 \\
0 & 0 & k_\gamma & 0 & 0 & 0 & 0 & (1+\rho^2)\gamma & -\rho\gamma^2 \\
0 & 0 & k_\gamma & 0 & 0 & 0 & 0 & -\rho & (1+\rho^2)\gamma
\end{bmatrix},$$

where

$$i_x = 5(1-\rho x) + 1 - \rho^2, \quad j_x = 1 - \rho x \quad \text{and} \quad k_x = j_x(x-\rho),$$

for $x = \alpha,\, \beta,\, \gamma$.

Let

$$A(d) = \frac{(1+\rho^2-\rho(d+1/d))(1+\rho^2+\rho d)}{1+\rho^2+\rho^4} \quad \text{and} \quad B(d) = \alpha(1/d).$$

We add 4th column multiplied by $A(\alpha)$ and 5th column multiplied by $B(\beta)$ and subtract the result from the 1st column. Then we add 6th column multiplied by $A(\beta)$ and 7th column multiplied by $B(\beta)$ and subtract the result from the 2nd column. Finally, we add 8th column multiplied by $A(\gamma)$ and 9th column multiplied by $B(\gamma)$ and subtract the result from the 3rd column. All these operations do not change the absolute value of the determinant of $\widetilde{\mathbb{Q}}$ and the resulting matrix is block diagonal with the following blocks on the diagonal

$$B_0 = \begin{bmatrix}
i_\alpha - j_\alpha(A(\alpha)+B(\alpha)) & i_\beta - j_\beta(A(\beta)+B(\beta)) & i_\gamma - j_\gamma(A(\gamma)+B(\gamma)) \\
j_\alpha - A(\alpha) + \rho\alpha B(\alpha) & j_\beta - A(\beta) + \rho\beta B(\beta) & j_\gamma - A(\gamma) + \rho\gamma B(\gamma) \\
j_\alpha - B(\alpha) & j_\beta - B(\beta) & j_\gamma - B(\gamma)
\end{bmatrix}$$

and

$$B_i = \begin{bmatrix} (1+\rho^2)d & -\rho d^2 \\ -\rho & (1+\rho^2)d \end{bmatrix}, \quad i = 1,2,3.$$

It suffices to prove that $\det B_i \neq 0$, $i = 0,1,2,3$.

Note that

$$\det B_i = d^2[(1+\rho^2)^2 + \rho^2] > 0 \quad i = 1,2,3.$$

Now we consider $G(\rho) := \det B_0$. Expanding the determinant of B_0 we arrive at a "polynomial of twelfth degree" in ρ

$$G(\rho) = -4 + 2s_1\rho - 2\rho^2 + (2s_1 + s_1^3 - 6s_1s_2 + 9s_3)\rho^3 - 2(1 - s_2^2 + s_1s_3)\rho^4$$

$$+2(s_1 - s_1s_2 + 2s_3)\rho^5 + (1 + 2s_2^2 - 2s_1s_3 - 6s_3^2 - s_2^3 + 3s_1s_2s_3)\rho^6 - 2(s_1s_2 - 2s_3)\rho^7$$

$$+2(s_2^2 - s_1s_3 - s_3^2)\rho^8 + (s_1s_2 - 2s_3 + s_3^3)\rho^9 - 2s_3^2\rho^{10} + s_3^2\rho^{12}.$$

where

$$s_1 = d_1 + d_2 + d_3, \quad s_2 = d_1d_2 + d_2d_3 + d_3d_1, \quad s_3 = d_1d_2d_3. \tag{24}$$

By (10) we get

$$G(\rho) = -4 + 2s_1\rho - 2\rho^2 + (2s_1 + s_1^3 + 6s_1^2s_3 + 9s_3)\rho^3 - 2(1 - s_1^2s_3^2 + s_1s_3)\rho^4$$

$$+2(s_1 + s_1^2s_3 + 2s_3)\rho^5 + (1 - s_1^2s_3^2 - 2s_1s_3 - 6s_3^2 - s_1^3s_3^3)\rho^6 + 2(s_1^2s_3 + 2s_3)\rho^7$$

$$+2(s_1^2s_3^2 - s_1s_3 - s_3^2)\rho^8 + (-s_1^2s_3 - 2s_3 + s_3^3)\rho^9 - 2s_3^2\rho^{10} + s_3^2\rho^{12};$$

Thet Viete formulas for W_3 give

$$t_2 = x_1x_2 + x_2x_3 + x_3x_1 = 2 - \rho^2 + 2\rho^4 > 0.$$

On the other hand

$$t_2 = \rho^2\left[\left(d_1 + \frac{1}{d_1}\right)\left(d_2 + \frac{1}{d_2}\right) + \left(d_2 + \frac{1}{d_2}\right)\left(d_3 + \frac{1}{d_3}\right) + \left(d_3 + \frac{1}{d_3}\right)\left(d_1 + \frac{1}{d_1}\right)\right].$$

By (24) and (10)

$$t_2 = \rho^2\left(\frac{s_2s_3 + s_1 + s_1s_2}{s_3} - 3\right) = \rho^2\left(\frac{s_1 - s_1s_3(s_1 + s_3)}{s_3} - 3\right).$$

Thus

$$s_1 = s_3\frac{3 + s_1^2 + t_2/\rho^2}{1 - s_3^2} \tag{25}$$

and consequently

$$s_1s_3 = s_3^2\,\frac{3 + s_1^2 + t_2/\rho^2}{1 - s_3^2} > 0$$

since $t_2 > 0$ and $|s_3| = |d_1d_2d_3| < 1$.

Since the coefficients of the polynomial W_3 depend only on ρ^2 then x_1 is a function of $|\rho|$. Since $x_1 < 0$ (see Lemma 1) it follows from (21) that

$$\rho d_1(\rho) < 0$$

and thus

$$\rho s_3 = \rho d_1(\rho) d_2(\rho) d_3(\rho) = \rho d_1(\rho) |d_2(\rho)|^2 < 0.$$

Moreover

$$\rho s_1 = \frac{\rho^2 s_1 s_3}{\rho s_3} < 0.$$

Transform (25) into

$$s_1^2 s_3 - (1 - s_3^2) s_1 + s_3(t_2/\rho^2 + 3) = 0$$

leading to

$$s_1 s_3 (1 - s_1 s_3) = s_1 s_3^3 + s_3^2 (t_2/\rho^2 + 3) > 0.$$

That is $0 < s_1 s_3 < 1$.

Now we will use the inequalities we have just derived

$$\rho s_1 < 0, \quad \rho s_3 < 0, \quad \text{and} \quad 0 < s_1 s_3 < 1.$$

to show that $G(\rho) < 0$ for any $\rho \in (-1, 1)$. To this end we split $G(\rho)$ into several terms and show that each of these terms is negative:

$$s_1 \rho < 0,$$

$$(2s_1 + s_1^3 + 6s_1^2 s_3 + 9s_3)\rho^3 < 0,$$

$$-2(1 - s_1^2 s_3^2 + s_1 s_3)\rho^4 + (1 - s_1^2 s_3^2 - 2s_1 s_3 - 6s_3^2 - s_1^3 s_3^3)\rho^6$$

$$= -\rho^4(2 - \rho^2) - 2s_1 s_3 (1 - s_1 s_3)\rho^4 - (s_1^2 s_3^2 + 2s_1 s_3 + 6s_3^2 + s_1^3 s_3^3)\rho^6 < 0,$$

$$2(s_1 + s_1^2 s_3 + 2s_3)\rho^5 + (-s_1^2 s_3 - 2s_3 + s_3^3)\rho^9$$

$$= 2s_1\rho^5 + s_1^2 s_3 \rho^5 (2 - \rho^4) + 2s_3\rho^5(1 - \rho^4) + s_3^3\rho^9 < 0,$$

$$2(s_1^2 s_3 + 2s_3)\rho^7 < 0,$$

$$2(s_1^2 s_3^2 - s_1 s_3 - s_3^2)\rho^8 = -2[s_1 s_3(1 - s_1 s_3) + s_3^2]\rho^8 < 0,$$

$$-2s_3^2\rho^{10} + s_3^2\rho^{12} = -s_3^2\rho^{10}(2 - \rho^2) < 0.$$

□

REFERENCES

CIEPIELA, P. (2004). Estimation of the mean in rotation schemes. MSc Thesis, Warsaw Univ. Techn., Fac. Math. Inform. Sci., Warsaw (in Polish)

KOWALSKI, J. (2009). Optimal estimation in rotation patterns. J. Statist. Plann. Infer. 139(4), 2429-2436.

KOWALSKI, J. (2010). Optimal recurrence estimation of the mean in rotation schemes. PhD Thesis, Warsaw Univ. Techn., Fac. Math. Inform. Sci., Warsaw (in Polish).

KOWALSKI, J. and WESOŁOWSKI, J. (2010). Recurrence optimal estimators for rotation cascade patterns with holes - unpublished manuscript

PATTERSON, H.D. (1955). Sampling on successive occasions. J. Roy. Statist. Soc. B 12, 241-255.

POPIŃSKI, W. (2006). Development of the Polish Labour Force Survey. Statist. Trans. 7(5), 1009-1030.

SZARKOWSKI, A. and WITKOWSKI, J. (1994). The Polish Labour Force Survey. Statist. Trans. 1(4), 467-483.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 287—312

ON SOME ROBUST ESTIMATORS FOR POLISH BUSINESS SURVEY[1]

Grażyna Dehnel, Elżbieta Gołata

ABSRACT

The paper presents attempts to use administrative data and non-standard techniques to estimate basic economic information about small business in the joint cross-section of Polish Classification of Economic Activities and regions. Due to many outliers and significant fraction of entities for which many variables are equal to zero, the considered distributions are heterogeneous. Therefore, Horvitz-Thompson estimates are compared with the robust ones: modified GREG, Winsor and local regression. Results obtained in the study present practical possibilities of adopting robust estimation techniques to small business data in Poland. Properties of the estimators are discussed.

Key words: Domain estimation, Robust estimators, Small business statistics.

1. Aim of the study, data and assumptions of the research

The objective of the referred study was to provide economic information about small business enterprises (employing up to 9 workers) across regions depending on the type of economic activity. Like in other countries, enterprises in Poland can be divided into three categories: big, medium and small, depending on their size measured by the number of employees[2]. Statutory duty to provide information on economic effects differs between these groups as concerns scope and frequency of supplying the reports to Central Statistical Office (CSO). According to the regulations, small business entities are exempt from statistical reporting and take part in a yearly survey which enables estimation for the whole country. The majority of small business entities are self-employed individuals, especially in commerce, which operate mainly on local markets. Although small

[1] The paper was presented during the 57th Session of ISI held in Durban, South Africa in August 2009, on IPM56: Modelling Economic Data to Produce Small Area Estimates.

[2] According to Polish regulations, the category of big enterprises comprises all those which employ at least 50 persons. Enterprises employing from 10 to 50 people are referred to as medium. And all those employing up to 9 people (inclusive) are called *small business*.

economic entities constitute the basis of local economy[1], there is no information about this group available at the regional level. The development of market economy stimulates information demand for more and more detailed data about small business at the regional level (Falorsi *et al*. (2000), Klimanek and Paradysz (2006), Pawlowska (2005)). In order to meet this demand, our research was aimed at widening the range of estimates about small business by providing information for regions and type of economic activity.

Domain estimation in business statistics is rather difficult due to undesirable properties of the estimated characteristics. Variables that describe small businesses (those which are estimated as well as those which are used as covariates) are usually of non-homogeneous distributions: extremely right skewed with high dispersion and strong kurtosis. The population of small business is heterogeneous also because of the existence of many outliers, in addition to the presence of quite a significant fraction of entities for which many variables are equal to zero. Outliers have a significant effect on estimation results, especially at a low level of aggregation. Under such conditions the properties of estimators like unbiasedness or effectiveness are not preserved. Although for some observations the variables take extreme values, they need not necessarily be false. Extremely large observations are a natural component in business surveys. That is why we tried to explore some alternative estimation techniques which are less sensitive to outliers. The following methods of estimation were analysed: generalised regression estimator GREG, its modification proposed by Chambers *et al.* (2001a), Winsor estimator as discussed by Mackin and Preston (2002) and local regression presented by Kim *et al.* (2001). Another detail objective of the paper was to examine the precision of small domain estimates in respect to different non-standard estimation techniques when applied to Polish data for small business.

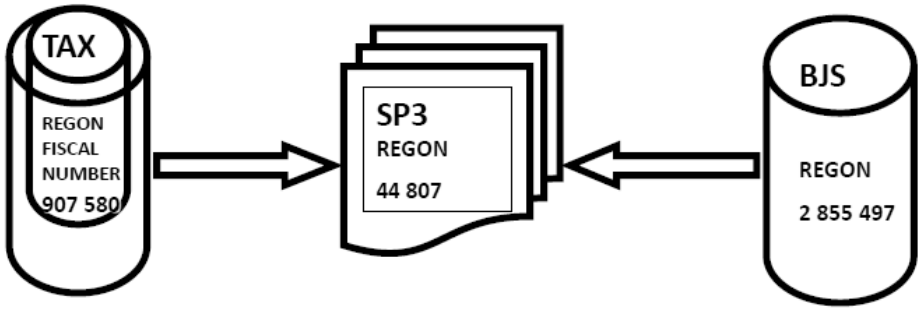
As concerns data, an attempt was made to apply auxiliary information from administrative files for more effective use of small businesses survey (SP-3). The analysis refers to data from the year 2001 for which databases were made available by Central Statistical Office (CSO). The main data source about small business units used in the study is SP-3 survey conducted by CSO on a 5% random sample chosen using a stratified sampling design. 114 thousand units were chosen, but only 44 807 responded (39.3% response rate). The information obtained in the survey comprise, among others, the following topics: costs, revenue, financial result, employment, investments, etc. It is assumed that the sample is drawn from an updated frame - the register called Database of Statistical Units (BJS). But in reality a considerable part of entities does not run their business, for example because of bankruptcy.

[1] According to Central Statistical Office there were 3.325 million enterprises in Poland in 2001. Most of them, i.e. 1.6 million, belonged to the small business category, employing up to 9 people. Most small economic entities, that is 1.545 million, are self-employed individuals. And only less than 4% of small business companies possess legal corporate personality. Over 3.2 million people, that is about 20% of the Polish labour force were employed in small business companies.

Since small business statistics is connected with a serious problem of non-response and incompleteness, our intention was to use supplementary data from registers providing complete and reliable information. Auxiliary variables used as covariates were taken from two administrative registers: Database of Statistical Units (BJS) and the Ministry of Finance's Tax Register (see figure 1). The Database of Statistical Units (BJS) was specially created by CSO to serve as a frame in all surveys conducted on the population of economic enterprises. BJS was constructed on the basis of Economic Units Register called REGON and includes additional information from other data sources. It is updated every year and in 2001 contained 2 855 497 records.

The Ministry of Finance's Tax Register called POLTAX served as the second auxiliary data source and was considered as especially reliable. Confidentiality of the data was specially secured. In the year 2001 for which data from SP-3 survey was available, the tax register contained information on tax statements from less than 1 million economic entities. The data base comprised 844 675 records referring to self-employed individuals (Personal Income Tax - PIT) and 62 905 records for business that possess legal corporate personality (Corporate Income Tax - CIT).

Figure 1. Integration of data from different sources



Source: Own compilation.

It was assumed that integration of all the three data sources: SP-3, BJS and POLTAX will enable a more detailed and complete analysis. A deterministic approach using a unique linkage key was applied (REGON number). But the percentage of matches was only 60. Differences in the number of observations in data sources used in the study were due to a high fraction of non-response in SP-3 survey, as well as incompleteness of the tax register and outdated information in BJS. Over 40% of units sampled could not be matched with the records in the tax register. The high fraction of missing linkage caused the danger of selectivity. It was very likely that the unmatched records constituted a specific group of enterprises, which differed significantly from the average in terms of the most important characteristics. This supposition was confirmed by the analysis of

difference in estimates of costs and revenue based on *(i)* integrated database, *(ii)* firms drawn to the sample but unmatched with the Tax Register and *(iii)* taking into account the whole sample (matched and unmatched records - see table 1).

Table 1. Comparison of cost and revenue estimates based on different data, small business, Poland, 2001

Variable (thousands PLN)	Estimation based on Integrated database *(i)*	Estimation based on firms not included in the Integrated database *(ii)*	Estimation based on the whole SP-3 survey *(iii)*
Cost per firm	318.8	155.4	253.8
Revenue per firm	369.1	184	282.1

Source: Own estimation based on SP3 Survey and Integrated database.

The overestimation obtained upon the integrated database made us apply a special procedure relying on assigning to unmatched firms values obtained from the survey. The correlation coefficients estimated for matched units showed a strong positive relation between variables form SP-3 survey and Tax Register $(r \in (0{,}75;0{,}99))$. This imputation procedure allowed us to take all records from SP-3 survey integrated with Tax Register data for further analysis.

Estimation was conducted for the following economic variables: average number of employees under a contract employed in a company, the amount of average monthly salary paid over one year per company, average revenue per company and average costs per company. As auxiliary variables we used information from additional data sources: BJS and Tax Register. For the two estimated variables: Revenue and Costs we applied Costs and Revenue from Tax Register as covariates. For the remaining two target variables: average number of employed and Gross Wage, as auxilary information we implemeted different combinations of: Number of Employees, Gross Wage, Revenue and Costs from Tax Register. The choice was made mostly upon the strength of relation between the variables.

The small domains were constructed on the basis of regions and types of economic activity and were defined as an 'intersection' of the joint distribution of economic entities by voivodships and sections of Polish Economic Activity classification (PKD section). Estimation was conducted for 176 domains: 17 regions (R) and 11 PKD sections (S) (**R**egion&**S**ection). Since the number of domains taken into account in the study is quite considerable, we present only selected results for one of the estimated variables - Gross Wage in the region of *Zachodniopomorskie voivodship* and in selected sections: *Industry, Construction, Trade, Hotels and Restaurants.* However, to evaluate results obtained in the study, synthetic characteristics across all domains were provided.

The results provided by the study were compared with the original GREG approach. In order to determine the estimation precision an approximated method based on samples and bootstrap was applied. The bootstrap method presented by

Efron (1979) for simple random sampling needs modifications when applied for complex samples. We applied the approach described in detail by Shao and Tu (1995). The simulation study consisted of 500 iterations. In each one, a subsample of size *n-1* was drawn and for each iteration sample modified weights were distinguished to estimate the unknown parameter $\hat{Y}_{*b}$, where $(*)$ stands for different robust estimators used in the study. The empirical variance was obtained according to the standard approach

$$Var(\hat{Y}_*) = \frac{1}{500-1}\sum_{b=1}^{500}(\hat{Y}_{Rb} - \hat{Y}_*)^2 \tag{1}$$

Estimation precision was evaluated on the basis of estimator's coefficient of variation:

$$CV(\hat{Y}_*) = \frac{\sqrt{Var(\hat{Y}_*)}}{(\hat{Y}_*)} \tag{2}$$

Coefficient $RedCV(\hat{Y}_*)$ was used to measure the degree of *CV* reduction obtained by robust estimators applied:

$$RedCV(\hat{Y}_*) = \frac{CV(\hat{Y}_*) - CV(\hat{Y}_{DIR})}{CV(\hat{Y}_{DIR})} \ . \tag{3}$$

The term CV was used instead of REE as we perceive that variation obtained in the bootstrap method does not include the bias in contrast to MSE.

$$deff\left(\hat{Y}\right) = \frac{Var\left(\hat{Y}_*\right)}{Var\left(\hat{Y}_{DIR}\right)} \tag{4}$$

Additionally, the *design effect* coefficient *deff* (proposed by Kish (1965), (1995)) was provided to measure the effectiveness of the examined estimator in comparison with the direct one. All the values smaller than unity indicate that the method applied is more effective. In the course of our research correlation analysis as well as model evaluation specially detecting residuals were naturally implemented. We made also an attempt to provide validation of the estimates obtained against information from other sources and compared the results for other levels of aggregation.

2. The GREG estimator and its modifications

In business statistics which is characterised by huge diversification, direct estimation procedures do not provide satisfactory results. Applying GREG estimation that use auxiliary information is perceived as a solution, which usually

increases precision considerably. Despite this advantage D. Hedlin (2004) draws attention to how important good modelling is. He indicates that for a set of real data, different GREG estimators might produce wildly different results. The difference between them lies entirely in the choice of the model. The modification proposed by Chambers *et al.* (2001a) refers to GREG estimators assuming heteroscedasticity. They introduce transformation that depends on including additional auxiliary variable – '*z*' to the regression model of *y* on *x*:

$$\hat{Y}_{GREG,d} = \sum_{i \in S_d} w_i g_i y_i \tag{4}$$

where:

d – domain,

g_i – weight of *i*-th individual observation defined as:

$$g_i = 1 + \left(X_d - \hat{X}_{HT,d}\right)' \left(\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma} \right)^{-1} \left(\mathbf{x}_i / z_i^{\gamma} \right) \tag{5}$$

X – auxiliary variable which, depending on the approach, is defined as: *(i)* number of employees or *(ii)* revenue (see table 2)

z – auxiliary variable which, depending on the approach, is defined as: *(i)* number of employees or *(ii)* revenue (see table 2)

$\hat{Y}_{GREG,d}$ – estimate of Gross wage in domain *d* obtained by applying the GREG estimator

$\hat{X}_{HT,d}$ – direct Horvitz-Thompson estimate of the total value of the auxiliary variable *x* in domain *d*

X_d – total value of the auxiliary variable *x* in domain *d,*

γ – coefficient characterising the degree of heteroscedasticity, for $\gamma = 0$ the original GREG estimator is obtained.

As it is known from other surveys, γ coefficient should be included in the interval $1 \le \gamma \le 2$ (Särndal (1992) p.255); the following estimators were analyzed:

1) $\hat{Y}_{DIR}$ – direct Horvitz-Thompson estimator (HT),
2) $\hat{Y}^0_{GREG}$ – $\gamma = 0 \Rightarrow z_i^0$ regression estimator based on linear regression model assuming homoscedasticity (GREG estimator),
3) $\hat{Y}^1_{GREG}$ – $\gamma = 1 \Rightarrow z_i^1$,
4) $\hat{Y}^{1,5}_{GREG}$ – $\gamma = 1{,}5 \Rightarrow z_i^{1,5}$,
5) $\hat{Y}^2_{GREG}$ – $\gamma = 2 \Rightarrow z_i^2$.

The estimators denoted by numbers *(3) (4)* and *(5)* are regression estimators based on linear regression model assuming heteroscedasticity. The direct estimator was considered as the reference value when comparing the effectiveness of GREG estimator and its modifications. The second one of the estimators *(2)* - $\hat{Y}^0_{GREG}$ takes the form of GREG estimator, but its value differs slightly in comparison with the original GREG formula. The difference results from the fact that not all of the units drawn to the sample take part in the estimation. Those observations, for which auxiliary variable '*z*' is equal to zero[1] are omitted. In the case of the remaining three estimators, the linear regression model assumes heteroscedasticity of a degree indicated by γ coefficient. The modification assumes that each of the auxiliary variables might take both roles, of '*x'* and '*z'* with the warranty that '*z'* does not equal zero. Many combinations were considered and finally the following approaches were proposed (see table 2). The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* and *(z)* we used: *(i)* the number of employees from BJS and *(ii)* revenue from Tax Register.

Table 2. Auxiliary variables *'x'* and *'z'* used in GREG's modification, small business, Poland, 2001

Approach	Auxiliary variable '*x*'/ *data source*	Auxiliary variable '*z*' / *data source*
1	Revenue / *Tax Register*	Revenue / *Tax Register*
2	Revenue / *Tax Register*	Number of employees / *BJS*
3	Number of employees / *BJS*	Number of employees / *BJS*

Source: Own estimation based on SP3 Survey and Integrated database.

The synthetic characteristics summarising estimates of Gross Wage in the cross-section of all domains are presented in table 3. They show that the highest precision was obtained for $\hat{Y}^0_{GREG}$, the estimator ignoring observations for which "*z*" was equal to zero. With the increase of γ, usually the maximum, median and the average value of *CV* also increases. The biggest proportion of domains for which the relative dispersion of the estimator was smaller than in the case of the direct one, was also observed for $\hat{Y}^0_{GREG}$, and then other estimators $\hat{Y}^1_{GREG}$, $\hat{Y}^{1.5}_{GREG}$ and $\hat{Y}^2_{GREG}$ were classified. Only for the model with *Revenue* as an auxiliary variable was the order of the estimators changed, which can be explained by a weaker relation of the estimated and auxiliary variables (see table 4).

[1] The variable '*z'* is assumed not to take zero value. Nevertheless, it happens in practice that zero values occur for variables for which they are not expected (for example the revenue might be equal to zero for an enterprise running a business).

We cannot unequivocally determine which of the estimators is characterized by the highest precision. In each case an appropriate model and the definition of '*x*' and '*z*' is necessary. Model misspecification may lead to negative estimates. Such an untypical situation was observed for domain *Trade* in *Dolnoslaskie voivodship* (see figure 2. B).

Table 3. Characteristics of the CV distribution, GREG's modification estimates of Gross Wage, all domains, small business, Poland, 2001

Characteristics	Estimator				
	$\hat{Y}_{DIR}$	$\hat{Y}^{0}_{GREG}$	$\hat{Y}^{1}_{GREG}$	$\hat{Y}^{1,5}_{GREG}$	$\hat{Y}^{2}_{GREG}$
x – number of employees z - number of employees					
min	0.0721	0.0665	0.0670	0.0671	0.0673
max	1.0964	1.0402	1.0522	1.1754	1.3406
average	0.3153	0.3019	0.3020	0.3030	0.3061
median	0.2861	0.2739	0.2773	0.2799	0.2811
proportion of domains for which $CV < CV_{DIR}$ *(%)*		73.30	72.16	71.59	68.18
x - revenue z - number of employees					
min	0.073	0.070	0.073	0.073	0.073
max	0.779	0.990	0.863	0.795	0.979
average	0.304	0.297	0.306	0.310	0.312
median	0.281	0.266	0.283	0.286	0.286
proportion of domains for which $CV < CV_{DIR}$ *(%)*		68.75	56.25	46.02	45.45
x - revenue z - revenue					
min	0.070	0.070	-1.731	-1.388	-1.332
max	0.793	13.556	5.093	3.203	11.370
average	0.305	0.421	0.358	0.316	0.401
median	0.280	0.274	0.276	0.280	0.302
proportion of domains for which $CV < CV_{DIR}$ *(%)*		57.95	65.34	63.07	55.11

Source: Own estimation based on SP3 survey data, BJS, Tax Register.

Table 4. Correlation coefficients between the estimated variable – Gross Wage and auxiliary variables from SP3 survey, BJS and Tax Register, 2001, Poland Zachodniopomorskie voivodship

Gross Wage by PKD sections	Auxiliary variables	
	Number of employees (BJS)	Revenue (Tax Register)
Industry	0,538	0,479
Construction	0,587	0,301
Trade	0,579	0,517
Hotels and restaurants	0,536	0,402

Source: Own estimation based on SP3 survey data and Tax Register.

Defining auxiliary variables '*x*' and '*z*' as Revenue, the estimators $\hat{Y}^{1,5}_{GREG}$ and $\hat{Y}^{2}_{GREG}$ provided negative estimates. As can be seen from figure 2.A, outliers in both dimensions (of the estimated as well as of the auxiliary variable) influence estimation results. The regression lines corresponding to $\hat{Y}^{0}_{GREG}$, $\hat{Y}^{1}_{GREG}$, $\hat{Y}^{1,5}_{GREG}$, $\hat{Y}^{2}_{GREG}$, presented on the diagram show the following regularity (which was observed also for other domains). The γ coefficient determines the degree to which outliers in the dimension of *'y'* are included in the model. The linear regression $\hat{Y}^{0}_{GREG}$ with γ=0 includes outliers in the dimension of *'x'*. With the increase of γ, outliers in the dimension of *'y'* are considered to a greater degree by assigning to them greater weights g_i, and less importance is attached to outliers in the dimension of *'x'*. The value of γ influences also the range of weights. The smallest dispersion of weights refers to γ =0 *(* $g_i 0 \in$ *(-0,13 ; 2,4))*, and the biggest range is observed for γ=2 *(* $g_i 2 \in$ *(-35 ; 55))* (see also tab. 5).

Table 5. Characteristics of weight's distribution:

A) domain defined as Industry in Dolnoslaskie voivodship

x –revenue z - revenue	Min	Max	Weight for outlier *(x)*	Average	Standard deviation
g_i *0*	0.912	1.997	1.249	1.039	0.123
g_i *1*	-4.705	8.384	1.290	1.057	0.595
g_i *1,5*	-16.917	25.764	1.087	1.090	1.607
g_i *2*	-35.176	50.416	1.016	1.129	3.008

B) domain defined as Trade in Dolnoslaskie voivodship

x –revenue *z - revenue*	Min	Max	Weight for outlier *(x)*	Average	Standard deviation
g_i *0*	0.964	2.448	0.984	1.020	0.067
g_i *1*	0.117	54.323	0.117	1.042	1.631
g_i *1,5*	-1.456	56.063	-1.456	1.042	1.687
g_i *2*	-3.170	55.812	-3.170	1.040	1.684

Source: Own estimation based on SP3 survey data and Tax Register.

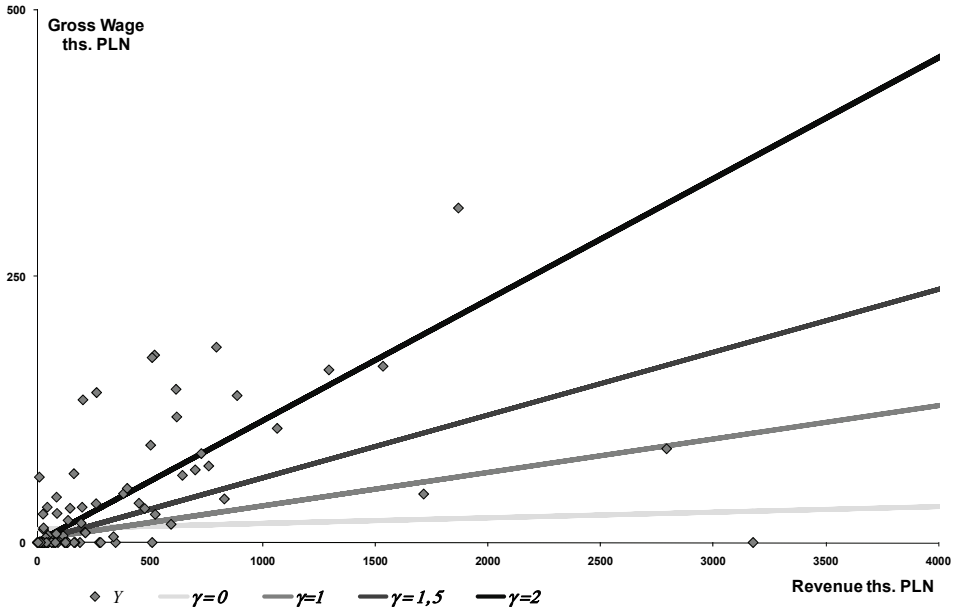
The g_i weights are sample dependent and it is not desirable that the weights are far from unity. Negative weights are particularly undesirable as they tend to increase the estimator variance. That is why some methods for restricting g_i weights were developed. There is also a close relationship between g_i weights of a sample unit and its influence on Beta. The most common measure of this influence is DFBETA defined by a change in the estimate of Beta when a unit is excluded from the sample:

$$DFBETA_i = \left(\sum_{i \in S} x_i x_i^{'} / z_i^{\gamma} \right)^{-1} \left(\frac{x_i}{z_i^{\gamma}} \right) \left(\frac{e_i}{1 - h_i} \right) \tag{6}$$

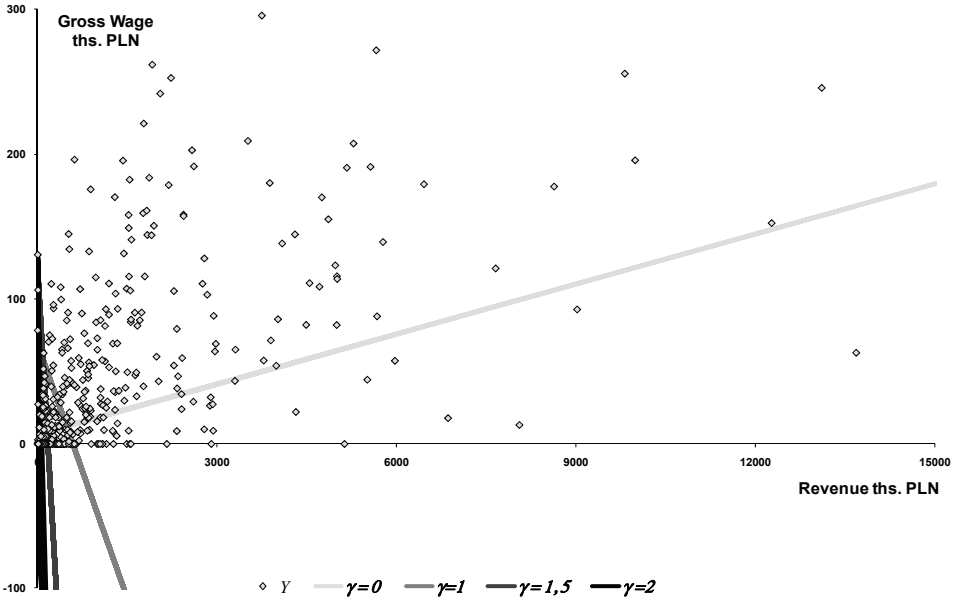
The main reason for negative estimates are outliers, i.e. those companies in which the Gross Wage is very big and the Revenue very small. They influence the model parameters and result in incomparably large weights assigned to those observations. DFBETA was used to identify the influential units. The outlying observations were modified before conducting further estimation.

Figure 2. Regression lines for $\hat{Y}^{\gamma}_{GREG}$ estimators in Dolnoslaskie voivodship:

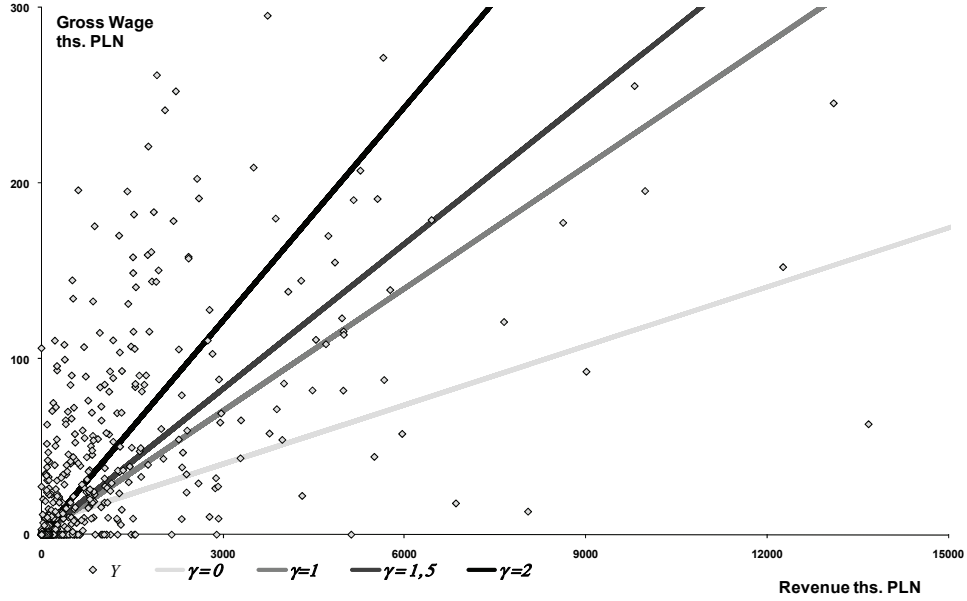
A) domain defined as Industry



B) domain defined as Trade



C) domain defined as Trade after substituting the outliers with median



Source: Own estimation based on SP3 survey data and Tax Register.

To avoid negative estimates, one proposed solution is to substitute outliers with median or implement post-stratification (Chambers *et al.* (2001a)). The observations were divided into three groups: outliers in the dimension of *'x'*, in the dimension of *'y'* and the remaining ones (this group desirably should be the largest). The results obtained by substitution are presented on the figure 2.C and by post-stratification in table 6.The modification considerably changed the model parameters for $\hat{Y}^{1,5}_{GREG}$ and $\hat{Y}^{2}_{GREG}$. Adapting both methods improved the estimation precision as well as reduced the variation of the estimator.

Table 6. Estimation precision for Gross Wage in Trade section of PKD in Dolnoslaskie voivodship

Estimator Coefficient of Variation – CV					*deff coefficient*			
$CV(\hat{Y}_{DIR})$	$CV(\hat{Y}^{0}_{GREG})$	$CV(\hat{Y}^{1}_{GREG})$	$CV(\hat{Y}^{1,5}_{GREG})$	$CV(\hat{Y}^{2}_{GREG})$	$deff(\hat{Y}^{0}_{GREG})$	$deff(\hat{Y}^{1}_{GREG})$	$deff(\hat{Y}^{1,5}_{GREG})$	$deff(\hat{Y}^{2}_{GREG})$
before modification								
0,104	0.086	1.729	-1.377	-1.045	0.8724	6.9087	33.7324	162.167
after substitution of outliers by median								
0,097	0.089	0.088	0.093	0.126	0.9635	1.0119	1.0895	1.5613
after post-stratification								
0,1063	0.1104	0.1111	0.1186	0.1753	1.1393	1.3669	1.5692	2.5939

Source: Own estimation based on SP3 survey data, BJS and Tax Register.

The analysis conducted so far allows us to draw the following conclusions:

- Implementation of Chambers *et al.* (2001a) modification provides the results which are very much differentiated depending on the auxiliary variables and the type of the model,
- Introducing the additional auxiliary variable '*z*' makes it possible to move the estimation problem from disjunctive values of the auxiliary variable to disjunctive values of the estimated variable,
- Application of the additional variable '*z*' creates also the possibility to estimate the basic economic information for small business at a lower aggregation level,
- Weights from the modified model may be characterized by great variation, which is contradictory to the general assumption that the product of weights g_i and w_i should be close to w_i,
- Due to outliers in the sample data, weights designated from the modified model, may obtain values less than zero, which might result in negative estimates,
- Gain in estimation precision is greater for domains of a smaller size (number of observations) and in the case of a stronger correlation between the estimated and the auxiliary variable,
- A significant improvement in estimation precision may be obtained by selecting an adequate model including variables '*x*' and '*z*', with properly designated auxiliary variables, which in the case of a large number of small domains makes application of the modified model rather difficult,
- Weights assigned to the GREG estimator are strongly related to the distance measure DFBETA, which enables identification of outliers,
- Post-stratification is proposed for domains, in case of which selecting a suitable model is difficult.

3. The Winsor estimation

The precision of GREG estimator is influenced by characteristics of enterprises in terms of the distribution of estimated variables and those which serve as auxiliary ones. The existence of outliers in the sample has also an undesirable effect. According to the modification proposed by Chambers *et al.* (2001a), the proportion of disjunctive observations may be reduced by their substitution or post-stratification. Both methods are rather time-consuming and require, in each case, identification of the outlying observations. That is why we also considered other methods, which apply a less individual procedure to solve the problem of outliers. As an example one may consider a group of methods aimed at making the estimator less sensitive to large residuals, which was described by Kokic and Bell (1994) or Chambers (1996). The individuals drawn into the sample, for which the variable takes a value beyond specified border points are changed.

The sample modification may be provided in different ways. One of the methods depends on modification of weights in such a way, that the proportion of outliers in estimation is very small (Hidiroglou, Srinath (1981)). Another method proposes the modification of the value of the estimated variable, for example the Winsor estimation. This estimation technique was introduces by Searls (1966).

The algorithm applied may be divided into several steps. The initial stage comes down to designating some border points K_{Ui} and K_{Li}, which serve to delimitate the individuals into two groups: of typical and untypical observations. In the second step the untypical observations are modified to obtain values which are close to the border points:

$$y_i^* = \begin{cases} \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Ui} & \text{if} \quad y_i > K_{Ui} \\ y_i & \text{if} \quad K_{Ui} \geq y_i \geq K_{Li} \\ \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Li} & \text{if} \quad y_i < K_{Li} \end{cases} \tag{7}$$

The last stage is estimation of the response variable using any estimation technique upon data from the modified sample:

$$\hat{Y}_{win} = \sum_{i \in s} w_i g_i y_i^* = \sum_{i \in s} \widetilde{w}_i y_i^* \tag{8}$$

where: $\widetilde{w}_i = w_i g_i$.

The biggest problem in Winsor estimation is the designation of adequate border points, which are crucial in indicating the outliers. In practice many different methods are proposed, for example robust regression. It is important to

underline that if the procedure applied in this step leads to proper delimitation, the estimation precision might be significantly improved. We decided to determine the limits by using four different techniques of robust regression: *(i) Trimmed Least Squares (TLS), (ii) Trimmed Least Absolute Value (LAV), (iii) Sample Splitting (TSS), (iv) Least Median of Squares (LMS)*. It was also assumed that the proportion of individuals removed from the sample was constant and in the case of each technique equal to 5%. The sample modification was conducted according to Winsor type II estimation and estimates from modified sample were provided by GREG estimator.

The *Trimmed least squares (TLS)* removes from the sample units with the biggest squared residuals. The *Trimmed Least Absolute Value (LAV)* method designates units characterised by the maximal absolute residuals. According to *Sample Splitting (TSS) Technique* the choice of individuals removed from the sample involves some random mechanism. In this method two regression models are estimated for observations drawn to the sample after a random division into two groups. The residuals for one group are calculated upon the model obtained for the other group. The final regression model is constructed after rejecting the observations with the biggest residuals. This approach makes the *TSS* technique more robust in comparison with others like *TLS* or *LAV*. The *Least Median of Squares (LMS)* is based on the bootstrap idea. It depends on selecting subsamples of size $n-1$ according to a simple random scheme with replacement. For each subsample a model is provided, estimates and squared residuals which serve to designate the median. Finally, the model with the smallest median of squared residuals was chosen.

Evaluating the influence of robust regression technique on the estimation precision we considered two approaches: determining both limits: upper and lower, and designation of one limit only - the upper one. The first proposition means that the modification refers to both types of outliers: in *'y'* as well as in *'x'* dimension. While in the second approach the individuals were modified with respect to the outliers in *'y'* dimension only. The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* we used Revenue from Tax Register.

Table 7. Comparison of Winsor and direct estimators precision using techniques: *TLS, LAV, SST, LMS*

Type of Winsor estimator $\hat{Y}_{Win}$ by robust regression technique	Indicator $W_{CV} = \frac{\overline{CV\left(\hat{Y}_{Win}\right)}}{\overline{CV\left(\hat{Y}_{DIR}\right)}}$	
	two border points	one border point
Trimmed least squares TLS	0.804	0.893
Trimmed least absolute value LAV	0.798	0.887
Sample Splitting TSS	**0.386**	**0.475**
Least median of squares LMS	0.843	0.960

Source: Own estimation based on SP3 survey data and Tax Register.

Assigning two border points provides a bigger gain in precision (see tab. 7). The greatest average reduction in *CV* was obtained for *TSS* technique. The W_{CV} ratios for *TLS* and *LAV* techniques are very similar. They indicate about 20% smaller variation than in the case of direct estimator (for two border points). It was found that *Sample Splitting Technique* $\hat{Y}_{TSS}$ provided the most efficient estimates.

To evaluate results of the research, similarly as in the case of GREG modification, the reference values were provided by direct Horvitz-Thompson (HT) estimates. Comparison analysis of Winsor estimators with GREG's modification is presented in Table 8. This time the estimated variable was defined as Revenue. To preserve comparability, we applied an identical set of variables in both GREG modification and Winsor estimator. The results show that the most efficient GREG estimates are obtained for the weight γ=1.5. Both mean and median values of the coefficient of variation are the lowest. But for bigger values of the parameter γ=2 efficiency is decreasing. If median is taken into account, a very high precision (close to the one for γ=1.5) is observed also for γ=0 and γ=1. Regardless of γ parameter, GREG estimator provides more precise estimates (in view of the median) than the direct one. The highest fraction of domains for which GREG estimates were more efficient is for γ=1 but with increase of γ this percentage decreases.

Basic distribution characteristics referring to the variation of Winsor estimator across all domains (**R**egion&**S**ection) show similar *CV* for those using techniques *TLS, LAV* and *LMS* and direct HT estimator. The mean belongs to an interval (0.263 – 0.303) and the median (0.210 – 0.232). The smallest variation is observed for *TSS* technique (mean equals to 0.225 and median 0.172). Also for *TSS* the highest fraction of domains for which it provided more efficient estimates was observed (almost 74%). This record is quite close to GREG with γ=1, and much better than for GREG with γ=1.5. It provides also absolutely better

estimates in terms of other characteristics of *CV distribution* like *min, average and median.*

Table 8. Characteristics of CV distribution: Winsor and modified GREG estimators, revenue, all domains, 2001 r.

Characteristics	*GREG Estimator*				
	$\hat{Y}_{DIR}$	$\hat{Y}^{0}_{GREG}$	$\hat{Y}^{1}_{GREG}$	$\hat{Y}^{1,5}_{GREG}$	$\hat{Y}^{2}_{GREG}$
y - Revenue					
x – Cost from TAX register			***z - Revenue from TAX register***		
min	0,088	0,074	0,075	-7,340	-3,834
max	0,902	1,407	1,417	8,148	4,474
average	0,276	0,273	0,278	0,230	0,298
median	0,232	0,212	0,213	0,210	0, 226
proportion of domains for which $CV < CV_{DIR}$ *(%)*		72,16	74,43	70,45	59,09
Characteristics	*WINSOR Estimator*				
	$\hat{Y}_{DIR}$	$\hat{Y}_{TLS}$	$\hat{Y}_{LAV}$	$\hat{Y}_{TSS}$	$\hat{Y}_{LMS}$
y - Revenue					
x – Cost from TAX register					
min	0,088	0,071	0,070	0,033	0,072
max	0,902	1,371	1,352	1,485	1,578
average	0,276	0,263	0,270	0,225	0,303
median	0,232	0,211	0,210	0,172	0,228
proportion of domains for which $CV < CV_{DIR}$ *(%)*		64,77	64,20	73,86	50,00

Source: Own estimation based on SP3 survey data and Tax Register.

Evaluating the precision of Winsor estimates we analysed a ratio relating mean and median of its *CV* to variation of direct HT estimator. Also *deff* coefficients relating variance of both estimators were considered (see tab. 9). Values that are less than one indicate improvement in estimation precision obtained as a result of implementing the robust approach.

Table 9. Synthetic measures of estimation effectiveness: Winsor and modified GREG estimators of revenue, all domains, 2001 r.

Estimation technique	$W_{\overline{CV}} \frac{\overline{CV\left(\hat{Y}_*\right)}}{\overline{CV\left(\hat{Y}_{DIR}\right)}}$	$W_{Me(CV)} \frac{Me\left(CV(\hat{Y}_*)\right)}{Me\left(CV(\hat{Y}_{DIR})\right)}$	$deff = \frac{Var(\hat{Y}_*)}{Var(\hat{Y}_{DIR})}$
Modified GREG			
$g_i(0)$	0,9891	0,9138	0,9913
$g_i(1)$	1,0072	0,9181	1,0080
$g_i(1,5)$	0,8333	0,9052	0,8334
$g_i(2)$	1,0797	0,9741	1,0795
Winsor Estimator			
TLS	0,9529	0,9095	0,9526
LAV	0,9783	0,9052	0,9786
TSS	0,8152	0,7414	0,8315
LMS	1,0978	0,9828	0,9913

Source: Own estimation based on SP3 survey data and Tax Register.

Comparing $W_{\overline{CV}}$, $W_{Me(CV)}$ and *deff* of different techniques, one should notice the highest reduction obtained for *Sample Splitting Technique*. It is also worth noticing that in terms of median the improvement was observed for all robust techniques, which is not always the case for mean and variance. Estimates obtained for $\hat{Y}^1_{GREG}$, $\hat{Y}^2_{GREG}$ and Winsor using *LMS* are less efficient than direct estimation. Interestingly enough, earlier estimates obtained by Dehnel (2008) for other variables provided other results. This shows ambiguity in the evaluation of robust estimation techniques with relation to estimated variables, availability of auxiliary information etc.

The appraisal of Winsor estimation showed that:

- Depending on the type of estimated characteristic, different methods might be applied to determine the border points used in the delimitation process: two (upper and lower) or one (upper). Both cases were examined showing that the estimation precision is bigger for two limits,
- Simulation research demonstrated the relation between efficiency and type of robust regression technique used. The more robust regression technique was applied, the more efficient estimates were produced. This was shown by comparison of TSS estimates with those obtained by TLS and LAV,
- Limits used in the delimitation process and therefore the type of robust regression technique, influence the precision of Winsor estimator. The highest precision in our study was observed for Sample Splitting Technique (TSS).

4, Local regression estimation

Apart from GREG modification or Winsor estimation, among other methods that are less sensitive to outliers, the local regression technique is often discussed. Its important advantage is its applicability in the case of a nonlinear relation between the estimated and auxiliary variables. Therefore, an attempt was made to present the possibility to implement local regression to estimate economic characteristics of small business for the joint distribution of regions and type of sections of economic activity (Region&Section). The research was conducted using the following estimator (Chambers, Dorfman, Wehrly (1993), Dorfman (2000)):

$$\hat{y}_{loc,i} = \mathbf{c}'_j \left(\mathbf{D}'_i \mathbf{W}_i \mathbf{D}_i\right)^{-1} \mathbf{D}'_i \mathbf{W}_i \mathbf{y}_s \quad i=1, 2, \ldots, N \tag{9}$$

where: U denotes population and s is the sample

$\mathbf{c}'_j$ is the vector of ones at the j-th position while the remaining are equal to zero,

$\mathbf{D}_i$, $i = 1, 2, \ldots, N$, are matrices of the dimension $n \times 2$ each, with $\begin{bmatrix} 1 & (x_j - x_i) \end{bmatrix}$ in the j-th row, $j = 1, 2, \ldots, n$,

$\mathbf{W}_i$, for $i = 1, 2, \ldots, N$, are diagonal matrices of the dimension $n \times n$ and $w_i b_i^{-1} K\left[(x_j - x_i) b_i^{-1}\right]$ at (j, j) position, where $K(\cdot)$ is the kernel function and b_i is the bandwidth for the i-th observation.

The main problems considered in local regression estimation are: to choose the kernel function and to determine the appropriate bandwidth. Once again many methods and suggestions can be found in this field (Chambers, Dorfman, Wehrly (1993), Chambers (1996), Kim, Breidt, Opsomer (2001)). In our study we decided to use the Epanechikov kernel (see also Hedlin, (2004)):

$$K(u_{ji}) = \max\left[0, \frac{3}{4}\left(1 - u_{ji}^2\right)\right] = \max\left[0, \frac{3}{4}\left(1 - \left(\frac{(x_j - x_i)}{b_i}\right)^2\right)\right]. \tag{10}$$

Different definitions were used for the bandwidth, for which the kernel function was determined. One of them was to assign one constant bandwidth for the whole sample according to: $b_i = \frac{1}{4}(x_{\max} - x_{\min})$ => $\hat{Y}_{loc}(\max, \min)$. In other cases, the bandwidth was determined by the value of an auxiliary variable x. Two types of *nearest neighbour bandwidth* suggested by Chambers (1996) were applied: substituting b_i with the difference between the value of the auxiliary variable[1] for the i-th observation (x_i) and for the observation identified by

[1] Units in the sample file were sorted by x_k in ascending order.

i + *20* (x_{i+20}) or *i* + *40* (x_{i+40}): $b_i = x_{i+20} - x_{i-20}$ => $\hat{Y}_{loc}(20)$ and $b_i = x_{i+40} - x_{i-40}$ => $\hat{Y}_{loc}(40)$. Additionally, we assumed the bandwidth to be more narrow than proposed by Chambers, that is : $b_i = x_{i+10} - x_{i-10}$ => $\hat{Y}_{loc}(10)$.

The comparison of robust estimation in business survey was supplemented by the analysis of how the bandwidth definition influences the estimation precision. To preserve comparability, we applied an identical set of variables as in the case of GREG modification and Winsor estimation. The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* we used Revenue from Tax Register.

Table 10. Characteristics of the CV distribution, local regression estimator, all domains, 2001

Characteristics	Estimator				
	$\hat{Y}_{DIR}$	$\hat{Y}_{loc}(10)$	$\hat{Y}_{loc}(20)$	$\hat{Y}_{loc}(40)$	$\hat{Y}_{loc}(\max, \min)$
min	0.10	0.09	0.08	0.09	0.08
max	0.63	0.60	1.77	1.78	0.73
average	0.27	0.25	0.36	0.36	0.27
median	0.25	0.23	0.21	0.21	0.24
proportion of domains for which $CV < CV_{DIR}$ *(%)*		73.30	71.59	72.16	59.66

Source: Own estimation based on SP3 survey data and Tax Register.

The characteristics of relative dispersion distribution for direct and local estimators are comparable, except the maximum. The relatively highest precision is observed for $\hat{Y}_{loc}(10)$ estimator, in terms of the characteristics of the distribution and as concerns the proportion of domains for which *CV* is smaller than obtained for the direct estimator.

The analysis of local regression estimation allowed us to make the following insights:

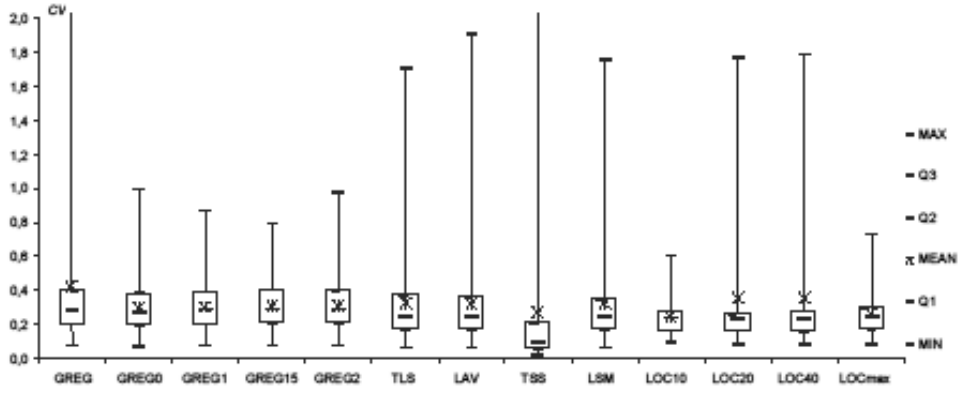
- As the bandwidth increases, the local regression estimates resemble GREG estimates more closely.
- The highest precision, in terms of CV, was obtained for local regression with bandwidth $\hat{Y}_{loc}(10)$ and $\hat{Y}_{loc}(20)$
- For a narrow bandwidth, the estimation is based on many local models, which lengthens the computing time
- As the bandwidth gets increasingly narrow, the local change in the estimated variable is taken into account to a greater degree. As the bandwidth increases, the smoothing effect is more significant

- The weights designated by the kernel function do not depend on the values of the estimated variable but on the auxiliary variables. This means that they can be applied for many estimated variables, in case the set of auxiliary variables is unchanged.

5. Comparison of the results obtained and conclusions

The GREG estimation is very popular and one of the most frequently used. But its application to business statistics is connected with a danger arising from outliers, whose presence results in a significant bias of the estimates (see Hedlin (2004)). Several modifications belonging to 'closer' and 'further' family of GREG estimators were applied and discussed. Among those closely related: estimation based on the model using inverse transformation and Winsor estimation introducing 'border' points to distinguish the outlier observation, were analysed. As concerns the 'further' family, we examined local regression, which has the ability to accommodate local departures from the linear model implemented. The estimation precision was evaluated on the basis of estimator's coefficient of variation *CV*, *RedCV* – coefficient measuring the degree of *CV* reduction and *deff* coefficient (see figures 3–5).

Figure 3. Characteristics of CV distribution across all domains, different robust estimators, 2001



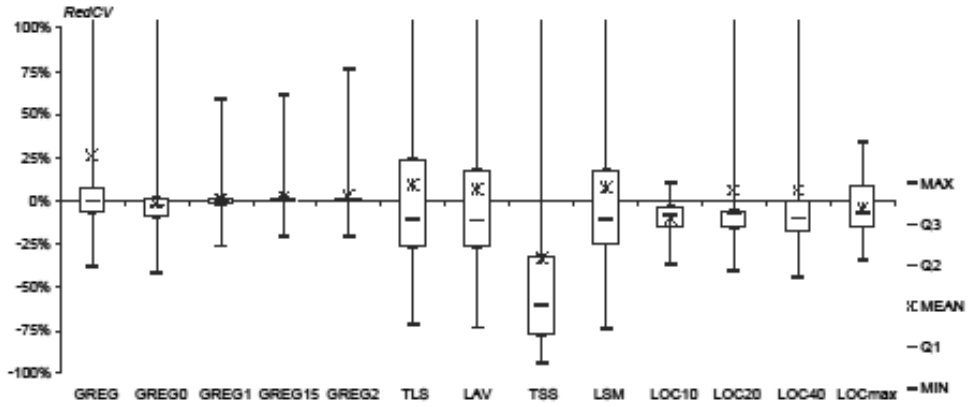
Source: Own estimation based on SP3 survey data, BJS and Tax Register.

The highest estimation precision expressed by the average value and the quartiles of the estimator's CV across all domains, is observed for the Winsor $\hat{Y}_{TSS}$ and local $\hat{Y}_{loc}(10)$ estimators. This remark does not take into account the extreme value observed for Winsor *TSS*, which was due to high dispersion in the

case of one of the bootstrap sub-samples. For the remaining Winsor estimators, local regression and GREG's modification, the relative dispersion, as measured by the quartiles and the average, is just a little bit smaller than for the original GREG estimator, but very similar.

The narrowest range of the CV is observed for two local estimators: $\hat{Y}_{loc}(10)$ and $\hat{Y}_{loc}(\max,\min)$ as well as the GREG modification by Chambers *at al*. For the remaining set of estimators examined, the variation is much more spread, but does not exceed the maximum range observed for the original GREG estimator (*CV=12%*).

Figure 4. Characteristics of *RedCV* distribution across all domains, different robust estimators, 2001



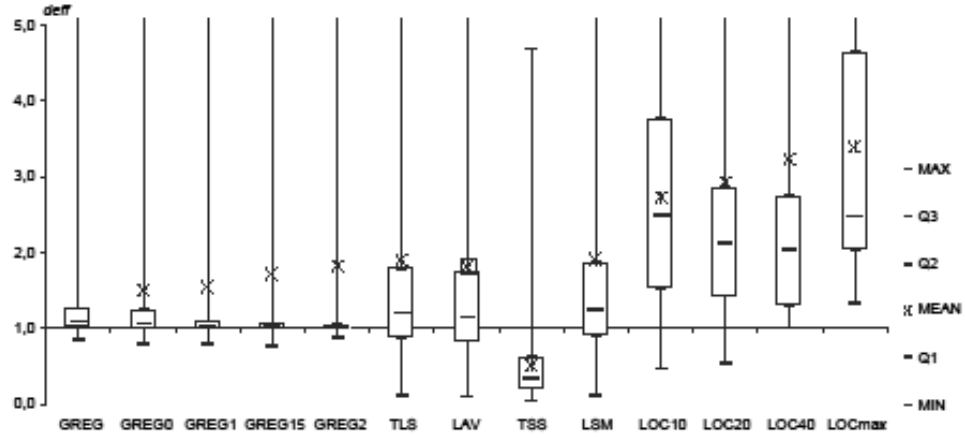
Source: Own estimation based on SP3 survey data, BJS and Tax Register.

Coefficient *RedCV (%)* represents the degree to which the variation of the estimators is reduced in comparison with the direct HT estimator (see figure 4). If the average value across all domains of study, or the quartiles of the distribution are considered, the reduction was obtained for the Winsor *TSS* estimator (on average by 33%, but for median by 61%) and the local regression estimators: $\hat{Y}_{loc}(10)$ (average - by 10%, median - by 9%) and $\hat{Y}_{loc}(\max,\min)$ (average - by 4%, median – by 8%). For the remaining estimation techniques applied, the average value of *CV* is bigger than in the case of direct HT estimator. This means that the techniques used did not result in improving the estimation precision. Generally, it is due to extreme values of *CV* observed for odd domains, which influence the mean. If the median, as a measure independent of outliers, was considered, reduction should be noticed for all Winsor and local regression estimators.

Another coefficient used to evaluate the gain in efficiency of an estimator is *deff*, which compares the variation of the estimator considered with respect to the

direct estimator (see figure 5). If its value is smaller than one, it means that the estimation technique applied is more efficient than the direct one. Evaluating the precision in terms of *deff* coefficient, it can be observed that the greatest gain was obtained by Winsor *sample splitting* estimator *TSS* (on average by 50%). The average variation across all domains, for the whole set of remaining estimators, is greater than that obtained using the direct one. Especially high value of *deff* coefficient is observed for all the local regression estimators. To explain this, we should bear in mind that although local models are constructed, in the estimation process all the observations are taken into account, including outliers. This has a direct influence on high variation of local estimators.

Figure 5. Characteristics of *deff* coefficient distribution across all domains, different estimators, 2001



Source: Own estimation based on SP3 survey data, BJS and Tax Register.

To complete the evaluation of the estimators, the bias analysis was conducted (see Chambers, Brown, Heady, Heasman (2001)). The procedure applied enabled the verification of the random character of the bias of the estimators in comparison with the unbiased direct estimates. Once again, the best results were obtained for the Winsor estimator.

Table 11 presents relation between direct and robust estimates obtained for revenue for a higher aggregation level. The estimates refer to domains defined as PKD Section of economic activity. It shows that local regression (Loc10) and *TSS* provided estimates with the smallest relative difference in case of all domains but one. This refers to GREG (γ =0) in section: *Human health and social work* .

Table 11. Comparison of the results for other levels of aggregation, Revenue, Type of Economic Activity (PKD Sections), 2001

Sections of Economic Activity	Relation between robust and direct estimates								
	GREG	$GREG^1$	$GREG^{1.5}$	$GREG^2$	TLS	LAV	TSS	LSM	LOC_{10}
Manufacturing	109,64	112,42	280,26	84,86	110,47	110,56	108,25	109,71	113,60
Construction	103,15	100,89	277,23	53,58	98,80	97,76	102,24	104,98	99,55
Wholesale and retail trade;	96,28	94,78	-15,30	118,78	96,49	96,45	96,92	95,89	98,75
Accommodation and food service activities	111,01	112,64	194,24	114,63	111,95	118,35	113,11	113,66	108,71
Transportation and storage	101,53	106,72	240,87	92,50	101,41	103,21	101,08	101,28	99,53
Financial and insurance activities	102,92	103,17	206,74	14,03	105,09	104,45	98,30	104,20	93,81
Real estate activities	108,12	113,14	389,46	82,06	105,27	104,22	108,78	107,62	96,98
Education	103,16	105,67	275,31	80,26	98,15	97,49	99,24	99,79	90,75
Human health and social work activities	101,51	102,32	185,47	92,34	105,55	107,75	106,20	103,33	107,17
Other service activities	103,50	105,55	26,16	150,02	113,44	113,99	101,13	106,16	101,39
Number of sections with smallest difference	1	0	0	0	0	0	**3**	1	**5**

Source: Own estimation based on SP3 survey data and Tax Register.

The idea of Winsor estimators differs in comparison with others used in this study in the sense that it modifies the outliers in the direction of the border points, whereas in the GREG's modification, as well as in the case of local regression, values of the auxiliary variables for the outlying observations are not modified. What is changed is their influence on the estimated variable thorough the value of weights. The results of the research show a negative influence of the outlying observations on estimation precision. These conclusions were supported by the correlation analysis between direct estimates and the estimates obtained by applying other robust techniques. In the case of Winsor estimators, a strong correlation is observed ($r \in (0{,}991;0{,}998)$), similarly for GREG modification but for local regression estimators the correlation is much weaker ($r \in (0{,}88;0{,}89)$).

Summing up the evaluation analysis, exploring the distribution of the coefficients *CV*, *RedCV* and *deff* across all domains, together with the bias analysis, the best record was obtained by the Winsor estimator using *Sample*

Splitting Technique (TSS). It is rather difficult to provide the ranking of the remaining estimators, as it changes depending on the criterion. It would be much easier to indicate the 'best' estimator in each group. As concerns the GREG modification it would be $\hat{Y}_{GREG}^{1,5}$, the one with *'z'* variable taking into account heteroscedasticity of the regression of *y* on *x* proportionally to $z_i^{1,5}$. But a well-known drawback of GREG – that it can often provide negative weights – should be borne in mind. Among local regression estimators we would point out $\hat{Y}_{loc}(10)$ with the narrowest bandwidth for which kernel functions were assigned. It was independently determined for each *i*-th observation by the *nearest neighbour bandwidth* technique within $(i \pm 10)$ observations. Characterizing the local regression it should be stressed that the bandwidth significantly influences the computing time. For Winsor estimators, the one implementing *Sample Splitting Technique* $\hat{Y}_{TSS}$ provided the most efficient estimates. In this method two regression models were estimated for sample observations randomly divided into two groups. The evaluation of such models in terms of residuals made us reject the most outlying observations. This approach makes the *TSS* technique more robust in comparison with others like *TLS* or *LAV*.

REFERENCES

BREIDT, F.J., OPSOMER, J.D. (2000) *Local Polynomial Regression Estimation in Survey Sampling.* The Annals of Statistics, **28**, 1026–1053.

CHAMBERS, R.L. (1996) *Robust case-weighting for multipurpose establishment Surveys*, Journal of Official Statistics, Vol.12, No.1, 3–32.

CHAMBERS, R., DORFMAN, A.H., WEHRLY, T.E. (1993) *Bias Robust Estimation in Finite Populations Using Nonparametric Calibration.* Journal of the American Statistical Association, **88**, 268–277.

CHAMBERS, R., KOKIC, P., SMITH, P. and CRUDDAS, M. (2000) *Winsorization for Identifying and Treating Outliers in Business Surveys*, Proceedings of the Second International Conference on Establishment Surveys (ICES II), 687–696.

CHAMBERS R., BROWN G., HEADY P., HEASMAN D. (2001) *Evaluation of Small Area Estimation Methods – an Application to Unemployment Estimates from the UK LFS*, Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective.

CHAMBERS, R.L, FALVEY, H., HEDLIN, D., KOKIC P. (2001a) *Does the Model Matter for GREG Estimation? A Business Survey Example*, Journal of Official Statistics, Vol.17, No.4, 527–544.

DEHNEL G. (2008) *Estymator GREG a estymacja typu Winsora w badaniach mikroprzedsiębiorstw (GREG and Winsor estimation in small business survey), [in:] Statystyka wczoraj, dziś i jutro (Statistics yesterday, today and tomorrow)*, Warszawa, Główny Urząd Statystyczny i Polskie Towarzystwo Statystyczne (Central Statistical Office and Polish Statistical Association), 58–71, Summ. - Bibliogr. ISBN 978-83-7027-431-3 (in Polish).

DORFMAN, A.H. (2000) *Non-Parametric Regression for Estimating Totals in Finite Populations.* Proceedings of the Survey Research Methods. American Statistical Association, 47–54.

EFRON, B. (1979) *Bootstrap methods: Another look at the jackknife*, [in:] Annals of Statistics 7, 1979, 1–26.

FALORSI P. D., FALORSI S., RUSSO A., PALLARA S. (2000) Small Domain Estimation Methods For Business Surveys, Statistics in Transition, June 2000, Vol. 4, No.5, 745–751.

HEDLIN D. (2004) *Business Survey Estimation*, R&D, Sweden.

HIDIROGLOU, M.H., SRINATH, K.P. (1981) *Some estimators of population total from simple random samples containing large units*, JASA, **76**, 690–695.

KIM, J.Y., BREIDT, F.J. and OPSOMER, J.D. (2001) *Local polynomial regression estimation in two-stage sampling.* Proceedings of the Section on Survey Research Methods, American Statistical Association, 55–61.

KISH L. (1965) *Survey Sampling,* Wiley.

KISH L. (1995) *Methods for design effects*, Journal Official Statistics*,* **11,** 55–77.

KLIMANEK T., PARADYSZ J. (2006) *Adaptation of EURAREA experience in business statistics,* "Statistics in Transition", Vol.7, No. 4.

KOKIC, P.N., BELL, P.A. (1994) *Optimal winsorizing cutoffs for a stratified finite population estimator*, Journal of Official Statistics, **10**, 419–435.

MACKIN, C., PRESTON J. (2002) *Winsorization for Generalised Regression Estimation*, Australian Bureau of Statistics.

PAWLOWSKA Z. (2005) *Role of small and medium enterprises in creating a demand on work*, [in:] "Wiadomosci Statystyczne"**,** No.2, 34–46 (in Polish).

SÄRNDAL C.E., SWENSSON B., WRETMAN J. (1992) *Model Assisted Survey Sampling*, Springer Verlag, New York.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 313—328

MODELING OF THE INTERNET SAMPLE

Getka-Wilczyńska Elżbieta[1]

ABSTRACT

In the paper we present the following ideas for the Internet mediated research – respondents' arrival process as a stream of events and a pure birth process and cut models of it for exponential distribution to interpret and analyse some stochastic properties of the Internet data collection process.

Key words: Internet sample, population, Markov process, pure birth process, model.

1. Introduction

Statistical research concerning sampling selection can be divided into representative surveys based on the probability sample (the sample is random and all statistical units should have a strictly positive probability of being selected to the sample) and surveys based on the non-probability sample. After choosing the kind of the sample selection the next stage of the survey is data gathering. In all surveys the data is collected by using an immediate interview, a telephone interview, a mobile phone, a post or by using a computer and in recent years an interview over the Internet (see www.WebSM.org.; www.aapor.org).

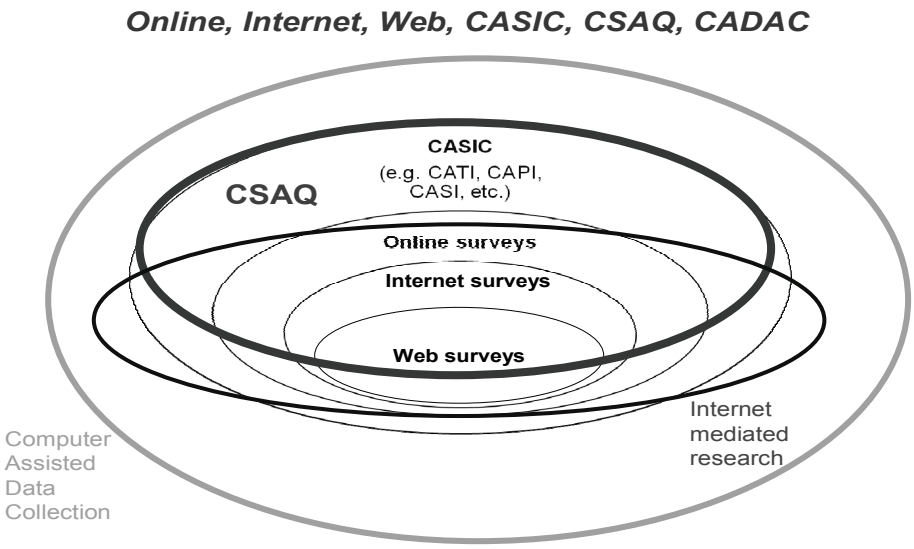
The Internet mediated research is in a process of intensive development and the key characteristic of it is its diversity. Collecting data in these surveys is useful for major corporations, small business, micro-business, academic researches and students, non-profit organizations, political organizations, individuals, government, private research societies and statistical agencies. The Internet mediated surveys have several advantages, such as low cost of collecting information, integration across devices, modes and processes, speed of the data transmission and opportunity to monitor it. Moreover, the computerized nature of Web surveys facilitates conducting experiments. The usage of the electronic questionnaire in the Internet mediated survey makes the interview more efficient, lowers the workload of the respondents and controls the response quality: an elimination of errors during data transcription, an implementation of advanced

[1] Warsaw School of Economics, Institute of Econometrics, Division of Mathematical Statistics, Al. Niepodległości 162, 02-554, Warsaw, Poland, Elzbieta.Getka-Wilczynska@sgh.waw.pl

features: automatic skips and branching, a randomization of questions. The first graphic browser (NCSA Mosaic) was released in 1992, with Netscape Navigator following in 1994 and Internet Explorer in 1995. The first published papers on Web surveys appeared in 1996. Since then, there has been a virtual increasing tendency of interest in the Internet generally, and World Wide Web specifically, as a tool of data collection. A special portal WebSM (www.WebSM.org) – Web survey methodology web site is a website dedicated to the methodology of Web surveys. It has been supported by the EU since 2002 and it includes bibliography lists and software database.

Figure 1 shows competing, overlapping, complementing, synonymous terms: Web surveys, Internet surveys, Internet mediated surveys, Internet mediated research, on-line surveys, CSAQ (computerized self-administrated questioners), CASIC (computer assisted survey information collection), Telesurveys, CADAC (computer assisted data collection).

Figure 1. Modes of data collection by using computer: Online, Internet, Web, CASIC, CSAQ, CADAC (Vehovar, 2007)



But the basic problem in the surveys over the Internet is concerned with collecting and analysing data sets according to classical methods of the sampling theory and statistical inference based on the probability sample.

Generally, differences between representative surveys and Internet mediated researches rely on the following aspects. In representative surveys based on the probability sample the frame of sampling or registers is available, respondents are drawn to the sample by a statistician according to sampling design (sampling

scheme) and the methods of sampling theory and statistical inference are applied to data analysis (e.g. Särndal et al. (1992), Tillě (2006)). If in representative surveys respondents are randomly selected to the probability sample and instead of the traditional modes of interview an electronic questionnaire is used, then it is only one of modes of data collection (one except – complete survey) and the correct usage of this data collection tool requires a suitable survey methodology.

In the most of the Internet mediated researches, the basic problem is concerned with imperfect frames, access to the computer (or the Internet), the self-selection error, issues on the representation, especially with regard to more general population and an application of statistical inference methods to the survey based on the non- probability sample.

If well-defined frames (that is, the current and complete list of identifiable respondents with access to the computer, the Internet) are not available in form suitable for sampling or do not exist, then drawing the probability samples is not possible. For example, drawing the sample is not possible because the preparation of the frame may not be possible – e.g. for users of computers, users of the Internet and different subpopulations of Internet users – users of Skype, Internet portals or Internet websites, consumers of company products, different virtual social groups as focus and chat groups, blogs and bulletin boards and so on (usually the size of the population is not known, the elements of the population are not identified). Sometimes drawing the probability sample is possible (e.g. there exists the frame of households), but there does not exist the frame of users with the access to the computer; there exists the data base of the population (e.g. the administrative data base of scientists), but large costs of preparing of the frame exclude the survey of this population based on the probability sample; the surveyed population is given by the frame, the register or the list, all units have the access to the electronic questionnaire (e.g. small business), all units of the frame are identified and then the frame – based Internet survey (the complete survey of the population) is possible.

The mentioned cases do not describe all possibilities, but generally, if the frame does not exist or is not available then drawing probability sample is impossible and units of the surveyed population cannot be included in the sample according with the classical sampling methods which are useful when it is possible to examine all the units of a finite population.

In these surveys the respondents who belong to the population of interest (the size of the population is known or not) are not randomly selected to the sample, but they participate in the survey according to their subjective decision and data are collected in an uncontrolled way. This is a source of the self-selection error and problem with representativeness of these surveys. The methods of the sampling theory based on the probability sample cannot be used for the data from such samples because the inclusion probabilities are not known (the exception – the complete survey) and statistics are calculated on the basis of the Internet data referring only to the population surveyed.

In theory and practice of the Internet survey two approaches to deal with this problem are identified by Couper and Miller (2008). The first one, the design approach, is mainly associated with the use of panels and is based on the idea of building probability – based on Internet panels by using other methods for sampling and recruitment and, if necessary, providing Internet access to those without. This approach is applied, e.g. by Knowledge Networks in the USA and CentERdata's MESS panel in the Netherlands. The second one, the model based approach, begins with a volunteer – in panel of Internet users, and attempts to correct representation biases using e.g. propensity score adjustment (Lee (2006)) or some other weighting method for assessing Web panel quality (Rosenbaum and Rubin (1983), Callegaro and Disogra (2008), Lee and Valliant (2009), Schonlau et al. (2009)). In both approaches the methodology of sampling theory and statistical inference to data analysis is used. The other interesting proposition is an application of the dynamic theory of decision making and the decision field theory to theoretical explanation of survey behaviour (Galesic (2006)).

Generally, we propose using stochastic processes (the birth and death processes) to study (the description and interpretation, modelling and possible analysis) models of the population and data collecting in the Internet mediated research. In this approach a stochastic nature of events which appear during the Internet mediated survey is presented in following way.

Firstly, when the process of the Internet data collection is treated as a process of registering questionnaires on the server, we assume that the moments of recording the questionnaires form a random sequence of arrivals (or responses) of these respondents who took part in the survey at these moments. This sequence can be modelled in different ways. Moreover, the respondents who took part in the survey form a random subset of the surveyed population called an uncontrolled sample, the Internet sample or the self-selection sample, and the size of this sample is defined as a counting process built on the moments of recording the questionnaires.

Secondly, the Internet mediated survey with the process of the Internet data collection is considered as a life test of the population of the finite size treated as coherent system and the basic characteristics of the population lifetime for proposed models are defined and calculated by using Markov methods (Getka-Wilczyńska (2009)).

The paper is structured as follows. Section 2 contains a notion and a definition of the size of the uncontrolled sample as a counting process. In section 3 the pure birth process and its properties as well as particular cases: the Poisson process and two models for the population of finite size are presented.

2. Preliminaries

We assume that the Internet mediated survey begins at the moment $t = 0$, when the electronic questionnaire is put on the website and the survey is

conducted for the time $T > 0$. A set $\{u_1, u_2, ...,\}$ denotes the population of potential respondents and the respondents fill in questionnaires independently. By $\tau_0 \leq \tau_1 \leq \tau_2 \leq ...$ are denoted the successive random moments of recording the questionnaires on the server after an initial moment $t = 0$, $\tau_0 = 0$. In this case, τ_k, $k \geq 1$, is interpreted as the random moment of an arrival (or a response) of the respondent u_k, $k \geq 1$, who belongs to the population of potential respondents.

For each $n \geq 1$ the population of finite size *n* is surveyed and the size of the population is random. In this case τ_k, $k = 1,2,...,n, n \geq 1$, is interpreted as the random moment of an arrival (or a response) of this respondent u_j, $j = 1,2,...,n, n \geq 1$, belonging to the population of size n, $n \geq 1$, who took part in the survey as the k th.

We assume that X_k, $k = 1,2,...,n$, $n \geq 1$, are nonnegative random variables with distribution function $F_k(t) = P(X_k \leq t)$, for $t \geq 0$, $k = 1,2,...,n$, $n \geq 1$, the probability density function $f_k(t) = F_k'(t)$, $F_k(t) = \int_0^t f_k(x)dx$.

Random variable X_k, $k = 1,2,...,n$, $n \geq 1$ is interpreted as a waiting time of the arrival (the response) of the *k*th respondent up to time *t* when the size of the uncontrolled sample is determined or a lifetime of *k*th element of the population of interest up to time *t* when the population lifetime is considered.

2.1. Random size of the uncontrolled sample

If the process of Internet data collection is considered as a process of registering questionnaires on the server in a fixed interval of the time $T > 0$ (the time of the survey conducted) then the size of the uncontrolled sample at the moment $t \geq 0$ equals total number of respondents' arrivals up to moment $t \geq 0$ and is defined as a counting process $\{N(t), t \geq 0\}$ built on the sequence $(\tau_0, \tau_1, \tau_2....)$ as follows.

We assume that no arrival has occurred up to time $t = 0$.

The number of respondents' arrivals up to time $t \geq 0$ is given by

$$N(t) = card\{k \geq 1 : \tau_k \leq t\} = \max\{k \geq 1 : \tau_k \leq t\}$$

and satisfies the conditions of definitions 1.1-3.1 below (Resnick (1998)).

Definition 2.1 The process $N = \{N(t) : t \geq 0\}$ is called a counting process (stream of events) if for all $t, h \geq 0$: 1. $N(0) = 0$ 2. $N(t) \in \mathrm{N}$ 3. $N(t) \leq N(t+h)$

The increments $N(t+h)-N(t)$ of the process N models the number of arrivals occurring in the interval $(t,t+h]$. Realizations of counting process are monotonically non-decreasing and right-continuous functions. The process N is completely defined, if for $n \geq 1$ and any nonnegative numbers $t_1, t_2,...,t_n$ a probability distribution of a random vector $(N(t_1), N(t_2),..., N(t_n))$ is determined.

Definition 2.2 (equivalent to definition 2.1). Let $\tau_1, \tau_2,....$, be successive moments of occurrence of events, $\tau_{k-1} \leq \tau_k$ for $k \geq 1$ and $\tau_0 = 0$.

Random variables $T_k = \tau_k - \tau_{k-1}, \quad k \geq 1$ denote the time between the *k*th and (*k*-1)st records of questionnaires and $\tau_k = \sum_{j=1}^{k} T_j$, $k \geq 1$, $T_0 \equiv 0$.

The process $N = \{N(t): t \geq 0\}$ is called a counting process (stream of events), if for all $n \geq 1$ and any nonnegative integer numbers $t_1, t_2,...,t_n$ a probability distribution of a random vector $(T_1, T_2,..., T_n)$ is determined.

Definition 2.3 Let $(T_k)_{k\geq 1}$ be a sequence of nonnegative random variables.

We denote $\tau_n = T_1 + T_2 + ... + T_n$ for $n \geq 1$, $\tau_0 = 0$

$$N(t) = card\{k \geq 1: \tau_k \leq t\} = \max\{k \geq 1: \tau_k \leq t\}.$$

A sequence $(\tau_n)_{n\geq 0}$ is called a renewal stream (renewal sequence).

The process $N = \{N(t): t \geq 0\}$ is called a renewal process and is completely defined if for each $n \geq 1$ and any nonnegative numbers $t_1, t_2,...,t_n$ a probability distribution of a random vector $(T_1, T_2,..., T_n)$ is determined.

Definitions 2.1–2.3 determine an infinite stream of respondents' arrivals i.e. for any integer number $M \geq 0$ exists the number $t \geq 0$, so that $\Pr(N(t) \geq M) > 0$.

For each $t \geq 0$ the value of the counting variable $N(t)$ is called the size of the uncontrolled sample (the Internet sample) at the moment *t*. Because the relation between the sequence $(\tau_0, \tau_1, \tau_2....)$ and the process $\{N(t): t \geq 0\}$ is given by $\{N(t) = n\} = \{\tau_n \leq t < \tau_{n+1}\}$ hence

$$P_n(t) = \Pr\{N(t) = n\} = \Pr\{\tau_n \leq t < t_{n+1}\} = \left\{ \sum_{k=1}^{n} T_k \leq t < \sum_{k=1}^{n+1} T_k \right\}$$

for $n \geq 0, t \geq 0$.

A general computational expression for the probabilities $P_n(t)$ is impossible because the interdependences $(T_n)_{n\in N}$ between the successive arrivals (responses) are not specified.

If we assume that $(T_n)_{n\in N}$ is the sequence of independent random variables we have the following possibilities:

1. $F_k(t) = P(T_k \le t)$, $k \ge 1$
2. $F_k(t) = F(t)$ for $k \ge 2$, $F_1(t) = P(T_1 \le t)$, the general renewal process
3. $F_k(t) = F(t)$, $k \ge 1$, the renewal process
4. $F_k(t) = F(t) = 1 - e^{-\lambda t}$, $\lambda > 0, t \ge 0$, $k \ge 1$, the Poisson process, the simplest renewal process.

Below are presented two cases of the infinite stream of the respondents' arrivals: a homogenous pure birth process (1) and a particular case of it – a homogenous Poisson process (4) as well as some cut models of it for the population of finite size.

3. The pure birth process

During Internet mediated surveys we observe arrivals of the respondents belonging to the population $\{u_1, u_2, ...,\}$ at the moments τ_k, $k \ge 0$ and for each $t \ge 0$ the process $N = \{N(t) : t \ge 0\}$. The random variable $N(t)$ (equals the number of arrivals up to time t), takes value 0,1,2,…

The Internet sample during the survey forms a random set of these respondents who participate in the survey. We say that the Internet sample is in the state E_n when $N(t) = n$ and $E = \{E_0, E_1, E_2, ...\}$ is a finite or a countable set of possible states of the Internet sample (E_0 denotes the state of the Internet sample in which $N(t) = 0$, E_1 denotes the state of the Internet sample, in which $N(t) = 1$,…. and so on). The transition from state E_n to state E_{n+1} occurs at the moment τ_n, $n \ge 1$, and denotes an increase of the size of the Internet sample by one. The Internet sample passes successively through the states $E_0, E_1, E_2 .....$. In the state $E_n, n \ge 0$ it is for the time $T_{n+1}, n \ge 0$, respectively, with the distribution $F_{n+1}(t) = P(T_{n+1} \le t)$ and next passes to the state E_{n+1} with the probability one.

According to the description stated above we postulate that the probability of the arrival of the respondent occurring at a given instant of time depends upon the number of the arrivals which have already occurred. It means that if during the interval $(0,t)$ n arrivals occur, then the probability of the new arrival in the

interval $(t, t+h)$ is equal to $\lambda_n h + o(h)$, where $\lim_{h \downarrow 0} \frac{o(h)}{h} = 0$ and the process $N = \{N(t) : t \geq 0\}$ is characterized by a sequence of positive numbers $\lambda_0, \lambda_1, \ldots$

Formally, the process $N = \{N(t) : t \geq 0\}$ is a Markov point process built on the sequence τ_n, $n \geq 0$, a homogeneous pure birth process. In this process the jumps intensity at any time *t* depends on the number of jumps before time *t* and does not depend on *t*. The process satisfies the following postulates:

1. N has the Markov property
2. $\Pr(N(h) = 1) = \Pr(N(t+h) - N(t) = 1 | N(t) = k) = \lambda_k h + o_{1,k}(h), \quad k \geq 0,$ $h \to 0+,$
3. $\Pr(N(h) = 0) = \Pr(N(t+h) - N(t) = 0 | N(t) = k) = 1 - \lambda_k h + o_{2,k}(h), \quad k \geq 0,$
4. $N(0) = 0,$
5. $\Pr(N(t+h) - N(t) < 0 | N(t) = k) = 0, k \geq 0.$

The left sides of 2 and 3 are $P_{k,k+1}(h)$ and $P_{k,k}(h)$ respectively, so that $o_{1,k}(h), o_{2,k(h)}$ do not depend upon *t.* Because the size of the Internet sample is defined by

$$N(t) = card\ \{n \geq 1:\ \tau_n \in [0,t]\} = \max\left\{ n \geq 1 : \sum_{j=1}^{n} T_j \leq t \right\}$$

we are interested in the derivation of $P_n(t) = \Pr(N(t) = n), n \geq 0$.

The system of Kolmogorov differential equations satisfied $P_n(t) = \Pr(N(t) = n)$ for $t \geq 0$ is given by $P_0'(t) = -\lambda_0 P_0(t)$ $P_n'(t) = -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t)$, $n \geq 1$, with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$. (e.g. Feller, t. I. p. 370, 1977; the solution, t. II, p. 431, 1978; Gnedenko et al. (1968)).

The first equation can be solved immediately and yields $P_0(t) = \exp(-\lambda_0 t) > 0$.

Since T_k is the time between the *(k-1)*th and *k-th* records (arrivals, responses), so that $P_n(t) = \left\{ \sum_{k=1}^{n} T_k \leq t < \sum_{k=1}^{n+1} T_k \right\}$ and $\tau_n = \sum_{k=1}^{n} T_j$ is equal to the time at which the *n*th record of the questionnaire (the arrival, the response) occurs. Therefore, $\Pr(T_1 \leq t) = 1 - \exp(-\lambda_0 t)$. The postulates 2-5 imply that random variable $T_k, k \geq 2$ has exponential distribution with parameter λ_{k-1} and T_k'*s* are mutually independent.

Therefore, the characteristic function of $\tau_k = \sum_{j=1}^{k} T_j$ is given by

$$\varphi_{\tau_k}(t) = E\left(\exp(-it\,\tau_k\right) = \prod_{j=1}^{k} E\left(\exp\left(-it\,T_j\right)\right) = \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\lambda_{j-1} + it}$$

$$\text{and} \quad Ee^{-\alpha\tau_k} = \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\lambda_{j-1} + \alpha} \quad \text{for } \alpha \neq 0 .$$

For a specific set of $\lambda_k \geq 0$ the solutions of the system of differential equations is given by the following formulae

$$P_k(t) = \lambda_{k-1} \exp\left(-\lambda_k t\right) \int_0^t \exp\left(\lambda_k x\right) P_{k-1}(x) dx , \; k = 1,2,\ldots .$$

Hence, all $P_k(t) \geq 0$ for any $k \geq 0$ and $t \geq 0$.

But there is a possibility that for some sequences $\lambda_0, \lambda_1, \ldots$ $\sum_{k=1}^{\infty} P_k(t) < 1$.

The intuitive argument for this fact is as follows: if the sum $\sum_{k=1}^{\infty} P_k(t)$ interpreted as the probability of the finite number of changes of the population states (records, arrivals, a response) up to the time t, then the difference $1 - \sum_{k=1}^{\infty} P_k(t)$ may be treated as the probability of the infinite number of the changes of the population states up to time t and, the process $N = \{N(t) : t \geq 0\}$ may explode, i.e. with positive probability $N(t) = \infty$. The necessary and sufficient conditions excluding the possibility of explosion is the equality (Feller, t. I, p. 373, 1977; t. II, p.431, 1978) $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} = \infty$.

More formal argument for this result is as follows: the time T_k between consecutive arrivals (responses) has exponential distribution with parameter λ_k. Therefore, the quantity $\sum_{k=0}^{\infty} \frac{1}{\lambda_k}$ equals the expected time before the populations become infinite. By comparison, $1 - \sum_{k=1}^{\infty} P_k(t)$ is the probability that $N(t) = \infty$.

If $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} < \infty$ the expected time for the population to became infinite is finite. It is then plausible that for all $t > 0$ the probability that $N(t) = \infty$ is positive.

If all $\lambda_0, \lambda_1, ... \lambda_{n-1}$ are distinct and positive numbers then the finite dimensional distribution of the process $\{N(t) : t \geq 0\}$ is given by formulae

$$P_k(t) = \Pr(N(t) = k) = \left(\prod_{j=1}^{k} \lambda_{j-1} \right) \sum_{i=1}^{k} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k} (\lambda_{j-1} - \lambda_{i-1})} -$$

$$- \left(\prod_{j=1}^{k+1} \lambda_{j-1} \right) \sum_{i=1}^{k+1} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k+1} (\lambda_{j-1} - \lambda_{i-1})}$$

for $k = 1,2,...,n-1$ and $P_0(t) = \Pr(N(t) = 0) = \Pr(T_1 > t) = e^{-\lambda_1 t}$ for $k = 0$.

If $\lambda_n = 0$ we have $P_n(t) = \Pr(N(t) = n) = \left(\prod_{j=1}^{n} \lambda_{j-1} \right) \sum_{i=1}^{n} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k} (\lambda_{j-1} - \lambda_{i-1})}$.

Finally, for this model the size of the uncontrolled sample is given by definition.

Definition 3.1 For $n \geq 1$ the size of uncontrolled sample until the moment $t \geq 0$ is given by

$$N(t) = card \{ n \geq 1 : \tau_n \in [0,t] \} = \max \left\{ n \geq 0 : \sum_{k=1}^{n} T_n \leq t \right\},$$

where $\tau_{k-1} \leq \tau_k$ for $k \geq 1$ are the successive moments of questionnaires record, $\tau_0 = 0$, $T_k = \tau_k - \tau_{k-1}$ for $k \geq 1$ are independent random variables with exponential distribution $Exp(\lambda_k)$.

The homogeneous pure birth process model is very interesting because it gives great flexibility for modelling the distribution of the number of arrivals as well as its evolution in time by proper choice of the sequence (λ_k). Different models determined by choosing specific sequence (λ_k) are given below.

Model 1. If $\lambda_1 = \lambda_2 = ... = \lambda_n = ... = \lambda$ then the process $N = \{N(t) : t \geq 0\}$ is a well known Poisson process (Kingman (1993)).

If the data collecting process is observed only for the population of finite size n we assume that the numbers $\lambda_1,...\lambda_n,....$ determining the homogeneous pure birth process are distinct and positive numbers and $\lambda_{n+1} = \lambda_{n+2} = ... = ... = 0$.

In this case the sequence τ_k, $k = 1,2,...,n$, $n \geq 1$, forms a limited stream of the respondents' arrivals and is obtained by a superposition (summing) of the streams of n respondents where each respondent generates a stream consisting of exactly one questionnaire. The particular cases of models for the population of the finite size are given below.

Models for the population of finite size

Model 2. If $\lambda_1,...\lambda_n$ are distinct and positive numbers and $\lambda_{n+1} = \lambda_{n+2} = ... = ... = 0$ then we obtain a general homogeneous pure death process.(Getka-Wilczyńska, 2009).

This model satisfies the following conditions:

1. N has the Markov property
2. $\Pr\big(N(t+h) - N(t) = 1 \big| N(t) = k-1\big) = \lambda_k h + o(h), \quad k \geq 1,\ h \to 0+,$
3. $\Pr\big(N(t+h) - N(t) = 0 \big| N(t) = k-1\big) = 1 - \lambda_k h + o(h), \quad k \geq 1,$
4. If $\Pr\big(N(t) = n\big)$ then the Internet survey ends and $\lambda_{n+1} = 0$

For this model the system of Kolmogorov differential equations satisfied by $P_n(t) = \Pr\big(N(t) = n\big)$ for $t \geq 0$ is given by

$$P_1'(t) = -\lambda_1 P_1(t),$$
$$P_k'(t) = \lambda_{k-1} P_{k-1}(t) - \lambda_k P_k(t),\ k = 2,3,...n,$$
$$P_{n+1}'(t) = \lambda_n P_n(t)$$

with boundary conditions $P_1(0) = 1, \quad P_k(0) = 0, k > 1$.

The solution is given by the formulae

$$P_{n+1}(t) = 1 - \lambda_1 \lambda_2 ... \lambda_n \sum_{k=1}^{n} \frac{e^{-\lambda_k t}}{\lambda_k w'(-\lambda_k)},$$

where $w(x) = (x + \lambda_0)(x + \lambda_1)...(x + \lambda_n)$.

Model 3. $\lambda_k = \lambda$ for $0 \leq k \leq n$ and $\lambda_k = 0$ for $k > n$. The series $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} = \infty$ is divergent, if any $\lambda_k = 0$. Then $\sum_{k=1}^{\infty} P_k(t) = 1$ and the system of Kolmogorov differential equations is given by

$$P_0'(t) = -\lambda P_0(t)$$
$$P_k'(t) = -\lambda P_k(t) + \lambda P_{k-1}(t) \text{ for } 1 \leq k \leq n,$$

$P'_{n+1}(t) = \lambda P_n(t)$ for $k = n+1$ with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$.

The solution is given by the equations

$$P_0(t) = e^{-\lambda t}, P_1(t) = \lambda t e^{-\lambda t}, P_2(t) = e^{-\lambda t}\frac{(\lambda t)^2}{2!} \ldots\ldots, P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{k!},$$

$$P_{n+1}(t) = 1 - \sum_{k=0}^{n} e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Model 4. $\lambda_k = (n+1-k)\lambda$ for $0 \le k \le n$ and $\lambda_{n+1} = 0$

The system of Kolmogorov differential equations is given by

$P'_0(t) = -(n+1)\lambda P_0(t)$

$P'_k(t) = -(n+1-k)\lambda P_k(t) + (n-k+2)\lambda P_{k-1}(t)$ for $1 \le k \le n$,

$P'_{n+1}(t) = \lambda P_n(t)$ for $k = n+1$ with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$.

The solution is given by equations

$P_0(t) = e^{-\lambda(n+1)t}$, $P_1(t) = (n+1)te^{-\lambda nt}\left(1 - e^{-\lambda t}\right), \ldots,$

$P_n(t) = (n+1)e^{-\lambda t}\left(1 - e^{-\lambda t}\right)^n,$

$P_{n+1}(t) = \left(1 - e^{-\lambda t}\right)^{n+1}.$

Properties of model 4

Let $\underline{X} = (X_1, \ldots, X_n)^T$, $n \ge 1$ be a sample of independent and identically random variables selected from the population having an exponential life distribution $F(t) = 1 - e^{-\lambda t}$. $X_k, k = 1,2,\ldots,n$, $n \ge 1$ is interpreted as the waiting time of an arrival (or a response) of u_k th respondent belonging to the population of the size n, $n \ge 1$ and $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ denotes the random vector of order statistics of the sample $\underline{X}$.

In the Internet survey we observe the successive arrivals of respondents at the moments of recording the questionnaires on the server

$$\tau_1 \le \tau_2 \le \ldots \le \tau_n, \quad \tau_0 = 0,$$

where $\tau_1 = X_{1:n}, \tau_2 = X_{2:n}, \ldots, \tau_n = X_{n:n}$ and

$T_1 = X_{1:n} \sim Exp(n\lambda)$, $T_2 = X_{2:n} - X_{1:n} \sim Exp((n-1)\lambda)$,…, $T_n = X_{n:n} - X_{(n-1):n} \sim Exp(\lambda)$, $T_{(0)} = X_{(0)} \equiv 0$.

Properties of the sequences $(T_1,...,T_n)$ and $(\tau_1,...,\tau_n)$(e.g. Barlow, Proschan (1975), Feller, v. II, (1978)).

1. $P(T_k \le t) = F_{(n-k+1)\lambda}(t) = 1 - e^{-(n-k+1)\lambda t}$ for $k = 1,2,...,n$, $T_1,...,T_k$ are independent random variables, $E\,T_k = \frac{1}{(n-k+1)\lambda}$, $Var(T_k) = \frac{1}{[(n-k+1)\lambda]^2}$,

$$f_{(T_1,T_2,...,T_n)}(t_1,...,t_n) = n!\prod_{i=1}^{n}\lambda\, e^{-\lambda(t_1+...+t_n)} = \prod_{i=1}^{n}(n-i+1)\lambda\, e^{-\lambda(n-i+1)t_i}.$$

2. $E(\tau_k) = E(X_{(k)}) = E\sum_{i=1}^{k}(X_{i:n} - X_{(i-1):n}) = \sum_{i=1}^{k}E(X_{i:n} - X_{(i-1):n}) = \sum_{i=1}^{k}\frac{1}{\lambda(n-i+1)}$,

$$Var(\tau_k) = Var(X_{k:n}) = Var\sum_{i=1}^{k}(X_{i:n} - X_{(i-1):n}) = \sum_{i=1}^{k}Var(X_{i:n} - X_{(i-1):n})$$

$$= \sum_{i=1}^{k}\frac{1}{[\lambda(n-i+1)]^2},$$

$$f_{\tau_k}(t) = \frac{n!}{(k-1)!(n-k)!}\lambda\left(1-e^{-\lambda t}\right)^{k-1} e^{-\lambda(n-k+1)t}.$$

Interpretation: Because for fixed $n \ge 1$ the size of the Internet sample at the moment $t \ge 0$ is given by

$$N(t) = card\,\{1 \le k \le n: \tau_k \in [0,t]\} = \max\{1 \le k \le n : \tau_k \le t\} = \sum_{k=1}^{n} \mathrm{I}_{[0,t]}(\tau_k),$$

and the value of the random variable $N(t), t \ge 0$ equals the number of arrivals up to the time $t \ge 0$, the process $\{N(t),\, t \ge 0\}$ can be described in the following way. Each respondent fills in only one questionnaire with the probability 1 in the interval $[0,T]$, independently from the others. The probability of filling in the questionnaire by certain respondent in the interval $[0,t] \subset [0,T]$ is equal to $P(X_k \le t) = 1 - e^{-\lambda t}$. In this way each respondent generates a stream consisting of only one questionnaire. The summary stream obtained by summing these streams forms a bound Bernoulli stream and consists of a finite number of events.

3. Since X_k, $k = 1,2,...,n$, $n \geq 1$, are independent random variables with exponential distribution $F(t) = 1 - e^{-\lambda t}$, $\lambda > 0, t \geq 0$

then $\Pr(N(t) = k) = \Pr(\tau_n \leq t < \tau_{n+1}) = \binom{n}{k}\left(1 - e^{-\lambda t}\right)^k \left(e^{-\lambda t}\right)^{n-k}$.

Hence, the number of respondents' arrivals up to the time t has binomial distribution with parameters n and $1 - \exp(-\lambda t)$, $N(t) \sim Bin(n, 1 - \exp(-\lambda t))$.

4. The process $\{N(t) : t \geq 0\}$ has no independent increments and it is a Markov process with transition probability: for $0 \leq t_1 \leq t_2 \leq .... \leq t_k$

$$\Pr\{N(t_k) - N(t_{k-1}) = k \mid N(t_1), N(t_2),..., N(t_{k-1})\} =$$

$$= \binom{n - N(t_{k-1})}{k}\left(1 - \exp(-\lambda(t_k - t_{k-1}))\right)^k \exp(-\lambda(t_k - t_{k-1}))^{n - N(t_{k-1}) - k}$$

5. The finite dimensional distribution of the process $\{N(t) : t \geq 0\}$ is equal to

$$\Pr(N(t_1) = n_1, N(t_2) = n_2,..., N(t_k) = n_k) = \binom{n}{n_1}\left(1 - e^{-\lambda t_1}\right)^{n_1} e^{-\lambda(n - n_1)t_1} \times$$

$$\times \binom{n - n_1}{n_2 - n_1}\left(1 - e^{-\lambda(t_2 - t_1)}\right)^{n2 - n_1} e^{-\lambda(n - n_2)(t_2 - t_1)} \times ... \times$$

$$\binom{n - n_{k-1}}{n_k - n_{k-1}}\left(1 - e^{-\lambda(t_k - t_{k-1})}\right)^{n_k - n_{k-1}} e^{-\lambda(n - n_k)(t_k - t_{k-1})}$$

with the covariance function

$$Cov(N(t), N(t+s)) = n \exp(-\lambda(t+s))(1 - \exp(-\lambda t)).$$

4. Conclusion

In this paper some stochastic models to the interpretation and study of Internet mediated research (survey) are proposed. Firstly, we assume that the random sequence of the moments of questionnaires record observed in the Internet data collection process is treated as the stream of events. Next, the size of the Internet sample is described as a counting process built on the points of this sequence. It is a homogeneous pure birth process and there are different versions of it – the Poisson process and the cut models for the population of finite size. The pure birth process and the Poisson process to study the countable population are proposed, the cut models of these processes to study the population of finite size.

Acknowledgments

The research was partially supported by the Ministry of Science and Higher Education programs 03/S/0074/08 and 03/E/0013/09, Warsaw School of Economics.

REFERENCES

BARLOW, R.E, PROSCHAN F. (1975). Statistical theory of reliability and life testing. Holt, Rinehart and Winston Inc., New York 1975.

CALLEGARRO M., DISOGRA CH. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 5, Special Issue, 1008–1032.

COUPER M.P., MILLER P.V. (2008). Web survey methods. *Public Opinion Quarterly*, 72, 5, Special Issue, 831–835.

FELLER W. (1977, 1978). An introduction to probability theory and its applications (in Polish). Vol. I, II, Polish Scientific Publisher, Warsaw.

GALESIC M. (2006). Dropout on the Web: effects of interest and burden experiences during an online survey. *Journal of Official Statistics*, 22, 313–328.

GETKA-WILCZYŃSKA E. (2009). Mathematical modeling of the Internet survey. Engineering the Computer Science and IT ed. by Safeeullah Soomro, In-Teh, Croatia.

http://www.sciyo.com/articles/show/title/mathematical-modeling-of-the-internet-survey.

GNEDENKO B. W., BIELAJEV A., SOLOWIEV A.D. (1968). Mathematical methods in reliability theory (in Polish). Technical –Scientific Publisher, Warsaw.

KINGMAN J. F.C. (1993). Poisson processes. Clarendon Press Oxford.

LEE S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web Surveys. *Journal of Official Statistics*, 22, 2, 329–349.

LEE S., VALIANT R. (2009). Estimation for volunteer panel Web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research*. Vol.37, No 3, February, 319–343.

RESNICK S.I. (1998). Adventure in stochastic processes. Birkhäuser Boston.

ROSENBAUM P.R., RUBIN D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometricka*, 70, 41–55.

SÄRNDAL C.E., SWENSSON B., WRETMAN J. (1992). Model assisted Survey Sampling. Springer-Verlag New York Inc., New York.

SCHONLAU M., van SOEST A., KAPTEYN A., COUPER M. (2009). Selection bias in Web surveys and the use of propensity scores. *Sociological methods and research*, Vol. 37, No. 3, 291–318.

TILLÈ, Y. (2006). Sampling algorithms. Springer Science-Business Media, Inc., New York.

VEHOVAR, V. (2007). Workshop on Internet Survey Methodology. Ullehammer, September 2007.

http://vv.ris.org, www.WebSM.org

http://www.aapor.org

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 329—359

CAUSALITY ANALYSIS BETWEEN PUBLIC EXPENDITURE AND ECONOMIC GROWTH OF POLISH ECONOMY IN LAST DECADE

Henryk Gurgul, Łukasz Lach[1]

ABSTRACT

This paper investigates the causal links between different kinds of budgetary expenditure and the economic growth of Poland. The empirical analysis was based on both the linear and nonlinear Granger causality tests and the aim was to evaluate the applicability of Wagner's Law and contrasting theory formulated by Keynes. We based our research on aggregate and disaggregate quarterly data with the sub–division of public expenditure on human resources (HR), physical resources (PR), net interest payment (NIP) and other remaining budgetary expenditure (OTHER) for the period Q1 2000 to Q3 2008. Linear causality analysis showed that relation between total budgetary expenditure and economic growth is consistent with Keynesian theory. However, for the examined sub–categories of expenditure mixed results were reported supporting Keynesian theory (NIP), Wagner's Law (OTHER) or none of them (HR and PR). Results of nonlinear causality analysis performed for unfiltered data also provided some support for Keynesian theory (HR and OTHER). However, after GARCH(1,1)–filtration of data nonlinear causality was not reported in any case.

Key words: government expenditure, linear and nonlinear causality, bootstrap techniques.

1. Introduction

In recent years one could observe rapid increase of government expenditure in both highly developed and emerging economies. This tendency was also reported in the case of transition countries e.g. Poland. Apart from the last year, it was possible in Poland to observe a rise in government expenditure along with a high growth in GDP rate. It may be of interest for economists to answer the following question: which of the two variables – economic growth and government

[1] Department of Applications of Mathematics in Economics, AGH University of Science and Technology, Cracow, Poland. E–mail addresses: henryk.gurgul@gmail.com, lukilach1983@o2.pl.

expenditure – is a causal factor for other one? The proper way to answer this question is the application of suitable causality analysis. In economics there are two main theories related to the interactions between public expenditure and economic growth. These contrasting propositions are known as Wagner's Law and Keynes's theory. Undoubtedly, the great achievement of Adolf Wagner was the establishment of a positive relationship between economic growth and public expenditure. An increase in economic activity reflected in economic growth leads to a rise in government activity, which in turn leads to public expenditure. This regularity is one of the best known explanations of Wagner's Law and leads to the further insight that government expenditure is an endogenous factor in economic growth. On the contrary, Keynes, claimed that public expenditure was an exogenous factor which could be used to influence economic growth. Further discussion and empirical investigations contained in this paper are dedicated to examining the applicability of these contrasting theories in the case of Polish economy.

This paper is organized as follows. In the next section we present the literature overview. Section 3 contains the formulation of the main research hypotheses to be tested by means of empirical analysis. Section 4 provides dataset description, which sets the context for the rest of the paper. In section 5 description of methodology of linear and nonlinear causality tests, details of the bootstrap technique as well as the results of some preliminary variable analysis are presented. Section 6 contains the outcomes of the causality analysis. The last section contains some final remarks.

2. Literature overview

The issue of examining causal links between public expenditure and economic growth has been a subject of number of papers in recent years. These studies dealt with both individual countries as well as groups of countries (applying panel datasets). The results vary significantly as for some countries the government expenditure was found to lead the economic growth, while for the others the opposite causal link was established. Naturally, for some countries there are papers reporting the existence of bidirectional causality or lack of causal link in any direction. The variety of conclusions depends surely on differences of the political and economic systems of the countries under study. At this place the sensitiveness of test results to sample size should also be mentioned. In this section various results of examination of GDP–expenditure links will be laid out. Alongside the total budgetary expenditure we shall also focus on its main sub–categories. Division of total public expenditure may be carried out in many ways. Researchers usually divide total expenditure on its main components, namely human resources expenditure, physical resources expenditure, net interest payment expenditure and other expenditure. However, there are also many studies

which focus on the link between economic growth and one specific expenditure area e.g. education, national defence etc.

The examination of Wagner's Law for the economy of Taiwan was a subject of Pluta's (1979) contribution. This research provided no evidence in favour of this theory. Demirbas (1999) examined the long–run relationship between government expenditure and GNP in the light of Wagner's Law for Turkey over the time period 1950–1990. Similarly to the previously mentioned paper in this case the suitable analysis also contradicted Wagner's Law. Mixed results for the direction of causality between GDP and public expenditure were delivered by contributions by Sinha (1998) for Malaysian data (1952–1992) and by Jackson et al. (1998) for Northern Cyprus data (1977–1996). Some results support Wagner's Law, while others favour Keynesian theory. On the other hand, contributions by Park (1996) for Korean data, by Khan (1990) for Pakistani data and the results obtained by Nagarajan and Spears (1990) for Mexican data provided convincing basis for the acceptance of Wagner's Law for the examined economies.

One of the most extensive research projects about the linkage between public expenditure and economic growth was conducted by Anwar et al. (1996). The authors examined 88 countries using unit root and co-integration techniques (over the period 1960–1992) finding unidirectional causality in 23 cases and bidirectional causality in 8 cases. The main conclusion arising from this contribution is the fact that for the majority of analyzed countries the causality between GDP and public expenditure does not run in any direction.

Saunders (1985) performed an analysis of links between economic growth and public expenditure for OECD countries for the period 1960–1981. In general the results of this research provided evidence in favour of Keynesian theory, however some weak evidence of causality running in opposite direction was also reported.

Every paper discussed so far (except for Pluta's (1979) contribution) dealt with aggregate data. However, there are also studies which investigate the relationship between the growth of GDP and specific sub–categories of public expenditure. In further part of this section some of the most important papers on this subject, not necessarily conducted by means of causality analysis, will be briefly reviewed.

Our review will start with the relationships between GDP and human resources. At this point we cannot neglect the fact that the main factor which is believed to have an impact on human capital is the level of education. Many economists claim that education increases the quality of human resources because it improves people's productivity and therefore speeds up economic growth. According to the endogenous growth theory the creation of new products or ideas is a function of human capital. The latter is reflected in accumulated skills, training and general knowledge. One of the most popular explanations of the influence of human capital on economic growth is as follows. The rise in government expenditure on research and development causes growth in physical capital which in turn is a direct stimulus to economic growth.

The notion that causation may run in the opposite direction, that is from economic growth to human resources (i.e. to education), is also relatively common. Investment in capital stock may result in sufficient economic growth providing the surplus which is necessary for further investment in the education sector. Many economists (e.g. Easterly et al. 1994, Caselli 1999) claim that demand for highly qualified staff is stimulated by investment in capital stock and new technologies. Moreover, some studies support the thesis that human resources and new technologies are complementary. In some papers it is pointed out that higher education supports the tendency towards the reduction in the current earnings in favour of higher future economic growth. The analysis of causal links between human resources and economic growth has been the subject of numerous contributions. De Meulemeester and Rochat (1995) performed an analysis of causality between higher education and economic growth in six countries. This research included Sweden, the United Kingdom, Japan, France, Italy and Australia and was conducted over different time periods. The causality running from higher education to economic growth was established in the case of Sweden, the United Kingdom, Japan and France while for Italy and Australia no causality was reported. For all analyzed countries the null hypothesis of no co-integration could not be rejected.

The relationship under study was also examined by In and Doucouliagos (1997), Asteriou and Agiomirgianakis (2001) and by Narayan and Smyth (2006). The last paper concerns causal relations between higher education, real income and real investment. The final conclusion following from this paper is that an increase in the rate of graduation from higher education has a positive effect on real income growth and on real investment.

It is a well known fact that one of most important factors which have significant influence on human capital is the shape of heath care system. Recently, the relationship between gross domestic product and health care expenditure has also been the subject of number of contributions, e.g. Culyer (1990), Hitris and Posnett (1992), Hansen and King (1996, 1998), Blomqvist and Carter (1997), McCoskey and Seldon (1998) Roberts (1999. The main problem arising from the mentioned literature is the inconclusiveness of empirical results. Devlin and Hansen (2001) examined GDP–health care relationship for 20 OECD countries. It was reported in the literature that health care expenditure appeared to cause GDP in some cases. However, for some countries evidence of causality running in the opposite direction was also reported.

The relationship between GDP and other sub–division expenditure has also been intensively examined in recent years. The paper by Benoit (1978) is believed to be the leading contribution concerning the effect of defence expenditure on economic growth. After some time, this relationship received a considerable attention and further research was performed by many economists (e.g. Smith 1980, Frederiksen and Looney 1982, Deger and Sen 1983, Lim 1983, Biswas and Ram 1986, Grobar and Porter 1989, Chen 1993, Kollias 1994, Dunne et al. 2005, Heo and Eger 2005, Lai et al. 2005, Reitschuler and Loening 2005, Kalyoncu and

Yucel 2006, Narayan and Singh 2007). Each empirical study provided some basis for describing causal relationships between analyzed variables and was a trysail for answering a question whether defence expenditure is associated with higher or lower GDP growth rates. Some researchers claim that the net effect of defence expenditure on economic growth is positive (Chang et al. 2001) while others treat defence expenditure as a reason for reduced savings and investment which lead to reduced economic growth (see e.g. DeRouen 1995 and Landau 1996). Generally, in most of the papers the Keynesian theory is believed to be a proper pattern for description of this relationship. However, there are also studies (e.g. Joerding 1986, Kalyoncu and Yucel 2006) which provide evidence that causation runs from economic growth to defence expenditure (this was found for Turkey, but not for Greece). In the more recent contribution to the subject by Narayan and Singh (2007) the authors report that their findings are consistent with the Keynesian school of thought.

The association between economic growth and public expenditure in the United States of America was examined by Liu et al. (2008). In general, the results of this research (conducted for aggregated and disaggregated data) are in line with Keynesian theory. The authors also formulate some policy recommendation claiming that the US government should invest more money in human resources in order to stimulate GDP growth.

The literature overview presented above refers only to some part of research on GDP–public expenditure links. One can easily see that this literature presents various points of view. Taking them into account in the next section of this paper we will formulate our main research hypotheses which are to be tested by means of causality analysis.

3. Main conjectures

The main goal of this paper is an investigation of the causal links between total public expenditure and economic growth as well as between economic growth and expenditure on selected sub–categories of public expenditure, namely human resources, physical resources, net interest payment and remaining expenditure. One important point that distinguishes our paper from other contributions on public expenditure and economic growth is that we did not use annual data but quarterly data. This is because the data covers only a few recent years due to the lack of previous years' data. Therefore, in order to get a sufficiently large data sample we chose quarterly data in spite of their high fluctuation. The two main hypotheses on causality between public expenditure and economic growth are known as Wagner's Law and Keynesian theory. The theories have contrasting propositions. Thus, it is of interest to test the following:

Hypothesis 1: Total public expenditure in Poland is an endogenous factor in economic growth, i.e. Wagner's Law holds true for Poland.

From the above literature overview it seems obvious that in numerous contributions no evidence in favour of Wagner's Law was found. In some of them causality was found in the exact opposite direction. Hence, the formulation of the following hypothesis seems to be justified:

Hypothesis 2: Total public expenditure is an exogenous factor influencing the economic growth of the Polish economy, i.e. Keynesian theory applies.

Economic growth is the basis for a rise in public expenditure which in turn stimulates economic growth in subsequent time periods. This theoretical possibility, especially in the case of aggregate data, seems to be more likely. Some authors report the existence of feedback between government expenditure and economic growth, i.e. that both Wagner's Law and Keynesian theory hold true. Therefore, we will also check the existence of feedback in the Polish economy.

After checking Wagner's Law and Keynesian theory for total public expenditure and economic growth we will examine the association between economic growth and the main sub–categories of public expenditure, i.e. human resources, physical resources, net interest payment and remaining expenditure. Since most of the previous studies provided relatively convincing support for claiming that for GDP and expenditure on human resources (or expenditure on main components of human resources) Keynesian theory holds true, we shall test the following:

Hypothesis 3: Spending on human resources Granger cause economic growth.

In the literature there are many quite conflicting results on the relation between expenditure on physical resources and economic growth. Some contributors claim that expenditure on physical resources has a positive effect on economic growth, while others claim that this effect is negative. Yet other contributors claim that economic growth is the source of expenditure growth on physical resources. The similar discussion is also ongoing in the case of the relationship between GDP and net interest payment. However, in some recent contributions the GDP–PR expenditure and GDP–NIP expenditure relationships are generally said to be in line with both the Keynesian theory and Wagner's Law. Therefore we formulate the next hypothesis for Poland:

Hypothesis 4: Expenditure on physical resources (net interest payment) Granger causes economic growth. Causality runs in opposite direction too.

Recent papers provided relatively convincing support for claiming that expenditure on remaining budgetary areas (aggregated in variable denoted as OTHER) are caused by movements of GDP growth rate in advanced economies. It may be interesting to check this regularity in case of analyzed emerging economy. Thus, we shall consider testing the following hypothesis for Poland:

Hypothesis 5: Economic growth Granger causes OTHER expenditure.

In order to check the size of the reaction of GDP growth rate on one s.d. shock we employed the Impulse Response Function technique. We formulate the next conjecture as follows:

Hypothesis 6: GDP growth is sensitive to one s.d. shocks for the categories of expenditure under study.

Since most of previous studies dealt with standard linear Granger causality tests it may be interesting to check whether the nature of causal link between GDP and budgetary expenditure is indeed linear. This may have an important practical meaning for Polish policymakers in terms of transporting fluctuations of expenditure to economic growth or vice versa. Therefore, we formulate the last conjecture as follows:

Hypothesis 7: All existing causal links between GDP and budgetary expenditure are strictly linear.

In the next section we describe the dataset applied.

4. Dataset overview

In this section we present a brief description of the dataset used in further computations. At the moment we shall underline some important facts. Firstly, contrary to the most of previous contributions concerned with GDP–expenditure links our analysis is not limited only to one specific relationship. Beside total budgetary expenditure we examined four budgetary sections. Hence, the defined dataset includes quarterly data of real growth rates of GDP, total public expenditure as well as budgetary expenditure on four subcategories, namely human resources, physical resources, net interest payment and other remaining expenditure for the period Q1 2000 to Q3 2008. The dataset contains 35 observations. All growth rates are calculated in comparison to corresponding quarter of previous year. Secondly, the application of real growth rates gives us the opportunity to examine the links between the variables of interest which are not affected by movements of the inflation rate. We followed a simple procedure to calculate real growth rates for variables describing budgetary expenditure. We first calculated the GDP deflators for all quarters (with the help of nominal and real GDP) and then we applied these quantities to filter out the impact of inflation on time series of budgetary expenditure. The following formula was used to calculate real growth rate of budgetary expenditure (it is in line with the method of calculating the real GDP growth rate):

$$BE_t^r(x) := \frac{\frac{BE_t(x)}{deflator_t} - BE_{t-4}(x)}{BE_{t-4}(x)} \cdot 100\% \qquad (1)$$

where $BE_t^r(x)$ denotes the real growth rate of budgetary expenditure on x (i.e. one of five possibilities) in quarter t, $BE_t(x)$ denotes the value of budgetary expenditure on x in quarter t (expressed in current prices) and $deflator_t$ denotes the value of GDP deflator in quarter t ($deflator_t := GDP_t \cdot (GDP_t^c)^{-1}$, where GDP_t stands for GDP in quarter t expressed in current prices and GDP_t^c stands for GDP in

quarter *t* expressed in constant prices of the previous year). It is easy to see that in order to construct all time series of real growth rates of budgetary expenditure for the period Q1 2000 to Q3 2008 the quarterly data of budgetary expenditure for the period Q1 1999 to Q3 2008 had to be used. The quarterly data describing GDP in Poland in the period under study was obtained from the Central Statistical Office in Poland, while time series of the budgetary expenditure (total and sub–categories) were collected from the Ministry of Finance of Poland.

Another important fact that distinguishes this paper from previous contributions concerned with similar topic is the application of quarterly data. Most of previous papers were based on applications of annual data. However, the application of lower frequency data may not be adequate for testing for short–run Granger causality between chosen variables, as some important interactions may stay hidden (for more details see e.g. Granger 2000). The data on GDP is published once a quarter, therefore the application of higher frequency data is not possible. In further parts of this paper we use abbreviations for all examined variables. Table 1 contains suitable information. Additionally, some short description of each variable is also presented:

Table 1. Abbreviations and short description of examined variables

Shortcut name	Description
GDP	Real GDP growth rate in Poland
BUDGET	Real growth rate of total budgetary expenditure in Poland
HR	Real growth rate of human resources expenditure in Poland (including education, health care, social security, sport, culture etc.)
PR	Real growth rate of physical resources expenditures in Poland (including manufacture, mining, construction of buildings, forestry, wholesale and retail trade, transport and communications, production and supply of energy, services, information technology etc.)
NIP	Real growth rate of net interest payment expenditure in Poland (including public debt service, subsidies for specific economic tasks, financial inflows from institutions and from individuals and connected spending etc.)
OTHER	Real growth rate of other expenditure in Poland (including science, public administration, public safety, national defence, justice administration, agriculture and fishing etc.)

In order to provide basic information about our dataset we present some descriptive statistics of all examined variables. For each time series some typical quantities were calculated. The following table contains suitable results:

Table 2. Descriptive statistic of examined variables

Variable / Quantity	GDP [%]	BUDGET [%]	HR [%]	PR [%]	NIP [%]	OTHER [%]
Minimum	0.50	–4.84	–18.65	–44.87	–13.54	–7.37
1st quartile	2.40	–0.47	–3.73	–11.65	–3.07	1.27
Median	4.40	4.36	2.78	4.66	5.16	5.04
3rd quartile	6.20	9.52	11.90	38.76	8.56	9.11
Maximum	7.50	20.63	32.44	219.32	18.34	37.32
Mean	4.25	4.98	4.70	19.97	3.66	6.43
Std. Deviation	2.09	6.82	11.75	51.33	6.79	9.90
Skewness	–0.30	0.42	0.26	2.14	–0.31	1.32
Excess kurtosis	–1.16	–0.62	–0.41	5.29	–0.02	1.94

We can remark some interesting information directly from this table. Firstly, we can see that in the period under study relatively stable development could be observed in Polish economy, since the real GDP growth rate was positive in each quarter. Moreover, periods of rapid development (GDP growth at the level of 7.50%) as well as stages characterized by relatively slow growth rate (at the level of 0.50%) were also perceived. In each quarter the total public expenditure was on average almost 5% greater than in the corresponding quarter of the previous year. Also one can easily note that the values of real growth rates of budgetary expenditure are much more varied than GDP growth. The biggest drop in expenditure (in comparison to corresponding quarter of previous year) was reported for physical resources expenditure time series and reached the value of 44.87%. The highest growth was also reported for PR series (219.32%, this huge value was reported for first quarter of 2001 when expenditure on sub–category under study reached value of over 3 million PLN in comparison to just one million PLN in corresponding quarter of the previous year). Furthermore, we shall note that the standard deviations of all time series are relatively large (except for GDP growth rate). All these facts together seem to prove that in the period under study the growth rates of expenditure on the chosen budgetary sections have evolved dynamically. This phenomenon could be interpreted as the effect of whole gamut of system transformation for the financing of the crucial budgetary sections which took place in Poland in recent years.

5. Methodology and preliminary analysis

In this article we use both the linear and nonlinear Granger causality tests to explore the short–run dynamic relationships between real growth rates of GDP, expenditure on major budgetary sections and total public expenditure in Poland.

The main goal of our analysis is to investigate which one of the theories presented in the introductory section, namely the Wagner's approach or Keynes's theory, seems to be more adequate for the case of Polish economy.

The definition of causality used in this paper is due to Granger (1969). One stationary time series, say X, is said to strictly Granger cause another stationary one, say Y, if past and current values of series X are helpful in predicting future values of time series Y. The definition of causality was intentionally formulated for stationary time series. As it was shown through empirical (Granger and Newbold 1974) and theoretical (Phillips 1986) deliberations, if the time series under study are indeed nonstationary then the results of typical linear causality tests may lead to spurious conclusions. Thus, testing the chosen time series for stationarity and identifying their order of integration is the initial part of standard causality analysis. In the first step we conducted Augmented Dickey–Fuller (ADF) unit root test. Table 3 contains results of ADF test with deterministic term including either constant or constant with linear trend. Before conducting the test we had set up the maximal lag length equal to 6 and then we used AIC information criterion to choose optimal lag length from the set {0, 1, …, 6}:

Table 3. Results of ADF tests (levels)

Variable	Only constant		Constant and linear trend	
	test statistic (*p*–value)	optimal lag	test statistic (*p*–value)	optimal lag
GDP	–1.54 (0.51)	4	–2.56 (0.29)	2
BUDGET	–6.28 (0.00)	0	–6.25 (0.00)	0
HR	–1.68 (0.43)	2	–1.58 (0.80)	2
PR	–2.73 (0.07)	3	–2.79 (0.20)	3
NIP	–5.38 (0.00)	0	–5.70 (0.00)	0
OTHER	–4.38 (0.00)	0	–5.45 (0.00)	0

From table 3 one can easily notice that only GDP, HR and PR time series were found to be nonstationary (at 5% significance level), regardless the form of deterministic term. However, at this place we should underline few important facts. Firstly, the results of ADF test are relatively sensitive to the incorrect establishment of lag parameter. Secondly, as it was shown in some papers this test tends to under–reject the null hypothesis pointing too often at nonstationarity. The low power against stationary alternatives was frequently reported (see e.g. Agiakoglu and Newbold (1992)). Therefore, the Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test was additionally conducted to check the results of ADF tests. Results of KPSS test are presented in table 4:

Table 4. Results of KPSS test of examined variables (levels)

Variable	With constant (test statistic*)	With constant and linear trend (test statistic**)
GDP	0.54	0.08
BUDGET	0.32	0.13
HR	0.22	0.14
PR	0.13	0.12
NIP	0.24	0.09
OTHER	0.55	0.04

* *critical values: 0.347 (10%), 0.463 (5%), 0.739 (1%).* ** *critical values: 0.119 (10%), 0.146 (5%), 0.216 (1%).*

As we can see results presented in table 4 lead to relatively different conclusions than the outcomes contained in table 3. This time only GDP and OTHER time series were found to be nonstationary (at 5% significance level). However, when time component was additionally included then the test pointed at stationarity (trend stationarity) of variables under study (also at 5% significance level).

The relatively different results of both tests forced us to use third test, namely the Phillips–Perron (PP) test. This test is based on nonparametric method of controlling for serial correlation when testing for a unit root. The initial point of this procedure is the equation of non–augmented DF test with lag parameter equal to zero. Then special modification of *t*–ratios is applied to avoid the influence of serial correlation on asymptotic distribution of test statistic (for more details see Phillips and Perron (1988)). We shall note that the null hypothesis refers to non- stationarity. The results of PP test are presented in following table:

Table 5. Results of PP test of examined variables (levels)

Variable	Only constant (*p*–value):	Constant and linear trend (*p*–value):
GDP	0.38	0.18
BUDGET	0.00	0.00
HR	0.00	0.00
PR	0.00	0.00
NIP	0.00	0.00
OTHER	0.00	0.00

After analyzing outcomes presented in table 5 one can easily see that all time series except for GDP were found to be stationary at reasonable significance levels. In order to make the final decision about orders of integration of all variables we re–run all applied tests for their first differences. Of course, this was

performed only in those cases for which the results of test conducted for the variables in their levels pointed at non-stationarity. Suitable outcomes are presented in the following table (Δ denotes differencing operator):

Table 6. Results of tests of stationarity of examined variables (first differences)

Variable	ADF with constant		ADF with constant and linear trend	
	test statistic (*p*–value)	optimal lag	test statistic (*p*–value)	optimal lag
ΔGDP	–2.96 (0.03)	3	–2.76 (0.21)	3
ΔHR	–9.69 (0.00)	1	–4.87 (0.00)	1
ΔPR	–3.83 (0.002)	4	–4.57 (0.001)	5
Variable	KPSS with constant (test statistic)		KPSS with constant and linear trend (test statistic)	
ΔGDP	0.14		0.11	
ΔOTHER	0.06		0.03	
Variable	PP with constant (*p*–value)		PP with constant and linear trend (*p*–value)	
ΔGDP	0.001		0.01	

The main conclusion arising from the analysis of outcomes presented in table 6 is the fact that the order of integration of all variables is no greater than one (assuming no deterministic trend). Taking into consideration all results presented in tables 3–6 we may state that only GDP time series is indeed integrated of order one (around constant). The non-stationarity of GDP was found in the results of all conducted tests while for other variables at least two of three conducted tests have pointed at stationarity. This final conclusion will be crucial for our further analysis as it is initial point of causality testing.

In this paper we use the Toda–Yamamoto (TY) approach to test for short–run linear Granger causality. This method has been commonly applied in recent studies (see e.g. Wolde–Rufael (2006)) since it is relatively simple to perform and free of complicated pretesting procedures, which may affect the test results especially while dealing with nonstationary variables. Another issue worth underlying is the fact that this method is useful for testing for causality between variables which are characterized by different orders of integration (which is true for most cases analyzed in this paper). In such cases the linear causality analysis cannot be performed by the application of suitable VEC model (as variables are characterized by different orders of integration).

In order to understand the idea of Toda–Yamamoto approach for causality testing consider the following *n*–dimensional VAR(*p*) process:

$$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t \tag{2}$$

where $y_t=(y_t^1,...,y_t^n)^{tr}$, $c=(c_1,...,c_n)^{tr}$ and $\varepsilon_t=(\varepsilon_{1,t},...,\varepsilon_{n,t})^{tr}$ are n–dimensional vectors (tr denotes transpose operator) and $\{A_i\}_{i=1}^{p}$ is a set of $n\times n$ matrices of parameters for appropriate lags. The order p of the process is assumed to be known. If this value is unknown then it may be established with the help of some standard statistical methods (like application of consistent model selection criterion, for more details see e.g. Paulsen (1984)). The Toda–Yamamoto (1995) idea of testing for causal effects is based on estimating the augmented VAR($p+d$) model:

$$y_t=c'+\sum_{i=1}^{p+d}A_i'y_{t-i}+\varepsilon_t' \tag{3}$$

In order to use Toda–Yamamoto approach we shall assume that the error vector ε' is an independent white noise process with nonsingular covariance matrix $\Sigma_{\varepsilon'}$ (whose elements are assumed to be constant over time). Additionally, we shall also assume that the condition $E\left|\varepsilon_{k,t}'\right|^{2+s}<\infty$ holds true for all k=1,…,n and some s>0. The value of parameter d is equal to the maximum order of integration of variables $y^1,...,y^n$. According to Toda and Yamamoto (1995) the number of extra lags (parameter d) is an unrestricted variable since its role is to guarantee the use of asymptotic theory. We say that the k–th element of y_t does not Granger–cause the j–th element of y_t ($k,j\in\{1,...,n\}$) if there is no reason for the rejection of the following hypothesis:

$$\mathrm{H}_0:\ a_{jk}^s=0 \quad \text{for s=1, …, p,} \tag{4}$$

where $A_s=\left[a_{pq}^s\right]_{p,q=1,...,n}$ for s=1, ..., p (note that the considered hypothesis of non–causality refers to non–augmented VAR model (i.e. model (2)). In order to present the test statistic we shall make use of the following compact notation (T denotes the sample size, circumflex indicates the OLS estimator):

Table 7. Compact notation used to formulate TY test statistic

Object	Description
$Y:=(y_1,..,y_T)$	$n\times T$ matrix
$\hat{D}:=(\hat{c},\hat{A}_1,...,\hat{A}_p,...,\hat{A}_{p+d})$	$n\times(1+n(p+d))$ matrix
$Z_t:=\begin{bmatrix}1\\ y_t\\ y_{t-1}\\ ...\\ y_{t-p-d+1}\end{bmatrix}$	$(1+n(p+d))\times 1$ matrix, $t=1,\ldots,T$
$Z:=(Z_0,..,Z_{T-1})$	$(1+n(p+d))\times T$ matrix
$\hat{\delta}:=(\varepsilon_1',..,\varepsilon_T')$	$n\times T$ matrix

The initial step of TY procedure is the calculation of $S_U:=\frac{\hat{\delta}\hat{\delta}^{tr}}{T}$ – the variance–covariance matrix of residuals from unrestricted augmented VAR model (i.e. model (3)). Then one may define $\beta:=vec(c,A_1,..,A_p,0_{n\times nd})$ and $\hat{\beta}:=vec(\hat{c},\hat{A}_1,..,\hat{A}_p,..,\hat{A}_{p+d})$ where $vec(\cdot)$ denotes column stacking operator and $0_{n\times nd}$ stands for $n\times nd$ matrix filled with zeros. Using this notation one can write the Toda–Yamamoto test statistic for testing causal effects between variables in y_t in the following form:

$$\mathrm{TY}:=\left(C\hat{\beta}\right)^{tr}\left(C\left(\left(ZZ^{tr}\right)^{-1}\otimes S_U\right)C^{tr}\right)^{-1}\left(C\hat{\beta}\right) \tag{5}$$

where $\otimes$ denotes Kronecker product and C is the matrix of suitable linear restrictions. In our case (i.e. testing for causality from one variable in y_t to another) C is $p\times(1+n(p+d))$ matrix which elements take only the value of zero or one. Each of p rows of matrix C corresponds to restriction of one parameter in β. The value of every element in each row of C is one if the associated parameter in β is zero under the null hypothesis, otherwise it is zero. There is no association between matrix C and last n^2d elements in β. This approach allows us to formulate the null hypothesis of Granger non–causality in the following form:

$$\mathrm{H}_0: C\beta^{tr}=0. \tag{6}$$

Finally we shall note that the TY test statistic is asymptotically χ^2 distributed (which holds true if previously mentioned assumptions like properties of error term, etc. are fulfilled) with the number of degrees of freedom equal to

number of restrictions to be tested (in our case this value is equal to p). In other words TY test is just a standard Wald test applied for first p lags obtained from augmented VAR($p+d$) model.

As we have already mentioned the application of Toda–Yamamoto method requires some specific modelling assumptions. If these assumptions are fulfilled then the test statistic is asymptotically chi–square distributed. At this place we should mention some drawbacks of this method of testing for Granger causality. Firstly, if the error term of augmented VAR model is not a white noise (e.g. heteroscedastic) then the application of asymptotic theory may lead to spurious results. Secondly, even if modelling assumptions are generally fulfilled the distribution of TY test statistic may be significantly different from chi–square while dealing with extremely small samples. In order to avoid these problems we have decided to use bootstrap technique additionally. This method is used for estimating the distribution of test statistic by resampling data. At this point we shall also underline some important facts. Firstly, the estimated distribution depends only on available dataset, therefore it may be reasonable to expect that none of the assumptions required for parametric methods has to be fulfilled for proper application of bootstrap technique. Secondly, the size and power properties of causality test based on bootstrap techniques remain relatively good even in cases of nonstationarity and various schemes of error term structure including heteroscedasticity etc. (for more details see Dolado [a]nd Lütkepohl (1996), Mantalos (2000), Hacker [a]nd Hatemi (2006), Lach (2010)). However, we may not forget that bootstrap methods have some drawbacks too and hence they cannot be treated as perfect tools for solving all possible model specification problems. The bootstrap approach is likely to fail in some specific cases and therefore should not be used without second thought (see e.g. Horowitz (1995), Chou and Zhou (2006)).

Every bootstrap simulation conducted for the use of this article is based on resampling leveraged residuals. We have decided to use leverages as this is just a simple modification of regression raw residuals which supports stabilization of their variance (more details on leverages may be found in Davison and Hinkley (1999)). For every pair of variables we estimated non–augmented bivariate VAR model through OLS methodology with the assumed null hypothesis that one variable does not Granger cause the other one. In fact this means that some elements of coefficient matrices were restricted to zero. In the next step we used leverages to transform regression raw residuals (vector of residuals modified by this transformation will be denoted as $\{\hat{\varepsilon}_i^m\}_{i=v_0,\ldots,T}$, T stands for sample size, v_0 is equal to VAR lag length plus one). Finally, the following algorithm was conducted:

- Drawing randomly with replacement (all points have equal probability $p_0=\frac{1}{T-v_0+1}$) from the set $\{\hat{\varepsilon}_i^m\}_{i=v_0,\ldots,T}$ (as a result the set $\{\hat{\varepsilon}_i^{**}\}_{i=v_0,\ldots,T}$ was obtained);

- Subtracting the mean in order to guarantee the mean of bootstrap residuals is zero (this way we create the set $\{\hat{\varepsilon}_i^*\}_{i=v_0,\ldots,T}$, such that $\hat{\varepsilon}_{k,i}^* = \hat{\varepsilon}_{k,i}^{**} - \frac{\sum_{j=v_0}^{T} \hat{\varepsilon}_{k,j}^{**}}{T-v_0+1},\ i=v_0,\ldots,T,\quad k=1,2$);
- Generating the simulated data through the use of original data, coefficient estimates from the regression of restricted non–augmented VAR model and the bootstrap residuals $\{\hat{\varepsilon}_i^*\}_{i=v_0,\ldots,T}$;
- Performing the TY procedure (for simulated data).

After repeating this procedure *N* times it was possible to create the empirical distribution of TY test statistic and to get empirical critical values (bootstrap critical values) next. In order to take part in academic discussion on how the number of bootstrap replications (parameter *N*) may affect performance of bootstrap techniques we examined several possibilities for this parameter. The suitable procedure written in Gretl is available from the authors upon request.

As a complacement of standard linear Granger causality tests we applied impulse response (IR) analysis additionally. The standard Granger causality analysis provides an opportunity to the establishment of direction of causal link between variables, however it does not tell anything about signs of this relationship. In order to examine the reaction of effect variable to the shock in the cause variable (which is transmitted through the dynamic structure of VAR model) we applied impulse response function based on orthogonal residuals (established through the application of Cholesky decomposition). In order to save the space we do not present all technical details (like definition and properties of Wold decomposition etc.) and results of suitable preliminary analysis (like analysis of Wold instantaneous causality etc.) which should be performed before applying orthogonal IR functions. The reader may find the theoretical background of this method in Lütkepohl (1993) and Hamilton (1994). Furthermore, complete results of all mentioned preliminary tests conducted before application of the IR techniques are available from authors upon request.

Besides the bootstrap–based linear causality test and IR analysis the nonlinear test for Granger causality was also used in this paper. There are two main facts justifying this decision. Firstly, standard linear Granger causality tests tend to have extremely low power in detecting certain kinds of nonlinear relationships (see e.g. Brock (1991), Gurgul [a]nd Lach (2009)). Secondly, since the traditional linear approach is based on testing the statistical significance of suitable parameters only in mean equation the causality in higher–order structure (for example causality in variance etc.) cannot be explored (Diks [a]nd DeGoede (2001)). The application of nonlinear approach may be a solution to this problem as it allows exploring complex dynamic links between variables of interest.

The idea of nonlinear procedure comes from Baek and Brock (1992). Their findings were thereafter modified by Hiemstra and Jones (1994). Diks and

Panchenko (2005) found the testing procedure proposed by Hiemstra and Jones (HJ test) mostly improper for testing for Granger causality. They managed to prove that the hypothesis examined by Hiemstra and Jones is in general not equivalent to the null hypothesis of Granger non–causality. Furthermore, their research led to the establishment of exact conditions under which the HJ test is a useful tool for causality analysis. They managed to bypass the above mentioned problem of testing for an incorrect hypothesis and provided detailed description of the asymptotic theory of their modified test statistic.

In this article we use nonlinear causality test proposed by Diks and Panchenko (2006). In our research we decided to use some typical values of bandwidth parameter, setting it at the level of 0.5, 1 and 1.5 for all conducted tests. These values were commonly used in previous papers (see e.g. Hiemstra and Jones (1994), Diks and Panchenko (2005) and (2006)). We have also decided to use the same lags for every pair of time series being analyzed establishing this lag at the order of 1 and 2. More details about meaning of technical parameters and the form of applied test statistic may be found in Diks and Panchenko (2006).

We performed our calculations on the basis of residual time series resulting from the appropriate augmented VAR model. Since the structure of linear dependences had been filtered out with application of suitable VAR models and TY procedure, residual time series reflect strict nonlinear dependencies (see e.g. Baek and Brock (1992), Chen and Lin (2004), Ciarreta and Zarraga (2007)). The time series of residuals were standardized, thus they shared a common scale parameter. Finally we must note that we used one–side test rejecting whenever calculated test statistic was significantly large. There are at least two main reasons justifying this choice. Firstly, in practice one–sided test is often found to have larger power than a two–sided one (see e.g. Skaug and Tjøstheim (1993)). Secondly, although significant negative values of test statistic also provide basis for rejection of the null hypothesis of Granger non–causality, they additionally indicate that the knowledge of past values of one time series may aggravate the prediction of another one. In contrast, the causality analysis is usually conducted to judge whether this knowledge is helpful (not aggravating) for prediction issues or not.

Finally, we shall note that the former research provided solid basis for claiming that the nonlinear causality test tends to over–reject in cases of presence of heteroscedastic structures in analyzed time series. (see e.g. Diks and Panchenko (2006)). Thus, we have decided to test all residual time series additionally for the presence of GARCH structures. Since we found significant proof of the presence of conditional heteroscedasticity in residuals of most VAR models, we decided to re–run nonlinear causality test for filtered series of residuals. Complete results of heteroscedasticity tests are available from authors upon request. At this point we shall also note that GARCH filtering shall be carried out carefully as it sometimes may lead to loss of power of the test, which derives from possible misspecification of conditional heteroscedasticity model. This of course may simply lead to spurious results of the test (Diks and Panchenko (2006)).

6. Analysis of empirical results

In this section the results of short–run linear and nonlinear Granger causality tests as well as the impulse response analysis are presented. These findings may be helpful in describing the structure of dynamic links between real GDP growth and crucial budgetary expenditure categories in Poland in the period under study. One may expect these outcomes to provide basis for judging which of two main concepts described in previous sections, namely Wagner's Law or Keynesian's theory, seems to be more adequate for Polish economy. We shall start the presentation of results of our research with the outcomes obtained from analysis of linear Granger causality. Tables 8–12 contain *p*–values obtained while testing for linear Granger causality through the application of bootstrap–based Toda–Yamamoto procedure. Numbers in brackets denote corresponding *p*–values obtained with the help of standard (chi–square) distribution of modified Wald test statistic. The value of *N* parameter denotes number of bootstrap replications used to construct the distribution of TY test statistic. For every pair of variables we first established the number of lags (parameter *p*) in non–augmented two–dimensional VAR model. For this purpose we followed a simple procedure. We set up maximal possible lag length at the level of 6 and then we used several information criteria (namely, AIC, BIC, HQ and SIC) to choose the optimal lag length. For all VAR models the optimal lag was always one of the elements of the following set {1, 4, 5}. If there were several possibilities indicated by information criteria for one specific model then we analyzed model residuals (in each variant) and rejected the value of that lag parameter for which the significant autocorrelation of error vector was reported. If all possibilities were rejected then we set up the lag parameter at the level of 4. This value was established arbitrarily and seemed to be a proper choice for quarterly data. This procedure (arbitrary establishment of lag parameter) is an alternative method to application of popular model selection criteria and it was commonly used in previous papers (e.g. see Granger (2000)). Significant autocorrelation may prove that the established lag length was too small and some important lagged parameters have been omitted in construction of VAR model. This may in turn affect both the causality tests (linear and nonlinear) as well as the IR analysis. One may expect that autocorrelation of error term should not be a serious problem for the application of bootstrap methods. However, in practical research it may significantly worsen performance of this approach. We shall note once again that since the GDP time series was found to be I(1), parameter *d* was set up to one in case of all examined pairs of variables. Whenever test results indicated the existence of causal link in certain direction (at 10% significance level) the shading was used to mark this finding.

The following table contains results computed by VAR model constructed for GDP and BUDGET time series:

Table 8. Results of Toda–Yamamoto test for linear Granger causality between GDP and BUDGET (set of lag lengths indicated by information criteria: {1, 5}, final lag length: p=5)

Null hypothesis	p–value		
	N=100	N=500	N=1000
GDP does not Granger cause BUDGET	0.58 (0.63)	0.63 (0.63)	0.68 (0.63)
BUDGET does not Granger cause GDP	**0.09 (0.06)**	**0.07 (0.06)**	**0.09 (0.06)**

As we can see test results strongly support hypothesis that BUDGET Granger causes GDP (at 10% significance level). Furthermore, test results provided no basis to claim that linear Granger causality runs in the opposite direction. It should be also noted that both these findings were reported by results of asymptotic– and bootstrap–based TY procedure (despite value of parameter N). The next table contains results computed by VAR model constructed for GDP and HR time series:

Table 9. Results of Toda–Yamamoto test for linear Granger causality between GDP and HR (set of lag lengths indicated by information criteria: {1, 5}, final lag length: p=5)

Null hypothesis	p–value		
	N=100	N=500	N=1000
GDP does not Granger cause HR	0.64 (0.52)	0.67 (0.52)	0.54 (0.52)
HR does not Granger cause GDP	0.32 (0.21)	0.41 (0.21)	0.39 (0.21)

After analyzing outcomes presented in table 9 one can easily see that the results of Toda–Yamamoto test provided no basis to claim that linear Granger causality runs in any direction for real growth rate of human resources expenditure and GDP growth variables. We shall underline that this finding was reported for both types of test statistic distribution, namely the $\chi^2(5)$ distribution and bootstrap–based distribution. It is worth mentioning that this phenomenon was reported for all used numbers of bootstrap replications.

The following table contains results gained after analysis of VAR model constructed for GDP and PR time series:

Table 10. Results of Toda–Yamamoto test for linear Granger causality between GDP and PR (set of lag lengths indicated by information criteria: {1, 4}, final lag length: p=4)

Null hypothesis	p–value		
	N=100	N=500	N=1000
GDP does not Granger cause PR	0.71 (0.65)	0.63 (0.65)	0.67 (0.65)
PR does not Granger cause GDP	0.38 (0.41)	0.44 (0.41)	0.39 (0.41)

Similarly to previous case, also for these two variables both variants of Toda–Yamamoto procedure indicated that there is no linear Granger causality running in any direction. Therefore, neither Keynesian approach nor Wagner's Law was found to be a proper pattern for dynamic relationship between GDP and PR. It is worth mentioning that this finding was obtained regardless of the number of bootstrap replications used.

The following table contains results gained after analysis of VAR model constructed for GDP and NIP time series:

Table 11. Results of Toda–Yamamoto test for linear Granger causality between GDP and NIP (set of lag lengths indicated by information criteria: {1}, final lag length: p=4)

Null hypothesis	p–value		
	N=100	N=500	N=1000
GDP does not Granger cause NIP	0.72 (0.67)	0.65 (0.67)	0.72 (0.67)
NIP does not Granger cause GDP	**0.08 (0.02)**	**0.04 (0.02)**	**0.06 (0.02)**

The outcomes presented in table 11 provided solid basis for claiming that there is no linear Granger causality running from GDP to NIP. This result was reported in both asymptotic– and bootstrap–based (once again nonetheless value of parameter N) variant of TY procedure. On the other hand, results of linear causality analysis provided relatively convincing arguments for the existence of causal link running from the real growth rate of budgetary expenditure on net interest payment to the real GDP growth rate. All these facts are in line with Keynes's approach to expenditure–GDP relationship.

The last VAR model was constructed for GDP and OTHER variables. The following table contains results of suitable causality analysis:

Table 12. Results of Toda–Yamamoto test for linear Granger causality between GDP and OTHER (set of lag lengths indicated by information criteria: {1}, final lag length: p=4)

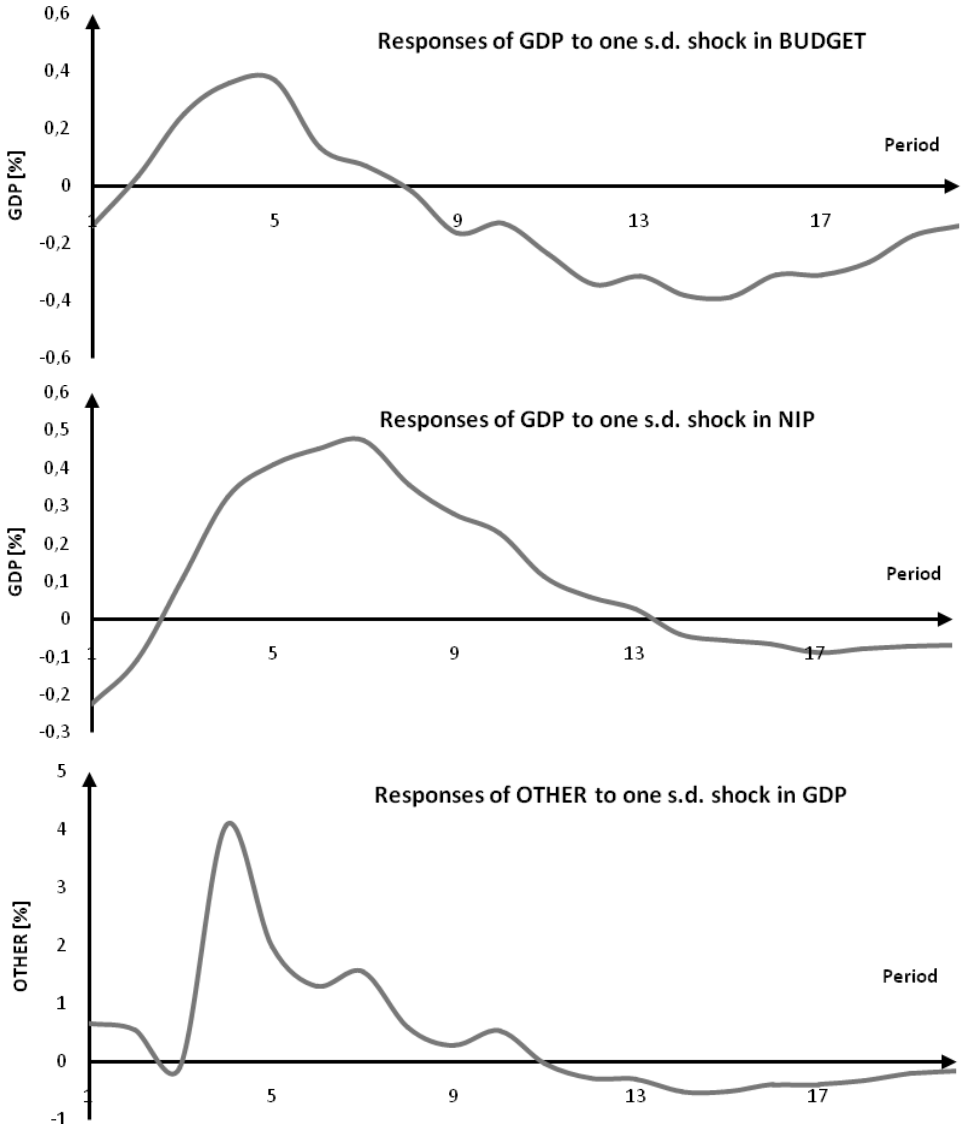
Null hypothesis	p–value		
	N=100	N=500	N=1000
GDP does not Granger cause OTHER	**0.03 (0.01)**	**0.04 (0.01)**	**0.05 (0.01)**
OTHER does not Granger cause GDP	0.61 (0.59)	0.67 (0.59)	0.51 (0.59)

The Toda–Yamamoto procedure based on asymptotic distribution theory provided no support for claiming that OTHER Granger causes GDP. The application of bootstrap–based distribution provided similar results (suitable p–value no less than 0.50). On the other hand, both variants of TY procedure strongly point at the existence of causal link in the direction from real GDP growth rate to growth rate of budgetary expenditure on examined category (OTHER variable). Unlike the previous case, this time the Wagner's Law was found as the suitable explanation of the GDP–expenditure relationship. It is worth mentioning that both these findings were obtained despite the number of bootstrap replications used. This robustness to technical parameters of bootstrap approach makes results of performed causality analysis even more convincing. At this place we shall underline that for other pairs of variables results of asymptotic– and bootstrap–based tests were also relatively similar.

After analyzing outcomes presented in tables 8–12 one can easily see that for the Polish economy the total public expenditure was found as a causal factor for movements of real growth rate of GDP. This is how the principals of Keynesian economy were found as the set of suitable rules for describing the relationship between GDP and total budgetary expenditure in Poland in the period under study. On the other hand, the causality analysis performed for expenditure sub–categories provided relatively mixed results. For HR and PR variables no recognizable pattern was found, while for NIP and OTHER categories results supporting competitive theories were reported.

However, the analysis of linear Granger causality in terms of TY procedure may not provide the complete information about the dynamic interactions between chosen variables. Therefore, the impulse response analysis was performed additionally. Every IR function illustrates the response of one variable (found as a caused variable through application of TY procedure) to one s.d. shock in time series of other variable (found as a causal factor in TY procedure). As we have already mentioned, the complete results of preliminary analysis are available from authors upon request. The following figure contains illustration of all responses:

Figure 1. Impulse responses of caused variables to one s.d. shocks in time series of causal factors



The one s.d. (6.82%) shock from BUDGET causes negative (–0.13%) response of GDP in the first quarter. However, the positive responses were reported in quarters 2 to 7. The highest positive response was reported for the fifth period and reached the value of 0.36%. Starting from eight period negative responses occur. The biggest drop in GDP was found for quarter 14 and reached the value of –0.37%.

The one s.d. (6.79%) shock from NIP causes negative responses of GDP in the first two quarters. However, in quarters 3 to 13 the positive responses were indicated. The highest positive response was reported for seventh period and reached the value of 0.47%. Starting from quarter 14 negative responses occur. However, these drops are relatively slight and do not exceed the value of 0.09%.

The one s.d. (2.09%) shock in GDP time series caused positive responses of OTHER variable in the first 10 quarters (except for slight negative response reported in third quarter). The highest positive response was reported for fourth period and reached the value of 4.07%. Starting from eleventh quarter negative responses had occurred, however they were not as significant as positive ones (drops no greater than 0.50%).

In addition to linear causality tests and impulse response analysis the nonlinear Granger causality tests were conducted as well. Results obtained for unfiltered residual time series are presented in the following table:

Table 13. Results of tests for nonlinear Granger causality between examined variables (unfiltered data)

Null hypothesis	*p*–value			
	ε=0.5	ε=1	ε=1.5	*l*
GDP does not Granger cause BUDGET	0.81	0.81	0.77	1
	0.77	0.71	0.45	2
BUDGET does not Granger cause GDP	0.96	0.91	0.92	1
	0.82	0.85	0.93	2
GDP does not Granger cause HR	0.40	0.38	0.31	1
	0.72	0.92	0.54	2
HR does not Granger cause GDP	0.23	0.21	0.17	1
	0.18	**0.08**	0.27	2
GDP does not Granger cause PR	0.39	0.27	0.67	1
	0.52	0.61	0.92	2
PR does not Granger cause GDP	0.24	0.33	0.78	1
	0.64	0.42	0.68	2
GDP does not Granger cause NIP	0.90	0.81	0.45	1
	0.86	0.62	0.67	2
NIP does not Granger cause GDP	0.95	0.88	0.85	1
	0.76	0.67	0.92	2
GDP does not Granger cause OTHER	0.82	0.82	0.42	1
	0.71	0.72	0.87	2
OTHER does not Granger cause GDP	0.20	0.27	**0.08**	1
	0.18	0.23	**0.09**	2

As we can see the test results provided solid evidence to claim that in most cases under study nonlinear Granger causality does not run in any direction. Some evidence of existence of causal link was found only for HR and OTHER variables. In both mentioned cases the direction of causal link was in line with fundamentals of Keynes's theory.

Finally, taking into consideration the fact that Diks and Panchenko's test was found to be sensitive to presence of heteroscedasticity in analyzed time series, we had performed GARCH(1,1)–filtration of suitable time series of residuals and then we re–ran nonlinear causality analysis. It is worth underlying that for almost every analyzed case the GARCH(1,1) structure was significantly present in residual time series. In order to save the space we do not present the results of these tests in this paper, however they are of course available from the authors upon request. The nonlinear causality was not found at reasonable significance levels for any analyzed pair of filtered variables. Therefore, for HR and OTHER variables applicability of Keynesian approach seems a bit uncertain. This phenomenon may somehow prove that nonlinear causality analysis is indeed sensitive to the presence of heteroscedastic structures in examined data which is in line with the outcomes presented in previous papers (e.g. Diks and Panchenko (2006)). However, we cannot forget that possible misspecification of heteroscedasticity model could be the reason for relatively different indications of tests conducted for unfiltered and filtered data.

7. Final remarks

Economists around the world make efforts to find sources of economic growth. Technically, it seems that conducting this type of research for developed economy is not a serious problem since necessary data is quite reachable. For countries like Poland, where economy is still in transitory phase, the problem of insufficient datasets occurs, which in turn causes many questions concerning economic growth impossible to answer.

In this paper we reached out to this problem as we decided to use quarterly data. This way the dataset of highest possible frequency was applied. Our goal was to test the applicability of two contrasting theories, namely Wagner's Law and Keynesian's theory, for the Polish economy. The application of TY procedure in both variants (asymptotic– and bootstrap–based) provided solid basis for claiming that for total budgetary expenditure and economic growth hypothesis 1 is false, but hypothesis 2 is true, which means that in this case Keynesian's theory applies. It is also worth mentioning that since results of nonlinear causality analysis provided no evidence of existence of causal link between examined variables in any direction then the relationship between GDP and total budgetary expenditure was found to have a strict linear nature.

Relatively mixed results were obtained for all four analyzed sub–division expenditure. Results of linear causality analysis provided relatively convincing

support for rejection of hypothesis 3. Both the asymptotic– and bootstrap–based (regardless value of parameter *N*) variants of TY procedure show evidence that linear causality does not run in any direction between HR and economic growth. However, nonlinear causality analysis provided weak support claiming that for this pair of variables the Keynesian theory applies. Since the later was reported only for unfiltered data this result may be due to the sensitivity of Diks and Panchenko's test to the presence of heteroscedastic structures which causes over–rejection. Therefore, applicability of Keynesian theory to HR and GDP variables seems quite doubtful and hypothesis 3 should rather be rejected.

We have also found relatively weak evidence in favour of hypothesis 4. Although the conducted tests do not allow to reject hypothesis of nonlinear causality between PR and economic growth, we found significant evidence of the unidirectional linear causal relation existence in the sense of Keynesian's theory for NIP and GDP variables. The application of nonlinear methods (for unfiltered and GARCH(1,1)–filtered data) provided no evidence of existence of causal link for PR and GDP as well as for NIP and GDP variables in any direction.

The application of asymptotic– and bootstrap–based TY tests strongly supports hypothesis 5, i.e. the existence of causal link in the direction from real GDP growth rate to real growth rate of budgetary expenditure on sub–categories included in OTHER variable. This finding supports the view that economic growth is driving public expenditure on science, national defence and public security. These results were reported at 5% significance level. Although for unfiltered data the nonlinear causality analysis provided relatively weak support for claiming that for GDP and OTHER variables Keynesian theory is also applicable, the GARCH(1,1)–filtration of time series led to different conclusion (no causality in any direction). As in GDP–HR case this phenomenon may provide some support for the hypothesis that nonlinear approach is indeed sensitive to presence of heteroscedasticity in analyzed time series.

The results by Impulse Response Function demonstrate sensitivity of economic growth rate to one s.d. shocks imposed on budgetary expenditure on NIP and total budgetary expenditure. The peak of economic growth response is located in 5–th (BUDGET) or in 7–th quarter (NIP). In further quarters the one s.d. shocks implies drop in economic growth rate. This is in favour of hypothesis 6. On the other hand, a one s.d. shock imposed on GDP time series causes significant positive responses of OTHER variable in first quarters. In this case negative responses have also occurred, but they were not as significant as positive ones.

To summarise, in the case of the Polish economy Keynesian's theory is in general more appropriate than Wagner's Law. This statement is mostly based on the results of linear causality analysis, as results of nonlinear tests, although providing some weak support, also seem to suffer due to uncertainty occurring in case of heteroscedasticity problems. Furthermore, one can claim, that rise in NIP growth rate can be linearly transmitted to GDP growth rate. This finding seems to

be an interesting advice for Polish policy makers and should gain a considerable attention.

Acknowledgments

The authors would like to acknowledge the help of the Ministry of Finance of Poland in obtaining the dataset.

REFERENCES

AGIAKOGLU, C. & NEWBOLD, P., 1992, Empirical Evidence on Dickey–Fuller Type Tests, *Journal of Time Series Analysis*, 13, pp. 471–483.

ANWAR, M.S., DAVIES, S. & SAMPATH, R.K., 1996, Causality between Government Expenditures and Economic Growth: An Examination Using Cointegration Techniques, *Public Finance*, 51, pp. 166–184.

ASTERIOU, D. & ARGIOMIRGIANAKIS, G.M., 2001, Human capital and economic growth: time series evidence from Greece, *Journal of Policy Modeling*, 23, pp. 481–489.

BAEK, E. & BROCK, W., 1992, A general test for Granger causality: Bivariate model, *Technical Report*, Iowa State University and University of Wisconsin, Madison.

BENOIT, E., 1978, Growth and defence in developing countries, *Economic Development and Cultural Change*, 26, pp. 261–280.

BISWAS, B. & RAM, R., 1986, Military expenditures and economic growth in less developed countries: an augmented model and further evidence, *Economic Development and Cultural Change*, 34, pp.361–372.

BLOMQVIST, A.G. & CARTER, R.A.L., 1997, Is health care really a luxury?, *Journal of Health Economics*, 16, pp. 2007–2029.

BROCK, W., 1991, Causality, chaos, explanation and prediction in economics and finance, In: Casti J., Karlqvist A. (eds.), *Beyond Belief: Randomness, Prediction and Explanation in Science*, CRC Press, Boca Raton, Fla.

CASELLI, F., 1999, Technological revolutions, *American Economic Review*, 89, pp. 78–102.

CHANG, T., FANG, W., WEN, L.F. & LIU, C., 2001, Defense expenditure, economic growth and temporal causality: evidence from Taiwan and mainland China, 1952–1995, *Applied Economics*, 33, pp. 1289–1299.

CHEN, A.S. & LIN, J.W., 2004, Cointegration and detectable linear and nonlinear causality: analysis using the London Metal Exchange lead contract, *Applied Economics*, Vol. 36, pp. 1157–1167.

CHEN, C.H., 1993, Causality between defence spending and economic growth: the case of mainland China, *Journal of Economic Studies*, 26, pp. 37–43.

CHOU, P.H. & ZHOU, G., 2006, Using Bootstrap to Test Portfolio Efficiency, *Annals of economics and finance*, 2, pp. 217–249.

CIARRETA, A. & ZARRAGA, A., 2007, Electricity consumption and economic growth: evidence from Spain, Working Paper, Dept. of Applied Economics III (Econometrics and Statistics), University of the Basque Country.

CULYER, A.J., 1990, Cost containment in Europe, Health Care Systems in Transition, OECD, Paris.

DAVISON, A.C. & HINKLEY, D.V., 1999, *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

DE MEULEMEESTER, J.–L. & ROCHAT, D., 1995, A causality analysis of the link between higher education and development, *Economics of Education Review*, 14, pp. 351–361.

DEGER, S. & SEN, S., 1983, Military expenditures, spin–off and economic development, *Journal of Development Economics*, 13, pp. 68–83.

DEMIRBAS, S., 1999, Cointegration Analysis–Causality Testing and Wagner's Law: The Case of Turkey, 1950–1990, Available at: http://www.econturk.org/safaaabi.pdf.

DEROUEN, K.R., 1995, Arab–Israeli defence spending and economic growth, *Conflict Management and Peace Science*, 14, pp. 25–47.

DEVLIN, N. & HANSEN, P., 2001, Health care spending and economic output: Granger causality, *Applied Economics Letters*, 8, pp. 561–564.

DIKS, C.G.H. & DEGOEDE, J., 2001, A general nonparametric bootstrap test for Granger causality, In: Broer H.W., Krauskopf W., Vegter G. (eds.), *Global analysis of dynamical systems*, Institute of Physics Publishing, Bristol, United Kingdom.

DIKS, C.G.H. & PANCHENKO, V., 2005, A note on the Hiemstra–Jones test for Granger non–causality, *Studies in Nonlinear Dynamics and Econometrics*, 9, No. 2, Article 4.

DIKS, C.G.H. & PANCHENKO, V., 2006, A new statistic and practical guidelines for nonparametric Granger causality testing, *Journal of Economic Dynamics & Control*, 30, pp. 1647–1669.

DOLADO, J.J. & LÜTKEPOHL, H., 1996, Making Wald tests work for cointegrated VAR systems, *Econometrics Reviews,* 15, pp. 369–386.

DUNNE, J.P., SMITH, R.P. & WILLENBOCKEL, D., 2005, Models of military expenditure and growth: a critical review, *Defence and Peace Economics*, 16, No. 6, pp. 449–461.

EASTERLY, W., KING, R., LEVINE, R. & REBELO, S., 1994, Policy, technology adoption and growth, In: Pasinetti L., Solow R. (Eds), *Economic Growth and the Structure of Long–term Development*, St Martins Press, New York.

FREDERIKSEN, P.C. & LOONEY, R.E., 1982, Defence expenditures and economic growth in development countries: some further empirical evidence, *Journal of Economic Development*, 7, pp. 113–125.

GRANGER, C.W.J., 1969, Investigating causal relations by econometric models and cross–spectral methods, *Econometrica*, 37, pp. 424–438.

GRANGER, C.W.J. & NEWBOLD, P., 1974, Spurious regression in econometrics*, Journal of Econometrics*, 2, pp. 111–120.

GRANGER, C.W.J., HUANG, B. & YANG, C., 2000, A bivariate causality between stock prices and exchange rates: evidence from recent Asian Flu, *The Quarterly Review of Economics and Finance*, 40, pp. 337–354.

GROBAR, L. & PORTER, R., 1989, Benoit revisited: defence spending and economic growth in less developed countries, *Journal of Conflict Resolution*, 33, pp. 318–345.

GURGUL, H. & LACH, Ł., 2009, Linear versus nonlinear causality of DAX companies, *Operations Research and Decisions*, 3, pp. 27–46.

HACKER, R.S. & HATEMI–J, A., 2006, Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application, *Applied Economics*, pp. 1489–1500.

HAMILTON, J.D., 1994, *Time Series Analysis,* Princeton University Press, Princeton.

HANSEN, P. & KING, A., 1996, The determinants of health care expenditure: a cointegration approach, *Journal of Health Economics*, pp. 127–137.

HANSEN, P. & KING, A., 1998, Health care expenditure and GDP: panel data unit root test results–comment, *Journal of Health Economics*, 17, pp. 377–381.

HEO, U. & EGER, R.J., 2005, The security–prosperity dilemma in the United States, *Journal of Conflict Resolution*, 49, pp. 699–708.

HIEMSTRA, C. & JONES, J.D., 1994, Testing for linear and nonlinear Granger causality in the stock price–volume relation, *Journal of Finance*, 49, pp. 1639–1664.

HITRIS, T. & POSNETT, J., 1992, The determinants and effects of health expenditure in developed countries, *Journal of Health Economics*, 11, pp. 173–181.

HOROWITZ, J.L., 1995, *Advances in Economics and Econometrics: Theory and Applications, Chapter 7: Bootstrap methods in econometrics: Theory and numerical performance*, Cambridge University Press, Cambridge.

IN, F. & DOUCOULIAGOS, C., 1997, Human capital formation and US economic growth: a causality analysis, *Applied Economics Letters*, 4, pp. 329–331.

JACKSON, P.M., MERYEM, D.F. & SAMI, F., 1998, Conitegration, Causality, and Wagner's Law: A test for Northern Cyprus, Available at http://www.le.ac.uk/economics/research/RePEc/lec/lpserc99–2.pdf.

JOERDING, W., 1986, Economic growth and defense spending, *Economics*, 21, pp. 35–40.

KALYONCU, H. & YUCEL, F., 2006, An analytical approach on defense and expenditure and economic growth. The case of Turkey and Greece, *Journal of Economic Studies*, 33, pp. 336–343.

KHAN, A.H., 1990, Wagner's Law and the Developing Economy: A Time Series Evidence from Pakistan, *Indian Economic Journal*, 38, pp. 115–123.

KOLLIAS, C., 1994, The economic effects of defense spending in Greece 1963–90: some preliminary econometric findings, *Spoudai*, 44, pp. 114–130.

LACH, Ł., 2010, Application of bootstrap methods in investigation of size of the Granger causality test for integrated VAR systems, to appear in *Managing Global Transition.*

LAI, C–N., HUANG, B–N. & YANG, C–W., 2005, Defense spending and economic growth the Taiwan straits: a threshold regression model, *Defence and Peace Economics,* 16, pp. 45–57.

LANDAU, D., 1996, Is one of the "peace dividends" negative? Military expenditure and economic growth in the wealthy OECD countries, *The Quarterly Review of Economic and Finance*, 36, pp. 183–195.

LIM, D., 1983, Another look at growth and defense in developed countries, *Economic Development and Cultural Change*, 31, pp. 377–384.

LIU, L.C., HSU, C.E. & YOUNIS, M.Z., 2008, The association between government expenditure and economic growth: Granger causality test of US data, 1947–2002, *Journal of Public Budgeting & Financial Management*, 20, pp. 537–553.

LÜTKEPOHL, H., 1993, *Introduction to Multiple Time Series Analysis,* 2nd ed., Springer–Verlag, New York.

MANTALOS, P., 2000, A Graphical Investigation of the Size and Power of the Granger–Causality Tests in Integrated–Cointegrated VAR Systems, *Studies in Nonlinear Dynamics & Econometrics*, 4, Issue 1, Article no. 2.

MCCOSKEY, S.K. & SELDON, T.M., 1998, Health care expenditures and GDP: panel data unit root test results, *Journal of Political Economy*, 70, pp. 129–157.

NAGARAJAN, P. & SPEARS, A., 1990, An Econometric Test of Wagner's Law for Mexico: A Re–Examination, *Public Finance*, 45, pp. 165–168.

NARAYAN, P.K. & SINGH, B., 2007, Modelling the relationship between defence spending and economic growth for the Fiji islands, *Defence and Peace Economics*, 18, pp. 391–401.

NARAYAN, P.K. & SMYTH, R., 2006, Higher Education, Real Income and Real Investment in China: Evidence from Granger Causality, *Education Economics*, 14, pp. 107–125.

PARK, W.K., 1996, Wagner's Law vs. Keynesian Paradigm: The Korean Experience, *Public Finance*, 51, pp.71–91.

PAULSEN, J., 1984, Order determination of multivariate autoregressive time series with unit roots, *Journal of time series analysis*, 5, pp. 115–127.

PHILLIPS, P.C.B., 1986, Understanding the spurious regression in econometrics, *Journal of Econometrics*, 33, pp. 311–340.

PHILLIPS, P.C.B. & PERRON, P., 1988, Testing for a Unit Root in Time Series Regressions, *Biometrika,* 75, pp. 335–346.

PLUTA, J.E., 1979, Wagner's Law, Public Sector Patterns, and Growth of Public Enterprises in Taiwan, *Public Finance Quarterly*, 7, pp. 25–46.

REITSCHULER, G. & LOENING, J.L., 2005, Modeling the defense–growth nexus, *World Development*, 33, pp. 513–526.

ROBERTS, J., 1999, Sensitivity of elasticity estimates for OECD health care spending analysis of a dynamic heterogenous data field, *Health Economics*, 8, pp. 459–472.

SAUNDERS, P., 1985, Public Expenditure and Economic Performance in OECD Countries, *Journal of Public Policy*, 5, pp. 1–21.

SINHA, D., 1998, Government Expenditure and Economic Growth in Malaysia, *Journal of Economic Development*, 23, pp. 265–283.

SKAUG, H.J. & TJØSTHEIM, D., 1993, Nonparametric tests of serial independence, In: Subba Rao T. (ed.), *Developments in time series analysis*, Chapman and Hall, London.

SMITH, R., 1980, ME and investment in OECD countries, 1954–1973, *Journal of Comparative Economics*, 4, pp. 19–32.

TODA, H.Y. & YAMAMOTO, T., 1995, Statistical inference in vector autoregressions with possibly integrated processes*, Journal of Econometrics*, 66, pp. 225–250.

WOLDE–RUFAEL, Y., 2006, Electricity consumption and economic growth: a time series experience for 17 African countries*, Energy Policy,* 34, pp. 1106–1114.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 361—379

MULTIDIMENSIONAL APPROACH TO POVERTY MEASUREMENT: FUZZY MEASURES OF THE INCIDENCE AND THE DEPTH OF POVERTY*

Tomasz Panek[1]

ABSTRACT

The paper presents a methodology for study of multi-dimensional aspects of poverty. In addition to the traditional uni-dimensional measures of poverty, exclusively estimated on the basis of the monetary variable (income or expenditure), other non-monetary variables are incorporated in analysis of the poverty phenomenon. The multidimensional approach has been based on the fuzzy sets theory in which the conventional poor/non-poor dichotomy is replaced by assessment of the degree of household poverty threat. The same methodology facilitates comprehensive monetary and non-monetary poverty analysis. In order to provide effective assessment of poverty the fuzzy measures of poverty were applied. The study employs the fuzzy measures to compare the degree of monetary poverty and deprivation threat in Polish voivodships in 2007 using micro data from the EU-SILC survey. The results of estimation of the fuzzy measures show that poverty in Poland has many dimensions and that its measurement solely from income standing vantage point is highly insufficient.

Key words: multi-dimensional poverty, fuzzy poverty measures, aggregation of indicators.

1. Introduction

The traditional approach to measuring poverty based on monetary indicators, whose foundations were set forth by the Material Welfare School (Marshall, 1920), dominated in nearly all research into this phenomenon up to the 1970s. In this approach the evaluation of the level of needs satisfaction was conducted exclusively on monetary variables (income or expenditure). Nevertheless, the

* This paper has been carried out with funding by the European Union within the project *Small Area Methods* for *Poverty and Living Conditions Estimates* (Proj. No. 217565-FP7-SSH-2007-1). All calculations have been done by Mariusz Dzieciątko (Warsaw School of Economics) and Robert Wieczorkowski (Central Statistics Office).

[1] Institute of Statistics and Demography, Warsaw School of Economics; Al. Niepodległości 162, 02-554 Warsaw, Poland; tompa@sgh.waw.pl.

viewpoint that the identification of impoverished persons exclusively on the basis of pecuniary categories is greatly insufficient gradually began to meet with considerable criticism. Nowadays, there is a widespread agreement that poverty is a multidimensional phenomenon and cannot be reduced solely to monetary dimension but has to be also explained by diverse non-monetary variables.

Many researchers have postulated the necessity of treating poverty multidimensionally. Townsend was one of the first persons to single out the imperfection inherent in identifying poverty exclusively on the basis of the income criterion. He proposed for poverty analyses to incorporate dwelling conditions, affluence, education as well as professional and financial resources (Abel-Smith and Townsend, 1973; Townsend, 1979). A broader look at the problem of poverty than just through the prism of income (expenditures) was also presented, among others, by Atkinson and Bourguignon (1982), Hagenaars (1986), Sen (1992), Panek (1996), Bourguignon and Chakravarty (2003), Tsui (2002), and Deuch and Silber (2005).

In this paper the multidimensional approach to poverty measurement based on application of the fuzzy set theory is presented. Moreover, the incidence and the depth of monetary and non-monetary poverty (deprivation) in Poland in 2007 by voivodships is analysed.

2. Characterization of the data

The bases for the analyses conducted are the data from the European Union Survey on Income and Living Conditions (EU-SILC) carried out by the Central Statistical Office (CSO, 2009). The main objective of EU-SILC is to supply EU comparable data on the living conditions of the population of the EU Members States. It is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal microdata on income, poverty, social exclusion and living conditions. In order to satisfy these needs EU- SILC is carried out with the use of the panel method in the four-year cycle. The households were selected for EU-SILC survey in Poland by means of two-stage stratified sampling scheme with different selection probabilities at the first stage. Before selection sampling units were stratified by voivodship, and then, within voivodships, according to the class of locality (rural areas and urban areas grouped by size of town). The first-stage sampling units (primary sampling units-PSU's) were census areas. During the second stage, dwellings were drawn systematically from a randomly generated list of dwellings, independently within each stratum created during the first stage. All the households, including all household members, from the selected dwellings entered the sample.

In the empirical results presented in the article, the data for Poland for 2007 were applied. In 2007, 14286 households were examined and individually 34898 members of these households age 16 or more. In order to generalize the results

over the entire population the appropriate system of weights was applied, which takes into consideration:

- selection probability for dwellings and households,
- survey completeness according to the class of locality,
- consistency of the composition of the household and individual populations according to age and gender, with the external demographic data available.

Estimation for standard errors was based on resampling approach. It used a bootstrap method with resamples 500 times from each stratum n_h-1 PSU's with replacement (McCarthy and Snowden, 1985), where n_h denotes the sample of PSU in the n-th strata.

3. Assumptions of the poverty analysis

In order to provide effective assessment of poverty the fuzzy incidence of relative poverty (monetary poverty), the fuzzy incidence of relative deprivation (non-monetary poverty) indicators, the fuzzy depth of relative poverty and the fuzzy depth of relative deprivation indicators were applied. The fuzzy incidence indicators are counterparts of the headcount ratio in the traditional approach defined as:

$$H = \frac{q}{n}, \tag{1}$$

where: q – the number of poor.

The fuzzy depth indicators are counterparts of the income gap index in conventional approach which is defined as follows:

$$I = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{z - y_i}{z}\right), \tag{2}$$

where:

y_i – equivalized income of the i-th individual,

z – poverty line,

n – the number of individuals (households or persons).

The construction of fuzzy measures bases on the fuzzy set approach introduced by Cerioli and Zani (1990) who drew inspiration from the theory of Fuzzy Sets initiated by Zadeh (1965). Cerioli and Zani's proposal was developed by Cheli and Lemmi (1995), and further has been followed by a number of applications (*Fuzzy Set Approach...*, 2006).

The fuzzy set approach makes it possible to avoid simple dichotomization of the population under research into poor and non-poor defined in the relation to some chosen treshold border (poverty line). Poverty is not defined in terms of presence or absence in the subset of poor individuals but as a matter of degree of

belonging to this sub-set. The *i*-th individuals' propensities to poverty is measured by the so-called *membership function* (*m.f.*) to the poverty sphere (μ_i). It assumes value 1 when an individual belongs to poverty set completely, value 0 when an individual does not belong to poverty set, and values between 0 and 1 when it belongs to poverty set partially. Its degree of membership of poverty set increases in proportion to the proximity of *m.f.* to 1.

4. Fuzzy Incidence Indicators

4.1. Fuzzy Monetary Incidence (FMI) Indicator

Cerioli and Zani defined the *m.f.* for the monetary dimension of poverty introducing two threshold values: an individual is definitely poor when he or she is below the first one and when he or she is above the second one . In the latter case, after crossing the threshold, the *m.f.* declines from 1 to 0 linearly.

In order to avoid arbitrary selection of threshold values Chelli and Lemmi (1995) defined the *m.f.* as the distribution function $F(y_i)$ of income, normalised (linearly transformed) so as to equal 1 for the poorest and 0 for the richest individual in the population. Betti and Verma (1999) modified the *m.f.* version proposed by Chelli and Lemmi taking the *m.f.* as the normalised (linearly transformed) Lorenz curve (function) of income $L(F(y_i))$. It varies from 1 for the poorest to 0 for the richest individual. Finally, Betti, Chelli, Lemmi and Verma (2006) combined their previous proposals into the *Integrated Fuzzy and Relative* (IFR) approach.

The Fuzzy Monetary Incidence (FMI) indicator defined under the IFR approach (serving simultaneously as the definition of *m.f.*), combines the (1- $F_{(MI)}$) indicator proposed by Chelli and Lemmi and the (1-$L_{(MI)}$) indicator proposed by Betti and Verma. The (1-$F_{(MI)}$) indicator is the proportion of individuals who are less poor then the individual concerned (their degree of poverty is less marked than the individual concerned):

$$\mu_i(y) = FMI_i = \left(1 - F_{(MI),i}\right)^{\alpha} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma}{\sum_{\gamma=2}^{n} w_\gamma} \right)^{\alpha}, \; i=1,2,...,n; \; \mu_n(y)=0, \qquad (3)$$

where:

y_i – equivalised income of the *i*-th individual,

$F_{(MI),i}$ – value of the income distribution function $F(y_i)$ for the *i*-th individual,

w_γ – weight of the *i*-th individual of rank γ in the ascending income distribution,

α – parameter.

The (1-$L_{(MI)}$) indicator is the share of total equivalised income received by all individuals who are not as poor as the person concerned:

$$\mu_i(y) = FMI_i = \left(1 - L_{(MI),i}\right)^{\alpha} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma y_\gamma}{\sum_{\gamma=2}^{n} w_\gamma y_\gamma} \right)^{\alpha}, \; i=1,2,...,n; \; \mu_n(y)=0, \qquad (4)$$

where:

$L_{(MI),i}$ – the value of the Lorenz curve of income $L(F(y_i)$ for the i-th individual.

Finally, the FMI indicator (the *m.f.*) is defined as a combination of the previous forms (3) and (4):

$$\mu_i(y) = FMI_i = \left(1 - F_{(MI),i}\right)^{\alpha-1}\left(1 - L_{(MI),i}\right), \; i=1,2,...,n. \qquad (5)$$

The parameter α is estimated so that the mean of the FMI indicator (the mean of *m.f.*) is equal to the head count ratio (1) computed for the official poverty line.

The FMI indicator (the incidence of relative poverty indicator) for the population takes the following form:

$$FMI = \frac{\sum_{i=1}^{n} FMI_i \cdot w_i}{\sum w_i}. \qquad (6)$$

5. Fuzzy Supplementary Incidence Indicator (FSI)

In addition to the monetary (income) variable, poverty in the multidimensional approach is also explained by non-monetary variables describing the standard of living of households and individuals. The starting point for including non-monetary variables in poverty analysis is the selection of variables that may be treated as deprivation symptoms (z_j; j=1,2,...,k_h) (to be included in the index or indices of deprivation dimensions) and grouping them into deprivation dimensions (h; h=1,2,...,m). Deprivation symptoms (variables) may take the form of dichotomous[1] or polychotomous variables[2]. The next step is to assign numerical values to each deprivation symptom ordered categories. Then, it is necessary to weight the deprivation symptoms scores in order to construct composite indicators and to scale the measures.

[1] The absence of certain goods or facilities due to financial reasons, for example a car or warm running water.

[2] For example, home mortgage loan defaults (from the absence of default, to default by one month to default of more than six months).

In the calculation of the FSI indicator procedure, numerical values (ranks) are assigned to each deprivation symptom category (c=1,2,...,u) after arranging the deprivation symptom categories from the most deprived (c=1) to the least deprived (c=u) situation[1]. Then, for each deprivation symptom we determine a quantitative deprivation score (assessment of the degree of deprivation) using the following formula:

$$e_{hj,i} = \frac{1 - F\left(c_{hj,i}\right)}{1 - F(1)}, \quad h=1,2,...,m; \ j=1,2,...,k_h; \ i=1,2,...,n, \tag{7}$$

where:

$c_{hj,i}$ – value of the category of the j-th deprivation symptom in the h-th dimension for the i-th individual,

$F(c_{hj,i})$ – value of the j-th deprivation symptom distribution function in the h-th dimension for the i-th individual.

The formula (7) is identical for dichotomous and polychotomous variables (deprivation symptoms).

Using a system of weights, an overall non-deprivation score indicating lack of deprivation, for the i-th individual and for each deprivation dimension separately, is determined:

$$e_{h,i} = \frac{\sum_{j=1}^{k_h} w_{hj}\left(1 - e_{hj,i}\right)}{\sum_{j=1}^{k_h} w_{hj}}, \quad h=1,2,...,m; \ i=1,2,...,m, \tag{8}$$

where:

w_{hj} – weight of the j-th deprivation symptom in the h-th dimension.

Weights are calculated separately within each dimension. The weighting procedure is based on a statistical consideration taking into account the dispersion of deprivation symptoms and its correlation with other deprivation symptoms in the given dimension (see also: Gianni and Verma, 2008). The weights to be given to deprivation indicators (scores) are calculated according to the following formula:

$$w_{hj} = w_{hj}^{a} \cdot w_{hj}^{b}, \quad h=1,2,...,m; \ j=1,2,...,k_h, \tag{9}$$

where:

w_{hj}^{a} – measure of dispersion of the j-th deprivation symptom in the h-th dimension,

[1] For example, for the absence of a car c=1 and for the possession of a car c=2, home mortgage loan defaults from more than sixth months (c=1) to the absence of default (c=u).

w_{hj}^{b} – measure of correlation of the *j*-th deprivation symptom in the *h*-th dimension with other deprivation symptoms in this dimension.

A measure of dispersion is defined as follows:

$$w_{hj}^{a} = V^{(k)}\left(e_{hj}\right) = \frac{S\left(e_{hj}\right)}{\overline{e}_{hj}},\ j=1,2,...,\ k_h;\ h=1,2,...,m, \tag{10}$$

where:

$S\left(e_{hj}\right)$ – standard deviation of deprivation score for the *j*-th deprivation symptom in the *h*-th dimension,

$\overline{e}_{hj}$ – arithmetic mean of deprivation score for the *j*-th deprivation symptom in the *h*-th dimension.

A measure of correlation is defined according to the following form:

$$w_{hj}^{b} = \left(\frac{1}{1+\sum_{j=1}^{k_h}\left|r_{e_{hj,hj'}}\right|\left|r_{e_{hj,hj'}} < r_{e_{hj}}^{*}\right.}\right)\left(\frac{1}{\sum_{j=1}^{k_h}\left|r_{e_{hj,hj'}}\right|\left|r_{e_{hj,hj'}} \geq r_{e_{hj}}^{*}\right.}\right),\ j,j'=1,2,...,k_h, \tag{11}$$

where:

$r_{e_{hj,hj'}}$ – correlation coefficient between deprivation indicators corresponding to the *j*-th and the *j'*-th deprivation symptoms in the *h*-dimension,

$r_{e_{hj}}^{*}$ – critical value (threshold) of the correlation coefficient of the *j*-th deprivation symptom in the *h*-th dimension.

The critical value of the correlation coefficient $r_{e_{hj}}^{*}$ is calculated on the basis of the set of absolute values of correlation coefficients between deprivation indicators (corresponding to the *j*-th and *j'*-th deprivation symptoms in the *h*-dimension) arranged in ascending order extended by including 0 and 1:

$$A = \{0, \left|r_{e_{hj,hj'}}^{(1)}\right|, ..., \left|r_{e_{hj,hj'}}^{(k_h-1)}\right|, 1\}.\ j,j'=1,2,...,k_h;\ j'\neq j. \tag{12}$$

In the next step an overall non-deprivation indicator (non-deprivation score) for any individual is calculated. Since we assume that all deprivation dimensions are of identical importance, they should be equally weighted using the following formula:

$$e_i = \frac{\sum_{h=1}^{m} e_{h,i}}{m},\ i=1,2,...,n. \tag{13}$$

As in the calculation of the FMI indicator, the overall FSI indicator (the *m.f.*), for the *i*-th individual, is defined as a combination of the (1-$F_{(SI),i}$) indicator and the (1-$L_{(SI),i}$) indicator:

$$\mu_i(z) = FSI_i = \left(1 - F_{(SI),i}\right)^{\alpha'-1}\left(1 - L_{(SI),i}\right),\ i=1,2,...,n. \tag{14}$$

The $\left(1 - F_{(SI),i}\right)$ indicator, for the *i*-th individual, is the proportion of individuals who are less deprived than the individual concerned (their degree of deprivation is lower than for the individual concerned):

$$\mu_i(z) = FSI_i = \left(1 - F_{(SI),i}\right)^{\alpha'} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma}{\sum_{\gamma=2}^{n} w_\gamma}\right)^{\alpha'},\ i=1,2,...,n;\ \mu_n(z)=0, \tag{15}$$

where:

$F_{(SI),i}$ – value of the non-deprivation score distribution function $F(e_i)$ for the *i*-th individual,

w_γ – weight of the *i*-th individual of rank γ in ascending non-deprivation score distribution,

α' – parameter.

The (1-$L_{(SI),i}$) indicator is the share of the total non-deprivation score assigned to all individuals less deprived than the person concerned (their deprivation risk is lower than for the individual concerned):

$$\mu_i(z) = FSI_i = \left(1 - L_{(SI),i}\right)^{\alpha'} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma e_\gamma}{\sum_{\gamma=2}^{n} w_\gamma e_\gamma}\right)^{\alpha'},\ i=1,2,...,n;\ \mu_n(z)=0, \tag{16}$$

where:

$L_{(SI),i}$ – value of the Lorenz curve of the non-deprivation score $L\left(F\left(e_i\right)\right)$ for the *i*-th individual.

The overall (for the population in question) Fuzzy Supplementary Incidence indicator (the incidence of relative deprivation indicator) is defined as:

$$FSI = \frac{\sum_{i=1}^{n} FSI_i w_i}{\sum_{i=1}^{n} w_i}. \tag{17}$$

As for the FMI indicator, the parameter α' in equation (14) is determined so as to make the overall FSI indicator equal to the head count ratio (1). The parameter α' is also used to calculate the FSI indicator for every single dimension.

The FSI indicator, in the *h*-th dimension, for the *i*-th individual is defined as:

$$\mu_i(z) = FSI_{hi} = \left(1 - F_{(SI),hi}\right)^{\alpha'-1}\left(1 - L_{(SI),hi}\right),\ h=1,2,...,m;\ i=1,2,...,n. \qquad (18)$$

Finally, the overall Fuzzy Supplementary Incidence indicator (the incidence of relative deprivation indicator) for the *h*-th dimension for the population is calculated as the following mean:

$$FSI_h = \frac{\sum_{i=1}^{n} FSI_{h,i} \cdot w_i}{\sum_{i=1}^{n} w_i};\ h=1,2,...,m. \qquad (19)$$

Fuzzy Depth Indicators

Fuzzy incidence indicators defined under the FR approach overlook the second basic aspect of poverty analysis, namely poverty depth. The necessity of also taking poverty depth into consideration in multidimensional analyses of poverty has been postulated by many researchers (see, for example Shorrocks and Subramanian, 1994). Panek (2009) proposed to extend the IFR approach by incorporating two additional indicators, namely the Fuzzy Monetary Depth indicator (FMD) and the Fuzzy Monetary Supplementary Depth indicator (FSD).

6. Fuzzy Monetary Depth Indicator (FMD)

The starting point for defining the FMD indicator is the calculation of the income (poverty) gap ratio for each individual:

$$v_i = \frac{z - y_i}{z},\ i=1,2,...,n, \qquad (20)$$

with the non-poor individuals v_i being assigned the value of zero.

In the next step, we define the degree of the lack of poverty gap (non-poverty gap score) for each individual:

$$d_i = 1 - v_i,\ i=1,2,...,n. \qquad (21)$$

d_i is a positive score indicating a lack of poverty gap and is analogous to y_i in the construction of the FMI indicator.

The FMD indicator is defined, similarly to the FMI indicator, as a combination of the $(1\text{-}F_{(MD)})$ indicator and the $(1\text{-}L_{(MD)})$ indicator.

The $(1\text{-}F_{(MD),i})$ indicator for the i-th individual is the proportion of individuals whose non-poverty gap score is higher (who are not as poor or better off) than the individual concerned:

$$\mu_i(d) = FMD_i = \left(1 - F_{(MD),i}\right)^{\beta} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma}{\sum_{\gamma=2}^{n} w_\gamma} \right)^{\beta}, \quad i=1,2,...,n; \ \mu_n(d)=0, \qquad (22)$$

where:

$F_{(MD),i}$ – value of the distribution function $F(d_i)$ of the non-poverty gap score for the i-th individual,

w_γ – weight of the i-th individual of rank γ in ascending non-poverty gap score distribution,

β – parameter.

The $(1\text{-}L_{(MD),i})$ indicator is the share of the total non-poverty gap score assigned to all individuals whose non-poverty gap score is higher (who are not as poor or are better off) than the individual concerned:

$$\mu_i(d) = FMD_i = \left(1 - L_{(MD),i}\right)^{\beta} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma d_\gamma}{\sum_{\gamma=2}^{n} w_\gamma d_\gamma} \right)^{\beta}, \quad i=1,2,...,n; \ \mu_n(d)=0, \quad (23)$$

where: $L_{(MD),i}$ – value of the Lorenz curve of the non-poverty gap score $L\left(F\left(d_i\right)\right)$ for the i-th person.

Finally, the degree of poverty gap, for the i-th individual, is defined as a combination of formulas (22) and (23):

$$\mu_i(d) = FMD_i = \left(1 - F_{(MD),i}\right)^{\beta-1}\left(1 - L_{(MD),i}\right), \ i=1,2,...,n. \qquad (24)$$

The overall (for the population in question) Fuzzy Monetary Depth indicator (the depth of relative deprivation indicator) is calculated as follows:

$$FMD = \frac{\sum_{i=1}^{n} \mu_i(d) \cdot w_i}{\sum_{i=1}^{n} w_i}. \qquad (25)$$

The parameter β in equation (24) is estimated so that the mean of the FMD indicator (for the entire population) is equal to the poverty gap index (2).

7. Fuzzy Supplementary Depth indicator (FSD)

The starting point for calculating the FSD indicator is the same set of deprivation symptoms as was established for the FSI indicator. Then, we determine a quantitative deprivation gap ratio for each deprivation symptom, and for each individual:

$$x_{hj,i} = \frac{\left(c_{hj} = r-1\right) - \left(e_{hj,i} - 1\right)}{\left(c_{hj} = r-1\right)}, \quad h=1,2,...,m; j=1,2,...,k_n; i=1,2,...,n, \qquad (26)$$

with the non-deprived individuals, with regard to the j-th symptom in the h-dimension, $x_{hj,i}$ being set to zero (for individual, for which rank assumes value $c_{hj,i} \geq r$; $c_{nj}=1,2,...,u$; r≤u), where:

$c_{hj}=r$ – rank of the j-th deprivation symptom category in the h-th dimension for which deprivation is not found.

The above formula is identical for dichotomous and polychotomous variables (deprivation symptoms).

In the next step the degree of the lack of deprivation gap (non-deprivation gap score) for each individual is calculated:

$$s_{hj,i} = 1 - x_{hj,i}, \; h=1,2,...,m; j=1,2,...,k_n; i=1,2,...,n. \qquad (27)$$

Then, we determine the deprivation gap score (assessment of the degree of deprivation gap) for each deprivation symptom:

$$g_{hj,i} = \frac{1 - F\left(s_{hj,i}\right)}{1 - F\left(1\right)}, \; h=1,2,...,m; j=1,2,...,k_h; i=1,2,...,n, \qquad (28)$$

where: $F(s_{hj,i})$ – value of the distribution function of the non-deprivation gap score, regarding the j-th deprivation symptom in the h-th dimension, for the i-th individual.

Using the system of weights, the same that was applied in the calculation of FMI indicator, the non-deprivation gap score for the i-th individual, and for each deprivation dimension separately, is determined:

$$g_{h,i} = \frac{\sum_{j=1}^{k_h} w_{hj}\left(1 - g_{hj,i}\right)}{\sum_{j=1}^{k_h} w_{hj}}, \; h=1,2,...,m; i=1,2,...,n. \qquad (29)$$

In the next step the non-deprivation gap scores (29) are aggregated into the overall deprivation gap score indicating lack of deprivation gap, for the each i-th person, as the unweighted mean:

$$g_i = \frac{\sum_{h=1}^{m} g_{h,i}}{m},\ i=1,2,...,n. \tag{30}$$

The FSD indicator, for the *i*-th individual, is calculated as a combination of the (1-$F_{(SD,i)}$) indicator and the (1-$L_{(SD),i}$) indicator:

$$\mu_i(s) = FSD_i = \left(1 - F_{(SD),i}\right)^{\beta'-1}\left(1 - L_{(SD),i}\right),\ \text{i=1,2,...,n.} \tag{31}$$

The (1-$F_{(SD,i)}$) indicator, for the *i*-th individual, is the proportion of individuals non-deprivation gap score is higher (which are less deprived) than the individual concerned:

$$\mu_i(s) = FSD_i = \left(1 - F_{(SD),i}\right)^{\beta'} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma}{\sum_{\gamma=2}^{n} w_\gamma}\right)^{\beta'},\ i=1,2,...,n;\ \mu_n(s)=0, \tag{32}$$

where:

$F_{(SD),i}$ – value of the distribution function $F(g_i)$ of the lack of the deprivation gap score for the *i*-th individual,

w_γ – weight of the *i*-th individual of rank γ in the ascending lack of the deprivation gap score distribution,

β' – parameter.

The (1-$L_{(SD),i}$) indicator, for the *i*-th individual, is the share of the total non-deprivation gap score assigned to all individuals whose non-deprivation gap score is higher than the individual concerned:

$$\mu_i(s) = FSD_i = \left(1 - L_{(SD),i}\right)^{\beta'} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma g_\gamma}{\sum_{\gamma=2}^{n} w_\gamma g_\gamma}\right)^{\beta'},\ i=1,2,...,n;\ \mu_n(s)=0, \tag{33}$$

where: $L_{(SD),i}$ – value of the Lorenz curve of the non-deprivation gap score for the *i*-th individual.

Finally, the Fuzzy Supplementary Depth indicator (the depth of relative deprivation indicator) for the population is defined as the following mean:

$$FSD = \frac{\sum_{i=1}^{n} \mu_i(s) \cdot w_i}{\sum_{i=1}^{n} w_i}, \tag{34}$$

The parameter β' is estimated so that the FSD indicator for the population is equal to the income gap index defined by (2).

The FSD indicator for the *i*-th individual and for the each *h*-th deprivation dimension is calculated as follows:

$$\mu_i(s_h) = FSD_{h,i} = \left(1 - F_{(SD)h,i}\right)^{\beta'-1}\left(1 - L_{(SD)h,i}\right),\ h=1,2,...,m;\ i=1,2,...,n. \quad (35)$$

Finally, the Fuzzy Supplementary Depth indicators for each *h*-th deprivation dimension for the population are defined as:

$$FSD_h = \frac{\sum_{i=1}^{n} \mu_i(s_h) \cdot w_i}{\sum_{i=1}^{n} w_i},\ h=1,2,...,m. \quad (36)$$

Empirical results

The fuzzy poverty (monetary) measures are based on the household equivalised income variable. The household equivalised income is defined as the household disposable income divided by the household equivalence scale. The household disposable income includes net monetary incomes gained by all the household members reduced by property tax, inter-household cash transfers paid and statements for the Treasury Office. The equivalence scales are the parameters which allow for comparing the income of households of various sizes and demographic composition. We have used the modified OECD equivalence scale, which assigns a weight of 1 to the first adult household member, 0.5 to every other adult household member, and 0.3 to every child in the household under 14.

The poverty line applied for calculation traditional poverty indicators (1) and (2) is that of 60% of median equivalised disposable income after social transfers.

To measure the incidence and the depth of deprivation (non-monetary poverty) two initial steps, as it was mentioned, were necessary. At first, using the rich EU-SILC data, a subset of substantively meaningful and useful indicators (deprivation symptoms) for deprivation analysis were selected. Next, applying factor analysis, they were grouped into five dimensions (Whelaan *et al*., 2001). The final list of the selected deprivation dimensions and deprivation symptoms is presented in the Table 1[1]. All these indicators are considered at the household level and subsequently are assigned to the household members.

[1] The selection of deprivation symptoms and grouping them into deprivation dimension has been done by Caterina Ferreti from CRIDIRE, University of Siena (Italy).

Table 1. Dimensions and symptoms of deprivation

No.	Dimensions and indicators
1	**Basic life style – symptoms relate to the lack of ability to afford most basic requirements:**
1.1	Paying for one week annual holiday away home
1.2	Eating meal with meat, chicken, fish (or vegetarian equivalent) every second day
1.3	Keeping home adequately warm
1.4	Ability to make ends meet
2	**Household arrears and unexpected financial expenses - symptoms relate to arrears that the household has experienced and to unexpected financial expenses it faced in the last 12 months:**
2.1	Arrears on mortgage or rent payments
2.2	Arrears on utility bills
2.3	Arrears on hire purchase instalments or other loan payments
2.4	Facing unexpected financial expenses
3	**Housing facilities and deterioration – symptoms relate to the absence of basic housing facilities and to serious problems with the dwelling:**
3.1	A bath or shower in dwelling
3.2	An indoor flushing toilet for sole use of household
3.3	Leaky roof, damp walls/floors/foundation, or rot in window frames or floor
3.4	Too dark, not enough light in dwelling
4	**Environmental problems – symptoms relate to the neighbourhood and the environment:**
4.1	Noise from neighbours or from the street
4.2	Pollution, grime or other environmental problems
4.3	Crime violence or vandalism in the area
5	**Equipment of households in durables – symptoms relate to the lack of possession of a widely – desired durables because of lack of resources:**
5.1	A telephone (including mobile phone)
5.2	A colour TV
5.3	A computer
5.4	A washing machine
5.5	A car

Table 2. Fuzzy Incidence Indicators in Poland by voivodships in 2007

Voivodships	Indicator values in percentages						
	FMI	FSI	$FSI_{h=1}$	$FSI_{h=2}$	$FSI_{h=3}$	$FSI_{h=4}$	$FSI_{h=5}$
Dolnośląskie	17.7	20.0	21.8	13.0	15.3	20.2	13.8
	6.2*	5.6*	6.2*	7.4*	6.3*	6.2*	6.5*
Kujawsko-pomorskie	17.8	16.7	20.4	11.7	11.8	15.4	12.7
	7.3*	8.3*	9.3*	10.6*	11.9*	8.7*	9.2*
Lubelskie	24.2	19.0	23.1	12.8	17.0	10.5	14.3
	5.6*	7.2*	6.8*	9.9*	8.7*	10.6*	7.5*
Lubuskie	16.7	17.4	28.6	9.9	11.1	10.4	12.6
	10.4*	10.5*	9.3*	18.0*	13.0*	17.8*	11.6*
Łódzkie	17.6	20.8	24.6	11.6	17.8	16.6	13.6
	5.7*	6.0*	5.7*	8.6*	7.0*	7.9*	6.1*
Małopolskie	17.4	18.9	26.5	10.8	10.9	15.8	13.9
	5.9*	5.8*	5.9*	8.9*	8.0*	8.0*	6.6*
Mazowieckie	14.8	15.2	17.8	9.7	14.2	14.4	11.3
	4.9*	5.1*	5.4*	7.0*	5.6*	5.8*	5.5*
Opolskie	14.2	14.6	18.4	8.6	10.5	15.1	10.6
	14.2*	13.8*	12.7*	16.9*	17.6*	16.3*	14.3*
Podkarpackie	23.5	18.6	28.4	11.5	11.9	10.5	12.8
	6.4*	7.7*	6.6*	9.5*	8.9*	12.5*	7.5*
Podlaskie	17.7	11.7	16.8	9.9	11.2	7.9	12.9
	12.8*	12.9*	11.3*	16.5*	19.7*	16.7*	13.2*
Pomorskie	16.6	18.6	21.2	10.9	13.5	17.8	13.0
	8.0*	7.0*	7.6*	13.3*	9.2*	8.6*	8.0*
Śląskie	13.5	17.3	21.7	10.3	10.0	18.7	10.3
	5.2*	5.0*	4.6*	6.9*	6.6*	5.4*	6.2*
Świętokrzyskie	20.9	17.9	28.1	9.1	13.8	10.2	11.0
	7.8*	8.5*	8.2*	13.7*	11.3*	13.6*	12.0*
Warmińsko-mazurskie	21.5	18.8	26.0	9.1	14.2	12.2	12.5
	8.6*	8.0*	7.6*	14.9*	11.2*	14.7*	11.8*
Wielkopolskie	16.3	14.1	19.3	10.5	10.7	13.8	11.5
	6.1*	6.7*	6.7*	9.4*	9.7*	8.58	7.4*
Zachodniopomorskie	16.4	17.3	23.9	10.9	11.1	13.7	11.2
	8.4*	8.2*	8.5*	10.6*	11.2*	11.1*	10.5*
Poland	17.3	17.3	22.3	10.8	12.8	14.8	12.3
	2.2*	2.2*	2.0*	2.2*	2.4*	2.0*	2.1*

**Relative standard error · 100.*

Source: Central Statistical Office, EU-SILC Survey data, wave 3, version dated of 01.08.2009. Survey co-financed by UE. The views expressed are solely those of the author and should not be attributed to the European Commission.

Table 3. Fuzzy Depth Indicators in Poland by voivodships in 2007

Voivodships	Indicator values in percentages						
	FMD	FSD	$FSD_{h=1}$	$FSD_{h=2}$	$FSD_{h=3}$	$FSD_{h=4}$	$FSD_{h=5}$
Dolnośląskie	6.3	6.5	10.5	4.0	5.8	8.7	4.0
	10.9*	10.7*	10.3*	13.7*	9.9*	11.3*	12.1*
Kujawsko-pomorskie	4.7	4.2	8.2	3.8	5.1	5.8	3.8
	16.1*	17.5	13.8*	17.1*	18.2*	15.3*	18.4*
Lubelskie	8.0	5.3	10.0	4.4	7.5	4.6	3.8
	11.0*	13.1*	11.5*	15.4*	14.7*	17.4*	12.6*
Lubuskie	3.8	4.4	15.5	2.8	3.5	4.0	2.7
	20.9*	23.2*	15.3*	30.2*	21.5*	36.6*	20.1*
Łódzkie	5.2	7.8	12.2	4.3	8.9	8.0	3.8
	12.7*	11.0*	9.8*	14.5*	10.7*	12.3*	13.5*
Małopolskie	4.7	5.2	13.3	3.3	3.9	6.7	4.6
	13.4*	10.9*	9.1*	15.8*	12.8*	13.7*	11.6*
Mazowieckie	4.0	4.6	8.6	3.6	6.2	5.7	3.2
	10.0*	10.6*	8.8*	13.8*	9.6*	10.9*	9.7*
Opolskie	4.3	3.1	6.4	2.1	2.8	6.6	2.8
	27.6*	23.6*	22.2*	30.0*	24.6*	26.6*	22.7*
Podkarpackie	6.8	4.8	12.8	3.2	4.5	3.4	2.9
	12.7*	14.6*	11.7*	19.8*	13.8*	21.8*	14.3*
Podlaskie	4.6	2.6	6.5	2.8	4.7	2.5	4.0
	30.9*	37.4*	25.7*	24.1*	27.5*	25.3*	28.0*
Pomorskie	5.0	5.7	10.5	3.4	4.5	8.2	3.7
	14.5*	14.0*	13.0*	20.0*	14.8*	13.3*	17.1*
Śląskie	4.1	5.7	10.5	3.7	3.4	9.4	3.0
	9.9*	9.0*	7.1*	11.7*	10.3*	8.9*	11.1*
Świętokrzyskie	5.8	4.7	14.5	2.2	7.1	2.6	2.7
	15.4*	16.9*	14.3*	23.4*	16.9*	21.4*	20.4*
Warmińsko-mazurskie	6.7	4.7	11.9	2.9	5.7	5.6	2.4
	16.2*	16.0*	13.4*	19.7*	16.6*	20.6*	20.1*
Wielkopolskie	4.3	3.5	8.7	3.4	4.1	5.1	2.5
	13.4*	15.0*	13.0*	16.6*	18.2*	14.8*	14.3*
Zachodniopomorskie	4.3	4.5	12.5	3.2	4.1	6.2	2.8
	16.3*	16.5*	13.7*	20.1*	17.2*	18.5*	17.0*
Poland	5.0	5.0	10.6	3.5	5.2	6.3	3.4
	3.2*	3.2*	3.1*	3.4*	3.5*	3.6*	3.3*

**Relative standard error · 100.*

Source: Central Statistical Office, EU-SILC Survey data, wave 3, version dated of 01.08.2009. Survey co-financed by UE.

Fuzzy poverty incidence in Poland by voivodships

Voivodships of the greatest fuzzy poverty incidence (Table 2) were in 2007 Lubelskie, Podkarpackie, Warmińsko-mazurskie and Świętokrzyskie voivodships, and those with the lowest were Śląskie, Opolskie and Mazowieckie. The hierarchy of voivodships according to the fuzzy deprivation incidence is different than according to the fuzzy poverty incidence. The highest level of fuzzy deprivation incidence was observed in Dolnośląskie and Łódzkie voivodships, and the lowest - in Podlaskie, Wielkopolskie and Opolskie voivodships.

The greatest Fuzzy Supplementary Indicator values for the deprivation dimensions were noted in the households arrears and unexpected financial expenses dimension, and the lowest in the equipment of households in durables dimension. The hierarchy of voivodships considering fuzzy deprivation incidence with regard to individual dimensions was diversified.

Fuzzy poverty depth in Poland by voivodships

The highest fuzzy poverty depth indicator values were observed in 2007 in Lubelskie, Podkarpackie, Warmińsko-Mazurskie and Dolnośląskie voivodships, while the lowest values were noted in Lubuskie and Mazowieckie voivodships (Table 3). When it comes to the fuzzy deprivation depth, the worst situation was in Łódzkie and Dolnośląskie voivodships, and the best was in Podlaskie, Opolskie and Wielkopolskie.

Voivodships noted the greatest fuzzy poverty deprivation in the household arrears and unexpected financial expenses dimension, and the lowest in the equipment of households in durables dimension. The order of voivodships according to the fuzzy deprivation depth is different in deprivation individual dimensions.

Concluding remarks

The aim of this paper has been to further develop the methodology of multidimensional analysis of poverty by using the fuzzy set theory. In addition to income, other non-monetary indicators were incorporated in the analysis. An important contribution of the paper is the inclusion into poverty analysis, apart from the poverty incidence, also the poverty depth. The study examines the incidence and the depth of monetary and non-monetary poverty in Poland 2007 by voivodships.

The empirical results of the analysis have shown that poverty viewed by income does not coincide with poverty as seen by non-monetary characteristics of this phenomenon. Moreover, the hierarchy of voivodships considering fuzzy deprivation with regard to individual dimensions was diversified. This confirms the hypothesis that poverty in Poland has many dimensions and that its

measurement exclusively form income standing vantage point is highly insufficient. The finding of this research should form a useful source of information in order to implement socio-economic policies to counteract poverty.

REFERENCES

ABEL-SMITH B., TOWNSEND P. (1973): The Poor and the Poorest, in: Atkinson A. B. (eds.) *Wealth, Income and Inequality*, Penguin Education, Harmondsworth.

ATKINSON A. B. AND BOURGUIGNON F. (1982): The Comparison of Multidimensional Distribution of Economic Status, *Review of Economic Studies*, 49: 183–201.

BETTI G., CHELI B., LEMMI A. AND VERMA V. (2006): Multidimensional and Longitudinal Poverty: An Integrated Fuzzy Approach, in: Lemmi A. and Bett G. (eds) *Fuzzy Set Approach to Multidimensional Poverty Measurement*, 111–137, Springer, New York.

BETTI G., VERMA V. (1999): Measuring the Degree of Poverty in a Dynamic and Comparative Context: A Multi-dimensional Approach Using Fuzzy Set Theory, *Proceedings*, ICCS-VI, Vol. 11: 289–301, Lahore, Pakistan, August 27–31, 1999.

BETTI G., VERMA V. (2008): Fuzzy Measures of the Incidence of Relative Poverty and Deprivation: a Multi-dimensional Perspective, *Statistical Methods and Applications*,17: 225–250.

BOURGUIGNON F. AND CHAKRAVARTY S. R. (2003): The Measurement of Multidimensional Poverty, *Journal of Economic Inequality*, 1: 25–49.

CENTRAL STATISTICAL OFFICE (2009): *Incomes and Living Conditions of the Population in Poland. Report from the EU-SILC Survey of 2007 and 2008*, Warsaw.

CERIOLI A., ZANI S. (1990): A fuzzy approach to the measurement of poverty, in: Dagum C. and Zenga M. (eds): *Income and wealth distribution, inequality and poverty*, 272–284, Springer Verlag, Berlin.

CHELI B., LEMMI A. (1995): A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty, *Economic Notes*, 24: 115–134.

DESAI M., SHAH A. (1988): An Econometric Approach to the Measurement of Poverty, *Oxford Economic Papers*, 40(3), 505–522.

DEUTSCH, J., SILBER, J. (2005): Measuring Multidimensional Poverty: An Empirical Comparison of Various Approaches, *Review of Income and Wealth,* 51(1): 145–74.

Fuzzy Set Approach to Multidimensional Poverty Measurement (2006): Lemmi A. and Betti G. (eds), Springer, New York.

HAGENAARS A. J. M. (1986): *The Perception of Poverty*, North Holland, Amsterdam.

MARSHALL A. (1920): *Principles of Economics*, 8th ed., Macmillan, London.

MCCARTHY P. J., SNOWDEN C. B. (1985): The Bootstrap and Finite Population Sampling, *Vital and Health Statistics*, 2(95), U.S. Government Printing Office, Washington.

PANEK T. (1996): A Multidimensional Analysis of the Poverty in Poland in 1995 and 1996, *Statistics in Transition*, (3)5: 979–1002.

PANEK T. (2009): Poverty Indices in Multidimensional Approach, *Wiadomości Statystyczne* (in Polish), 12: 1–19.

SEN A. K. (1999): *Development as Freedom*, Oxford University Press, Oxford.

TOWNSEND P. (1979): *Poverty in the United Kingdom*, Penguin Books, Middlesex.

TSUI K. Y. (2002): Multidimensional Poverty Indices, *Social Choice and Welfare*, 19(1): 69–93.

WHELAN C. T., LAYTE R., MAITRE B., NOLAN B. (2001): Income, Deprivation and Economic Strain: An Analysis of the European Community Household Panel, *European Sociological Review*, 17: 357–372.

ZADEH L. A. (1965): Fuzzy Sets, *Information and Control,* 8: 338–353.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 381—402

USING ROC CURVES TO FIND THE CUT-OFF POINT IN LOGISTIC REGRESSION WITH UNBALANCED SAMPLES

Janusz Śmigielski[1][2], Anna Majdzińska[3], Witold Śmigielski[3]

1. Introduction

Logistic regression is widely used in many fields of science, e.g. medicine, psychology and anthropology. It was introduced in the 19th c. and its graphic form was developed by P. F. Verhulst and R. F. Pearl, who were the first to use the logistic model in practice to model population increase. The full model of logistic regression was used by Finney in 1972 [Stanisz, 2000, p. 205].

Logistic regression is used in models with a dichotomous endogenous variable, i.e. one taking only values 0 and 1 (e.g. healthy persons Y_i=0, sick persons Y_i=1). The probability of variable Y_i taking value 0 or 1 can be estimated by means of maximum likelihood method (see section 2). As far as the dichotomous variable is concerned, a frequently occurring problem is unbalanced sample, i.e. having the number of the values Y_i=0 considerably different from the number of values Y_i=1, for example, the number of healthy persons is usually much larger than of the sick ones. The classical method to find the cut-off point for the estimated probability P(Y_i=1) (in order to transform this probability into values 0 or 1 of the endogenous variable), may turn out quite ineffective. Therefore, the optimal cut-off value should be sought with methods other than the classical ones.

In the paper we propose using the concept of the receiver operator characteristic (ROC) curves in order to find an optimal cut-off point in logit models based on unbalanced samples. The proposed method will be compared with some other popular methods discussed in the literature.

[1] The authors wish to express their gratitude to Prof. dr. hab. Agnieszka Rossa for her valuable comments and guidelines, as well as all other assistance she extended during the writing of the article.
[2] Department of Medical Informatics and Statistics, Medical University in Łódź.
[3] Department of Demography, University of Łódź.

2. Logistic regression – the model's theoretical foundations

2.1. The form of a logistic regression model

It has already been mentioned in the introduction that the logit models are used to describe the influence of a number of exogenous variables $X_{1i}, X_{2i},..., X_{ki}$ on a dichotomous variable Y_i, which is a variable presenting two states of a described phenomenon observed for an i-th unit, i.e. 1 (a success) and 0 (a failure).

One reason for using a logit model in such a case flows among others from the undesirable properties of the linear regression model having a dichotomous variable as a dependent variable. One of the main drawbacks of the linear regression model with a binary dependent variable is that the model does not guarantee probability estimates $P(Y_i=1)$ within the interval [0,1]. Moreover, because the random error is heteroscedastic in this case and does not have the normal distribution, applying the Least Squares Method (LSM) to estimating the model parameters is not effective and the estimates of the error's variance can even be negative.

The traditional linear regression model is as follows:

$$Y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} + \varepsilon_i, \tag{1}$$

where $x_{1i}, x_{2i},..., x_{ki}$ are known values of the variables $X_{1i}, X_{2i},..., X_{ki}$, while ε_i stands for a random term whose expected value equals 0.

When variable Y_i is dichotomous, the conditional expected value of variable Y_i (given assumed values of the explanatory variables) is expressed by the formula:

$$E(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 1 \cdot P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$$

$$+ 0 \cdot P(Y_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \tag{2}$$

where $\boldsymbol{X_i} = [X_{1i}, X_{2i},..., X_{ki}]^T$, $\boldsymbol{x_i} = [x_{1i}, x_{2i},..., x_{ki}]^T$.

Denoting $p_i = P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$, we obtain from (1) and (2):

$$p_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} \tag{3}$$

The structural parameters in (3) show the amount of change in probability p_i (i.e. its increase or decrease) for a unit growth in the value of the exogenous variable at the given parameter, when the values of all other variables stay unchanged. The conclusion can be formulated that the structural parameters in the model (3) should not be outside the interval [-1,1], otherwise a unit growth in the exogenous variable will be accompanied by a change in probability p_i larger than a unit, which is unacceptable for probability. Therefore, if the obtained parameter

estimates do not belong to the interval, then their interpretation, as presented above, becomes pointless [Gruszczyński, 2002, pp. 54–56; Pruska, 2001, p. 91].

To eliminate such interpretation problems, the right-hand side of equation (3) can be transformed to obtain estimates p_i always within the interval [0,1]. This can be done by replacing (3) with the model:

$$p_i = F(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki}) \tag{4}$$

or, equivalently,

$$Y_i = F(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki}) + \varepsilon_i . \tag{5}$$

where F is a continuous and increasing function having following properties $\lim_{z\to\infty} F(z) = 1,\ \lim_{z\to-\infty} F(z) = 0.$ Such a model is called a binomial model.

In practice, the cumulative distribution function (CDF) of the logistic distribution[1] is frequently applied, which is defined by the formula:

$$F(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} . \tag{6}$$

Assuming therefore that F is the CDF of the logistic distribution, equation (4) can be written as follows [see Stanisz, 2000, pp. 207–208]:

$$p_i = P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji}\right)}}{1+e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_{ji} x_{ji}\right)}} \tag{7}$$

and equivalently

$$q_i = 1 - p_i = P(Y_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 1 - \frac{e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji}\right)}}{1+e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_{ji} x_{ji}\right)}} \tag{8}$$

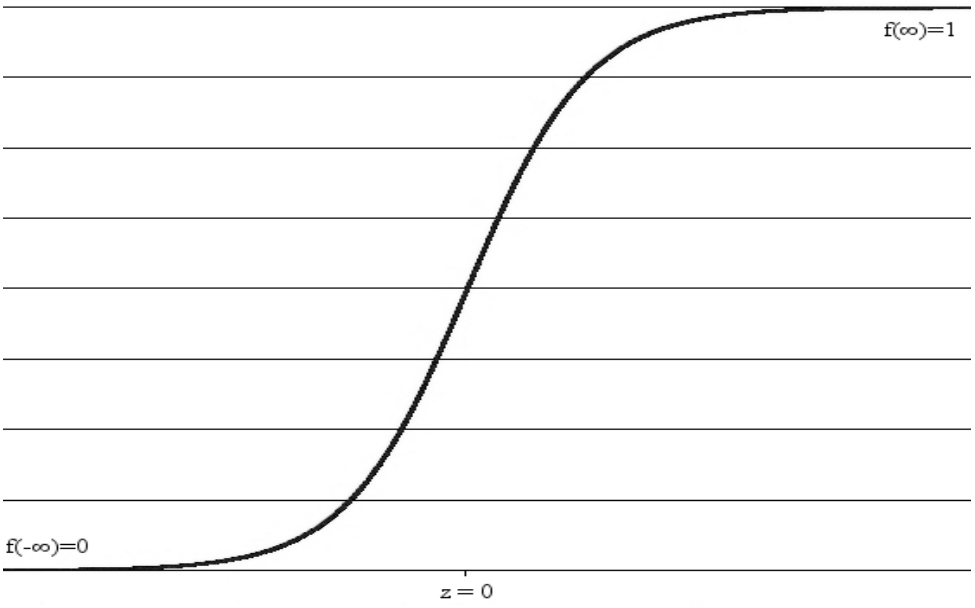
where α_j, $j = 0,...,k$ are the regression coefficients. The special case of binomial model defined in (7) is called a logistic regression model or a logit model.

Further, probabilities p_i and q_i in the models (7) and (8) will be generally denoted as $p(y_i \mid x_1, x_2, ..., x_k)$, where $y_i = 1$ or $y_i = 0.$

The example of the logistic function is graphically presented in Figure 1.

[1] To this end, the CDF of any continuous distribution can be applied, but usually the CDFs of the logistic or normal distribution are used.

Figure 1. Graphic representation of the logistic function



Source: developed by the authors.

The logistic curve is *S*-shaped. The probability p_i varies within the interval [0,1], so it can be treated as the probability of success occurring for an object characterised by the linear combination $z_i = \alpha^T \boldsymbol{x}_i$ [Gruszczyński et al., 2009, p. 142]. Initially, the function takes values close to 0, but after reaching some threshold value it grows rapidly and takes values approaching 1.

As mentioned, in the logistic model, the dichotomous dependent variable Y_i for the *i*-th unit takes two values denoted as 1 (a success) and 0 (a failure). In other words, the logistic model relates the probability of occurrence of either of the two possible results of variable Y_i to variables $\boldsymbol{X}_i$.

2.2. The likelihood function and parameter estimation for the logit model

Coefficients α_j of the logit model are usually estimated using the Maximum Likelihood Method (ML). A sufficiently large sample is needed for the purpose, i.e. $n > 10(k+1)$, where n is the sample size and k stands for the number of model parameters.

Let $[x_{1i}, x_{2i}, ..., x_{ki}]^T$ be the observation vector for the *i*-th sample unit and $\alpha_1,...,\alpha_k$ the unknown parameters of the logistic regression function to be estimated based on the sample. The likelihood function is given by the formula:

$$L = \prod_{i=1}^{n} p(y_i \mid x_{1i},...,x_{ki}) \tag{9}$$

where $p(y_i \mid x_{1i},...,x_{ki})$ stands for the probability of the dependent variable Y_i in the given regression model taking value y_i for known values x_{1i}, x_{2i}, ..., x_{ki}.

When Y_i equals 1 or 0, then the probability (9) can be written as:

$$L = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} \tag{10}$$

where p_i denotes the probability of $Y_i = 1$. In the case of a logistic regression model this probability is given by (7).

It is assumed in this paper that the estimators of the structural parameters $\alpha_1,...,\alpha_k$ are the ML-estimators [see Stanisz, 2000, pp. 208-209]. „An ML-estimator is a vector of parameters ensuring the largest probability of obtaining the observed values of the variables" [see Welfe A., 2003, p. 55].

When probabilities p_i are given by the formula (7), then finding the values of the parameters $\alpha_1,...,\alpha_k$ that maximise the function L (or, equivalently, the function ln L) involves the application of relevant numerical methods, such as the Newton-Raphson algorithm.

2.3. The odds ratio and the interpretation of logit model's parameters

After transforming (7) , we obtain

$$\ln \frac{p_i}{1-p_i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} \tag{11}$$

The expression under logarithm on the left-hand side in (11)

$$\frac{p_i}{1-p_i} \tag{12}$$

is known as an odds ratio, as it shows the probability of event (a success) occurrence in relation to the probability of its non-occurrence (a failure) [see Gruszczyński, 2001, p. 58; Gruszczyński et al., 2009, p. 169]:

Model (11) is equivalent to model (8). It is linear with respect to its parameters $\alpha_1,...,\alpha_k$ and with respect to the explanatory variables. The left-hand side of equation (11) is called a logit[1], i.e. it is the logarithm of the odds ratio.

[1] Analogous transformation of (4) where F stands for the CDF of the normal distribution is called a probit model.

The odds ratio is a tool applied to interpreting the parameters of the logit model. In general terms, the interpretation is as follows: if an *j*-th explanatory variable increases by a unit, then the odds ratio will change e^{α_j} times; in other words, the value of e^{α_j} shows how many times the odds ratio will increase or decrease for the *j*-th variable increasing by a unit (for *j*=1,…,*k*), *ceteris paribus*. A value of $e^{\alpha_j} > 1$ means that the odds of $Y_i = 1$ are growing, while for $e^{\alpha_j} < 1$ the odds are falling. The parameter α_j alone indicates the increase in the logarithm of the odds ratio [Gruszczyński et al., 2009, p. 169].

2.4. Forecasting the explained variable with a logit model

The binomial model, i.e. one describing a dichotomous variable Y_i (the logit model in this case), is mainly used for forecasting probabilities p_i, that is the probabilities of obtaining „a success", as well as forecasting the value of the dependent variable for new units with unknown values of Y_i.

The estimated probability p_i is usually transformed into a dichotomous variable Y_i using the so-called *standard forecasting method* that fixes the threshold value *c* at the level 0,5, i.e. with accepting the rule that $\hat{Y}_i = 1$ when $\hat{p}_i > 0{,}5$ and $\hat{Y}_i = 0$ when $\hat{p}_i \leq 0{,}5$ [Jeziorska-Pąpka, 2007, p. 277], where $\hat{p}_i$ denotes the estimated probabilitity that Y_i=1, and $\hat{Y}_i$ denotes the predicted outcome of the dependent variable Y_i. The above rule flows from the simple assumption that $\hat{Y}_i = 1$ when $\hat{p}_i > \hat{q}_i$, which reduces to the inequality $\hat{p}_i > 0{,}5$.

The standard rule performs well with a balanced sample, i.e. with a sample for which the number of observed successes is equal or "close" to the number of observed failures [Gruszczyński, 2001, p. 80]. However, when we deal with an unbalanced sample, then this approach may lead to poor predictive properties. „Upon fitting a logit model it is then invariably found that the estimated prediction probabilities $\hat{p}_i$ are quite high for Y_i=1, the outcome with the greater share, and very low for the outcome with lesser share" [Cramer, 1999, p. 3].

In the literature the following solutions are proposed to find a threshold value *c* for an unbalanced sample [Gruszczyński, 2001, p. 81; Jeziorska-Pąpka, 2007, p. 277; Dudek and Dybciak, 2006, p. 85]:

a) *Cramer's optimal threshold value*: $\hat{y} = 1$ when $\hat{p}_i > c$ and $\hat{y} = 0$ when $\hat{p}_i \leq c$ where c is the threshold value representing the share of 1s in the sample.
b) *Sample balancing*, by drawing two subsamples of the same size separately for both classes of the dependent variable. This sample composition is called a matched sample [Gruszczyński, 2001, p. 68]. In such a case the optimal cut-off point *c* is the value 0,5.

In our paper, we shall present the concept of using another method of finding the cut-off point c for the logit model's outcomes when the model is estimated from an unbalanced sample. The proposed method uses the properties of the so-called Receiver Operating Characteristic Curves (ROC) . The relevant theory can be found in section 3.

2.5. Goodness-of-fit of logit models

The literature of the subject usually proposes the following measures of model's goodness-of-fit [see Gruszczyński, 2001, pp. 64-66; Maddala, 1992, pp. 332-334]:

a) The Efron's determination coefficient R^2:

$$\text{b) } R^2_{Efrona} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n}{n_1 n_0}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2 \tag{13}$$

since

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = n_1 - n\left(\frac{n_1}{n}\right)^2 = \frac{n_1 n_0}{n},$$

where n_1 and n_0 are the numbers of actual outcomes in a sample for which $Y_i=y_i$, where $y_i = 1$ and $y_i = 0$, respectively.

c) The McFadden's determination coefficient R^2 based on the maximum likelihood function:

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}_{UR}}{\ln \hat{L}_R} \tag{14}$$

where $\ln \hat{L}_R$ is the maximal value of the logarithm of the likelihood function in a model containing only a constant, while $\ln \hat{L}_{UR}$ represents a maximal value of the logarithm of the likelihood function for a unrestricted model.

d) *Count* R^2 – is a measure of prediction accuracy, indicating the share of the correctly predicted outcomes of the model in their total number:

$$R^2 = \frac{n_{00} + n_{11}}{n} \tag{15}$$

where n_{00} and n_{11} are the numbers of observations for which, respectively, $\hat{Y}_i = Y_i = 0$ and $\hat{Y}_i = Y_i = 1$ (see table 1).

Table 1. The crossclassification of predicted and observed values of Y in a sample

Empirical	Forecasted		Total
	Y=1	Y=0	
Y=1	n_{11}	n_{10}	$n_{1.}$
Y=0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Source: Gruszczyński et al., 2009, p. 171.

The above measures take values within the interval [0,1].

e) The *ex post* crossclassification table (see table 1) allows constructing another measure of prediction accuracy (used in this paper). The measure is founded on the Yule's coefficient Q [see Domański, 2001, p. 184]. Using the notation from table 1, the measure Q can be defined as follows:

$$Q = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{11} \cdot n_{00} + n_{10} \cdot n_{01}} \tag{16}$$

The coefficient Q takes values in the interval [-1, 1]. The more it approaches 1, the better prediction accuracy of a given model.

3. The theoretical foundation of *ROC* curves

The *ROC* curve is a statistical tool that is used in many important fields of life and science. It was used for the first time during World War II for receiver's signal analysis. In the 1950s, the *ROC* curve was employed in signal detection theory and psychophysics, whereas in the 1979s and 1980s it turned out to be of use in medicine, particularly radiology, cardiology and epidemiology, and then in other fields of science, e.g. psychology, and the earth sciences, as well as financing and insurance [Krzanowski and Hand, 2009, p. 13].

ROC curves are used to assign the objects to two subpopulations according to the value of a continuous random variable, as they allow selecting a threshold value ensuring a possibly high probabilities of correct classifications.

Let us assume that a population of objects G consists of two subpopulations G_0 and G_1, such that $G = G_0 \cup G_1$ and $G_0 \cap G_1 = \emptyset$, where G_1 is the subpopulation of objects for which Y=1, and G_0 is the subpopulation of objects with Y=0. In other words, the variable Y indicates unit's membership to either of the two subpopulations – G_0 or G_1.

Let us assume that a unit has been drawn from a population G with an unknown value of the indicator Y. We can assign the unit to either G_0 or G_1 using the estimate $\hat{p}$ of the probability p=P(Y=1) that has been obtained from a fitted

logit model. In doing this, we are assuming that the unit will be assigned to G_1, when the inequality $\hat{p} > c$ is met, but for $\hat{p} \le c$ it will go to G_0. The point c will be called the decision threshold (or the cut-off point).

Let us consider the estimator $\hat{p}$ of the probability p derived from the adjusted logit model. Assume that H stand for the CDF of the estimator $\hat{p}$, i.e.

$$H(y) = P(\hat{p} \le y) \tag{17}$$

We will denote by H_0, H_1 the conditional CDFs of $\hat{p}$ in the subpopulation G_0 and G_1, respectively:

$$H_0(y) = P(\hat{p} \le y \mid G_0) \tag{18}$$

$$H_1(y) = P(\hat{p} \le y \mid G_1) \tag{19}$$

It is worth noting that for a given decision threshold $y=c$ the values $\overline{H}_1(c) = 1 - H_1(c)$ and $H_0(c)$ represent probabilities that a unit from G_1 or G_0, respectively, is correctly classified according to the classification rule just described.

Using the assumed notation, a *ROC* curve can be defined as follows:

$$ROC = \left\{ \left(\overline{H}_0(y), \overline{H}_1(y) \right), \quad y \in R \right\},$$

where

$$\overline{H}_1(y) = 1 - H_1(y), \ \overline{H}_0(y) = 1 - H_0(y). \tag{20}$$

In other words, the *ROC* curve is a set of points on a plane with the coordinates $(\overline{H}_0(y), \overline{H}_1(y))$ for $y \in R$.

In general, the theoretical ROC curve is usually not known, because the distributions H_0 and H_1 are unknown, so it is replaced by its empirical counterpart, i.e. by ROC_{emp} of the form:

$$ROC_{emp} = \left\{ \left(\hat{\overline{H}}_0(y), \hat{\overline{H}}_1(y) \right), \quad y \in R \right\}, \tag{21}$$

where $\hat{\overline{H}}_0(y)$ and $\hat{\overline{H}}_1(y)$ are the estimators of, respectively, $\overline{H}_0(y)$ and $\overline{H}_1(y)$. The estimators can be, for instance, derived as the complements of the empirical cumulative distribution functions from a training sample, i.e. a sample with known population membership indicators Y. Such a sample can be written as the following sequence:

$$(\hat{p}_1, Y_1), (\hat{p}_2, Y_2), \ldots, (\hat{p}_n, Y_n), \tag{22}$$

where Y_i, (i = 1, 2, …, n) are the observed outcomes of the indicator variable Y, while $\hat{p}_i$ are the estimated probabilities p_i that Y_i =1, i = 1, 2, …, n (i.e. probabilities derived from the logit model).

By plotting the ROC_{emp} curve we can find the empirical decision threshold $y=c$, i.e. a point producing as high empirical probabilities $\widehat{H}_0(y)$ and $\widehat{\overline{H}}_1(y)$ as possible, where $\widehat{H}_0(y) = 1 - \widehat{\overline{H}}_0(y)$. This procedure reduces to finding $y=c$ with coordinates $(\widehat{\overline{H}}_0(y), \widehat{\overline{H}}_1(y))$ on the plotted empirical ROC curve located the closest to the left upper corner of the unit square, i.e. the closest to the point with coordinates (0,1). Then $y=c$ is treated as the optimal decision threshold.

3. Examples of application of the presented cut-off rules

3.1. Characteristics of the databases and logit models used in the experiments

For the purposes of this article, four different datasets were used to fit the logit models. Each of the models contained two explanatory variables (statistically significant at the significance level 0.05). All the analysed datasets were unbalanced and were assembled based on observations of actual events. The first three models were fitted by using "moderately-sized" datasets (with n below 200). The statistical data were obtained mainly from own questionnaire surveys. The fourth dataset, containing 594 observations, was gathered from databases published in the Internet.

In the first model, a binary dependent variable „Emigration" represented the attitudes of students in their final years of study to economic emigration. The questionnaire survey[1] asked the students the following question: „If you received identical job offers, one in Poland and one from abroad, which one would you choose?". If the student declared to work in Poland, the answer was arbitrarily assigned 0. Answers indicating respondents' inclination to work abroad were given 1[2]. The explanatory variables used in the logit model were respondent's sex and the expected level of earnings (4 different levels of earnings were included). Table 2 presents results of the statistical inference received for the model. Since the article does not aspire to make a detailed analysis of the conclusions offered by the model, let us only state that males expecting substantial salaries were more frequent to choose emigration.

[1] The dataset that the authors used in their research was compiled on the basis of a questionnaire survey that was conducted among the students of the Faculty of Sociology and Economics, University of Łódź, and of the Medical University in Łódź in 2006 for the purposes of Joanna Śmigielska's master thesis: „Dobór i rekrutacja pracowników". The authors were authorised by the thesis author to use the database in this article.

[2] Observations where the marked answer was „I don't know" were omitted from the investigation.

Table 2. Results of the statistical inference for the logit model with the dependent variable „Emigration"

N=95	Model: Logistic regression (logit) (Emigration) Dependent variable: Emigration Loss: Maximum likelihood estimator, MSE max. 1 Final loss: 46,913010240 Chi2(2)=17,683 p=,00014 Constant	Earnings	Sex
Estimate	-5,78256	1,1367	1,33638
Standard error	1,46061	0,4538	0,51454
t(92)	-3,95902	2,5046	2,59725
Significance level p	0,00015	0,0140	0,01094
-95%CL	-8,68345	0,2353	0,31447
+95%CL	-2,88167	2,0380	2,35830
Wald's chi-square	15,67381	6,2729	6,74570
Significance level p	0,00008	0,0123	0,00940
Odds ratio	0,00308	3,1164	3,80525
-95%CL	0,00017	1,2653	1,36953
+95%CL	0,05604	7,6754	10,57295

Source: calculated by the authors with the Statistica software.

In order to explain some selected results contained in the table 2, we refer here to the odds ratio corresponding to the binary, explanatory variable "Sex". It follows from the last column and 9th row of the table 2 that the odds of having emigration inclination is 3,80525 times greater for males (Sex=1) than for females (Sex=0).

In the second model, the binary dependent variable „Kids" was the students' attitude to having multi-children families in the future. As in the first survey, the respondents were students in the final years, studying at the tertiary education institutions in Łódź. For a respondent wishing to have at least three children the observation was assigned 1, otherwise it was 0. The explanatory variables in this model were respondent's sex and the place of permanent residence. The statistical inference results for the model are presented in table 3. In this case, let us only state that males coming from small towns or countryside declared their wish to have a multi-children family the most frequently[1].

[1] More information on the same survey and the conclusions it provided can be found in the article: A. Majdzińska, W. Śmigielski, Wpływ religijności na decyzje dotyczące planowania życia rodzinnego studentów Uniwersytetu Łódzkiego, [in:] (ed.) J. T. Kowaleski, A. Rossa, Przyszłość demograficzna Polski, Folia Oeconomica 231, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2009 or on the website:
http://www.demografia.uni.lodz.pl/pubonline/Folia_Majdzinska_Smigielski.pdf.

Table 3. Results of the statistical inference for the logit model with the dependent variable „Kids"

	Model: Logistic regression (logit) (Kids) Dependent variable: Kids. Loss: Maximum likelihood estimator. MSE max. 1 Final loss: 61,550783827 Chi2(2)=9,2604 p=,00976		
N=129	Constant	Sex	Origin
Estimate	-4,10963	1,090751	0,64740
Standard error	1,02160	0,472151	0,27923
t(126)	-4,02275	2,310176	2,31850
Significance level p	0,00010	0,022505	0,02203
-95%CL	-6,13134	0,156379	0,09481
+95%CL	-2,08792	2,025124	1,19999
Walds chi-square	16,18249	5,336914	5,37544
Significance level p	0,00006	0,020885	0,02043
Odds ratio	0,01641	2,976510	1,91056
-95%CL	0,00217	1,169269	1,09945
+95%CL	0,12395	7,577050	3,32007

Source: calculated by the authors with the Statistica software.

In the third model, the dependent variable „Taxes" illustrated students' declarations as to their compliance with tax regulations after they become economically active. The data were collected in the course of a questionnaire survey[1]. If respondents declared that they would not pay taxes in the future according to the laws in force, the value assigned to the variable "Taxes" was 1, otherwise it was given 0. The explanatory variables that were statistically significant in this case were binary variables called „Ethics" and „Law". The former took the value 1, when the respondent believed that tax evasion was immoral, otherwise the value was 0. The variable „Law" was assigned 1 when the respondent believed that the possibility of being penalised by the tax administration if irregularities were found was the reason for paying taxes. For other answers, the variable took the value 0. The inference results for the model are presented in table 4.

[1] The database that we used in the investigation was compiled during the questionnaire survey conducted in 2006 among the students of the Faculty of Economics and and Sociology, University of Łódź, and of the Medical University in Łódź for the purposes of W. Śmigielski's master's thesis: „Unikanie opodatkowania jako zagadnienie etyczno-moralne w nauczaniu społecznym Kościoła Katolickiego".

Table 4. Results of the statistical inference for the logit model with the dependent variable „Taxes"

	Model: Logistic regression (logit) (taxes) Dependent variable: Taxes. Loss: Maximum likelihood estimator. MSE max. 1 Final loss: 33,443165686 Chi2(2)=25,873 p=,00000		
N=176	Constant	Ethics	Law
Estimate	-2,46129	-3,36931	1,69001
Standard error	0,60362	1,06492	0,71031
t(173)	-4,07757	-3,16390	2,37925
Significance level p	0,00007	0,00184	0,01844
-95%CL	-3,65270	-5,47122	0,28802
+95%CL	-1,26989	-1,26740	3,09200
Walds chi-square	16,62655	10,01028	5,66083
Significance level p	0,00005	0,00156	0,01735
Odds ratio	0,08532	0,03441	5,41951
-95%CL	0,02592	0,00421	1,33378
+95%CL	0,28086	0,28156	22,02098

Source: calculated by the authors with the Statistica software.

As far as the parameter estimates corresponding to variables „Ethics" and „Law" are analysed (table 4), it is worth noting that they have different signs. Almost all persons believing that tax evasion was immoral declared themselves as honest tax payers. On the other hand, the belief that the fear of the tax administration is what makes people pay taxes was frequently coupled with respondents' unwillingness to declare that they will pay their taxes in future as required by the law. This may suggest that the respondents' opinion about the effectiveness of the Polish tax inspectors is rather moderate.

In the fourth model, the binary dependent variable „Juventus" was the score of a football game played by Juventus Torino in the Italian *Serie A*. If the team did not lose the game, then the dependent variable was given 1, otherwise the assigned value was 0. The explanatory variables in the model were „Goals" and „Home/away". The first of them represented the number of goals the team scored during a game and the second variable was a binary one indicating whether the team played at home (Home/Away=1) or away (Home/Away=0)[1]. Table 5 presents detailed results of the statistical inference concerning this model.

[1] Match scores starting with the season 1991/92 until the last games in 2009 were obtained from the website www.wikipedia.pl for „Serie A".

Table 5. Results of the statistical inference for the logit model with the dependent variable „Juventus"

	Model: Logistic regression (logit) (Juventus) Dependent variable: Juventus. Loss: Maximum likelihood estimator. MSE max. 1 Final loss 184,75434375 Chi2(2)=128,84 p=0,0000		
N=594	Constant	Home/Away	Goals
Estimate	-0,92362	0,566605	-1,56049
Standard error	0,49799	0,274142	0,19004
t(591)	-1,85471	2,066829	-8,21158
Significance level p	0,06414	0,039185	0,00000
-95%CL	-1,90165	0,028194	-1,93372
+95%CL	0,05442	1,105016	-1,18726
Walds chi-square	3,43993	4,271783	67,43002
Significance level p	0,06365	0,038758	0,00000
Odds ratio	0,39708	1,762274	0,21003
-95%CL	0,14932	1,028595	0,14461
+95%CL	1,05593	3,019274	0,30505

Source: calculated by the authors with the Statistica software.

Analysing the parameter estimates (table 5) we intuitively arrive at rather obvious conclusions that Juventus odds of having a favourable result (i.e. not losing a match) increase for a match played at home and with the number of goals the team scores.

3.2. Determination of the optimal cut-off points for logit models and assessment of the models' prediction accuracy

The problem of low prediction accuracy for a dependent variable in an unbalanced sample in the case of the logit model was already brought up by J. Cramer in his article [Cramer J.S, 1999]. It should be noted that the prediction accuracy involving a binary endogenous variable depends not only on the appropriate specification of model (7) and on the correct estimation of its parameters, but also on the methodology one uses to find the cut-off point c, in which the estimates $\hat{p}_i$ of probabilities p_i are transformed into values 0 or 1 of the variable Y_i. The problem was already discussed in section 2.4.

Regarding the question of finding an optimal cut-off point c, Cramer proposed an alternative method to the standard forecasting method. Regarding the eight logit models that the author analysed, the comparison of *count* R^2 for forecasts with a standard cut-off point c at 0,5 and a cut-off point found using the Cramer's approach indicates that the standard method provides a higher *count* R^2 in each of the models. However, the Cramer's method produces more accurate predictions for one of two values of dependent variable Y which is characterized by a smaller share in the training sample. In one model only the result was actually identical. The Cramer's approach will be called hereafter the Cramer's rule.

The conclusion is that the standard method of finding the cut-off point generally provides good results for the case of unbalanced samples as far as the *count* R^2 is concerned; however, the frequency of correct forecasts for one of the dependent variable's values (i.e. 1 or 0) is quite frequently below 0,5, and sometimes it is exactly 0. This especially applies to this value of the dependent variable that has a smaller share in the unbalanced sample. The Cramer's rule for finding the cut-off point almost always improves the frequency of the correct *ex post* forecasts for such a value, however, the frequency of the correct forecasts for the second, predominating value of Y is worse. As a result, the *count* R^2 is usually smaller then in the standard method. Considering the context, however, the *count* R^2 does not seem to offer the complete picture of the model quality. It is therefore recommended to use in the comparisons also other measures of model's goodness-of-fit, such as Efron's R^2, which is equivalent to the ordinary coefficient of determination as far as the binomial model is concerned.

In the next section, we shall examine a proposed method for seeking an optimal cut-off point for unbalanced samples that builds on the properties of the ROC curves. This method will be called the *ROC rule*. The *ex post* prediction results obtained by means of the *ROC* rule and those provided by the *standard method* and the *Cramer's rule* are contrasted in tables 6 and 7.

Table 6 shows the frequencies with which the predicted values of the dependent variable agree with its actual values, i.e. the accuracy of the *ex post* forecasts for the four logit models discussed in section 3.1. Then, table 7 presents the basic prediction accuracy measures for the particular models. Since both the *count* R^2 values and the values of the *Q coefficient* depend on the method used to find the cut-off point, the point was determined separately for each of the three different rules. The best results are printed in bold to facilitate the analysis of the tables 6–7.

Table 6. Correct *ex post* forecasts for the dependent variable in the four logit models as produced by the three methods for determining the cut-off point

Dependent variable (logit model)	Sample characteristics		Percentages of correct predictions of outcomes (for two values *WD* and *WN* of the dependent variable)					
			WD			*WN*		
	n	WN%	standard method	Cramer's rule	ROC rule	standard method	Cramer's rule	ROC rule
Emigration	95	**27%**	**81,16%**	79,71%	79,71%	38,46%	**57,69%**	**57,69%**
Kids	129	**21%**	**96,08%**	50,00%	50,00%	18,52%	**70,37%**	**70,37%**
Taxes	145	**9%**	**100,00%**	67,38%	67,38%	0,00%	**92,31%**	**92,31%**
Juventus	594	**15%**	**94,47%**	75,49%	91,70%	39,77%	**80,68%**	52,10%

n – sample size. D – value of the dependent variable accounting for the larger share in the sample. WN – value of the dependent variable accounting for the smaller share in the sample. WN %– the percentage share of WN's values in a dataset.

Source: calculated by the authors with the Statistica software.

Table 7. Selected measures of prediction accuracy for the four logit models as produced by the three methods for determining the cut-off point

	Sample characteristics			Count R^2			Q coefficient		
Dependent variable (logit model)	n	WN%	Efron's R^2	Standard rule	Cramer's rule	ROC rule	Standard rule	Cramer's rule	ROC rule
Emigration	95	**27%**	0,14	0,69	**0,74**	**0,74**	0,46	**0,68**	**0,68**
Kids	129	**21%**	0,08	**0,80**	0,54	0,54	**0,70**	0,41	0,41
Taxes	145	**9%**	0,21	**0,93**	0,69	0,69	×	**0,92**	**0,92**
Juventus	594	**15%**	**0,25**	0,86	0,76	**0,87**	0,84	0,86	**0,87**

n – sample size. WN – value of the dependent variable accounting for the smaller share in the sample. WN % – the percentage share of WN's values in a sample. × – denotes that the percentage of correct predictions took value 0 and the Q coefficient could not be computed.

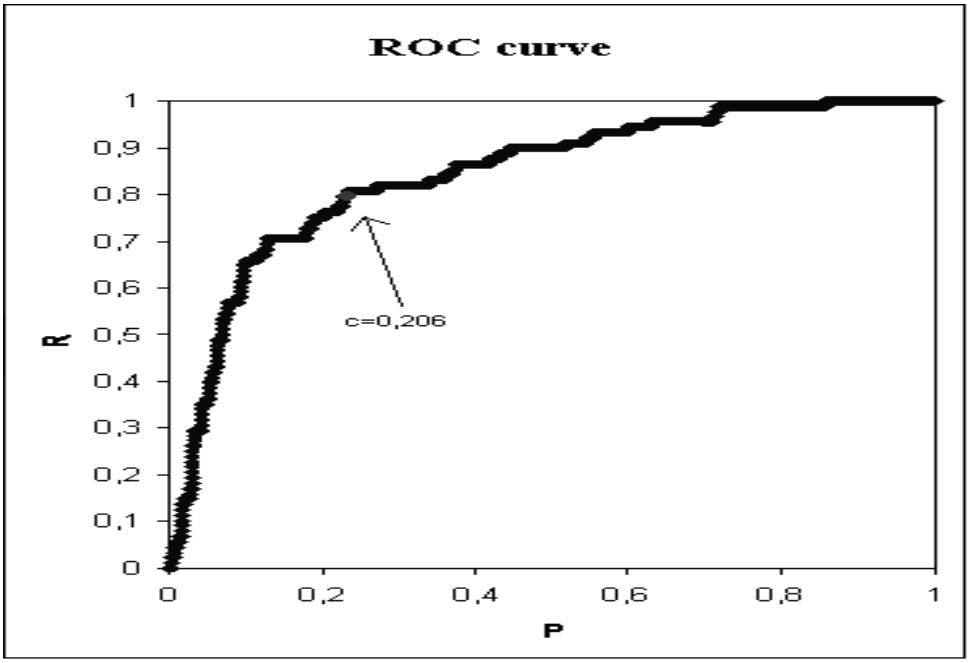
Source: calculated by the authors with the Statistica software.

It is worth noting that the *count* R^2 computed using the standard rule for three of the four models exceeds 0.8 (thus the results are similar as discussed in the Cramer's article), which is a relatively high value. On the other hand, the percentage of consistency between the predicted values of dependent variables and their actual values which represent the smaller share in the respective datasets

is lower than 50% in the case of the standard rule for each of the four models, but for the variable „Taxes" it is 0% (see table 6).

Another notable fact is that the prediction accuracy measures (i.e. *count* R^2 and *Q coefficient*) obtained by employing the *Cramer's rule* and the *ROC rule* is identical for the first three models. In the authors' opinion, this situation arises from the relatively limited sample sizes. As far as the forecasts of the variable „Juventus" are concerned, the *ROC rule* performs the best, as it provides the highest values of *count* R^2 and *Q coefficient*. Figure 2 illustrates the empirical ROC curve for this case. The empirical ROC curves for the other models are included in Appendix 1.

Figure 2. The empirical ROC curve with the optimal cut-off point for the logit model with the dependent variable „Juventus"



Legend[1]*:* $P=\hat{\overline{H}}_0(y)$, $R=\hat{\overline{H}}_1(y)$.

Source: calculated by the authors with the Microsoft Excel worksheet.

Table 8 compares the cut-off point values calculated using the *Cramer's rule* and the *ROC rule* for the four models.

[1] The same symbols apply to the other ROC curves.

Table 8. The cut-off points as produced by the *Cramer's rule* and the *ROC rule* for the four logit models

Dependent variable (model)	Cut-off point		Absolute diference $\|c_1-c_2\|$
	Cramer's rule (point c_1)	ROC rule (point c_2)	
Emigration	0,274	0,262	0,012
Kids	0,209	0,217	0,008
Taxes	0,074	0,079	0,005
Juventus	0,148	0,206	0,058

Source: calculated by the authors with the Microsoft Excel worksheet.

Analysing the data in table 8 we find that for training samples containing relatively small numbers of observations the cut-off point provided by the *ROC rule* is almost fully consistent with that obtained using the *Cramer's rule* (hence, the forecasts provided by both rules in the first three models are the same). However, in the model with the dependent variable „Juventus" the points found using both rules are noticeably different, which affects the *ex post* forecast results.

As shown by the Efron's R^2 values[1] the best fit of the model is in the cases of the variable „Juventus". Hence, we decided to assess the quality of the model again, this using the data from the so-called test sample that was not involved in model estimation. The test sample was compiled using the scores of the matches played by *Juventus Torino in Serie A* in the four successive seasons: from 1987/88 to 1990/1991 (n=132 matches in total). Table 9 presents results for the logit model with parameter estimates given in table 5, but forecast results derived with respect to the test sample using the three discussed rules.

Table 9. Forecast results for the dependent variable „Juventus" received for the test sample

Rule	Percentage of correct forecasts in the test sample (for two values WD_t and WN_t of the dependent variable)		Values of prediction accuracy measures in the test sample	
	WD_t	WN_t	Count R^2	**Q**
Standard	**84,16%**	70,97%	**0,81**	**0,86**
Cramer's	67,74%	**75,25%**	0,73	0,73
ROC	**84,16%**	67,74%	0,80	0,84

WD_t – value of the dependent variable accounting for the larger share in the test sample.
WN_t – value of the dependent variable accounting for the smaller share in the test sample.

Source: calculated by the authors with the Microsoft Excel.

[1] As we can read in statistical articels Effron's R^2 value is Fast always not very high. See more: Morrison D. G., *Upper Bounds for Correlation Between Binary Outcomes and Probabilistic Predictions*, JASA, vol. 67, no. 337/.

It is worth noting that the Q values in table 9 are close to 1, however they are higher than for the training sample data (table 7). Compared with the first model, though, the percentage of the correct forecasts for those values of the dependent variable, which are represented by the smaller portion of the test observations, is higher (this concerns values of the dependent variable denoted by WN_t in table 9). The best, i.e. the greatest, value of Q is obtained with the *standard rule*, although it is only slightly better than the *ROC rule*. The poorest results were obtained for the cut-off point determined using the *Cramer's rule.*

4. Summation and final conclusions

When the logit models are based on unbalanced samples then it becomes necessary to use methods allowing the determination of a cut-off point that are alternative to the *standard rule*. The *Cramer's rule* usually increases the pecentage of the correct forecasts of the dependent variable, but it frequently involves lower values of *count* R^2 and/or of the *Q coefficient*. The *ROC rule* application to computing the optimal cut-off points that the authors proposed has not yielded better results for the models with relatively small sample sizes, but for the model based on a large number of observations (almost 600 in this study) the procedure improved the frequency of the correct forecasts concerning those values of the dependent variable which are represented by a smaller part of the observations. It allowed also obtaining higher values for both *count* R^2 and *Q coefficient* (it must be borne in mind, though, that *count* R^2 provides rather general information about the prediction accuracy).

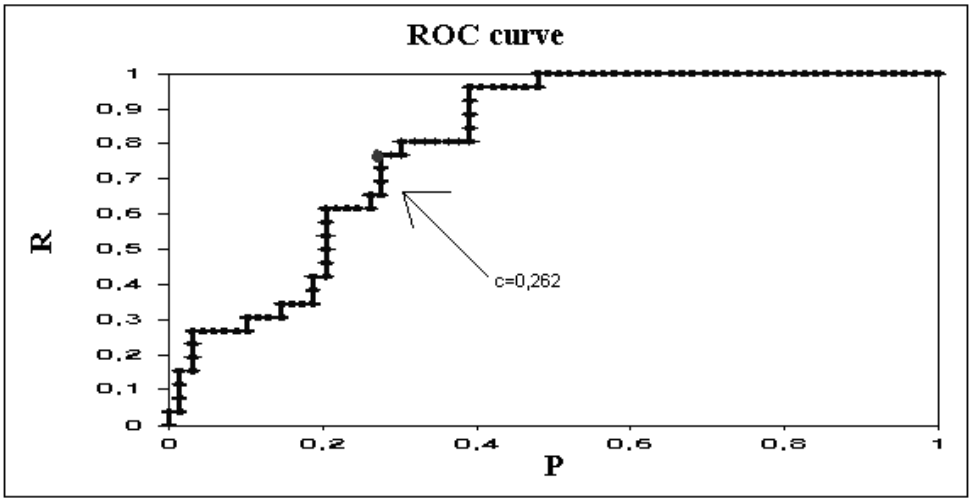
The authors are aware that further research is needed to confirm the conclusions offered by the study. This article aims at demonstrating that using the ROC curve is justified when the binary dependent variables are forecasted using, for instance, the logit models. The presented examples show that the method can produce results that are better or comparable with those offered by methods that are more common in the literature and applied more often.

Acknowledgements

We acknowledge an anonymous reviewer for valuable comments and suggestions which improved the manuscript.

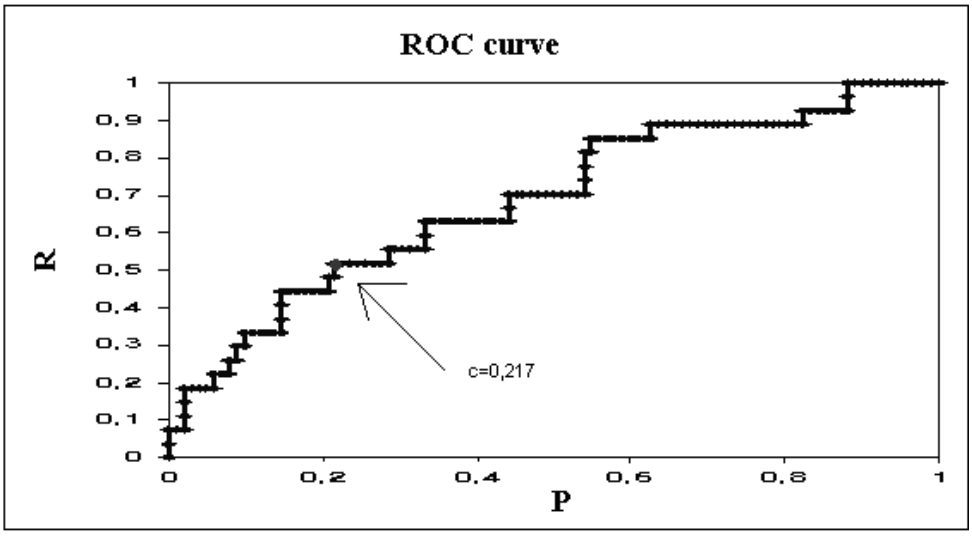
Appendix

Figure A.1. The empirical ROC curve for the model with the dependent variable „Emigration"



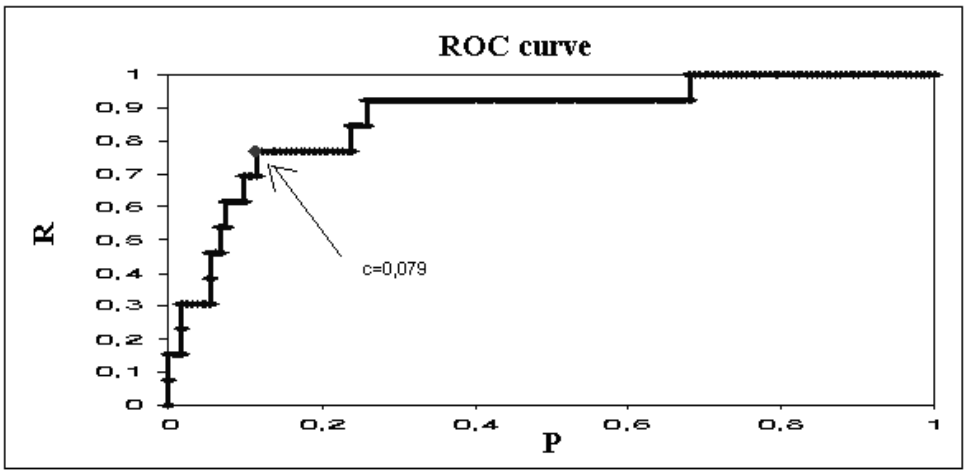
Source: calculated by the authors with the Microsoft Excel worksheet.

Chart 3. The empirical ROC curve for the model with the dependent variable „Kids"



Source: developed by the authors using the Microsoft Excel worksheet.

Chart 4. The empirical ROC curve for the model with the dependent variable „Taxes"



Source: calculated by the authors with the Microsoft Excel worksheet.

REFERENCES

CRAMER J. S. (1999). Predictive performance of the binary logit model in unbalanced samples, [in]: Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 48, No. 1, pp. 85-94, Blackwell Publishing, Oxford.

DOMAŃSKI CZ. (ed) (2001) Metody statystyczne. Teoria i zadania, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

DUDEK H., DYBCIAK M. (2006) Zastosowanie modelu logitowego do analizy wyników egzaminu [in]: Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Ekonomika i Organizacja Gospodarki Żywnościowej, nr 60, Wydawnictwo SGGW, Warsaw.

GRUSZCZYŃSKI M. (2001) Modele i prognozy zmiennych jakościowych w finansach i bankowości, Monografie i opracowania, no. 490, Szkoła Główka Handlowa, Warsaw.

GRUSZCZYŃSKI M., KUSZEWSKI T. AND PODGÓRKSA M. (2009) Ekonometria i badania operacyjne. Podręcznik dla studiów licencjackich, PWN, Warsaw.

JEZIORSKA-PĄPKA M. (2007) Zastosowanie modeli dwumianowych do opisu asymetrii informacji na rynku ubezpieczeń na przykładzie polis komunikacyjnych OC [in]: Dynamiczne Modele Ekonometryczne, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

KRZANOWSKI W., HAND D. (2009) ROC curves for continuous data, Monographs on statistics and applied probability, 111, CRC Press, New York.

MADDALA G. S. (1992) Introduction to Econometrics – 2nd ed., Macmillan Publishing Company, New York.

MAJDZIŃSKA A., ŚMIGIELSKI W. (2009) Wpływ religijności na decyzje dotyczące planowania życia rodzinnego studentów Uniwersytetu Łódzkiego, [in:] (ed.) J. T. Kowaleski, A. Rossa, Przyszłość demograficzna Polski, Folia Oeconomica 231, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

MORRISON D. G. (1972) *Upper Bounds for Correlation Between Binary Outcomes and Probabilistic Predictions*, JASA, vol. 67, no. 337.

PRUSKA K. (2001) Modele probitowe i logitowe w programach nauczania studiów ekonomicznych [in:] Metody analizy cech jakościowych w procesie podejmowania decyzji (conference proceedings), Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

ROSSA A. (2007) Asymptotic tests for receiver operating characteristic curves [in]: Statistics in Transition – new series, vol. 8, no. 3, GUS, Warsaw.

STANISZ A. (2000) Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny, vol. II, Kraków.

WELFE A. (2003) Ekonometria. Metody i zastosowanie, PWE, Warsaw.

WELFE A., BRZESZCZYŃSKI J. AND MAJSTEREK M. (2002) Angielsko-polski, polsko-angielski słownik terminów metod ilościowych, Polskie Wydawnictwo Ekonomoiczne, Warsaw.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 403—422

INFLUENCE OF NON-RESPONSE IN BUSINESS TENDENCY SURVEYS ON PROPERTIES OF EXPECTATIONS

Emilia Tomczyk[1], Barbara Kowalczyk[2]

ABSTRACT

Expectations concerning key economic variables certainly influence decisions undertaken by economic agents. Since assumption of rationality forms the basis of neoclassical economic theory, question of whether expectations of industrial enterprises are indeed formed rationally deserves careful attention. Direct data on expectations are available mainly through business tendency surveys. Non-response problem is present in almost every survey, and much higher non-response rates are observed for expectations than for realizations. Weighting systems used to control for size of respondents may also introduce bias into expectations data derived from business tendency surveys. In this paper, we analyze two basic properties of expectations rational in the sense introduced by J. F. Muth – that is, unbiasedness and orthogonality – taking into account issue of non-response and weighting schemes. We propose several sample balance statistics to correct for changing sample structure that results from non-response; it depends on time, and is different from general population. We find that rationality of expectations of Polish industrial firms is not sensitive to these factors: independently from non-response and weighting issues, expectations concerning relative changes in production remain unbiased but not efficient with respect to freely available information.

Key words: expectations, rationality, tendency surveys, industrial production, non-response, weighting, survey data, qualitative data, post-stratification

1. Introduction

Economic agents are usually assumed to be rational; and while the term itself is variously defined, ranging from strict to bounded to imperfect rationality, the rationality assumption remains at the core of modern economics. Yet its accuracy

[1] Warsaw School of Economics, Poland, e-mail: Emilia.Tomczyk@sgh.waw.pl.
[2] Warsaw School of Economics, Poland, e-mail: Barbara.Kowalczyk@sgh.waw.pl.

and realism are often called into question, and tests of rationality constitute a major branch of modern economic research.

In this paper, we analyze properties of rational expectations, as introduced in 1961 by J. F. Muth, taking several weighting systems and non-response issues into consideration. Our previous work on influence of non-response on rationality (Kowalczyk, Tomczyk [2008]) was based on contingency tables summarizing individual-level data; now we turn to classical rationality tests and in addition to non-response problem, we also consider several weighting systems employed to scale survey data.

In section 2 we briefly describe Muth's Rational Expectations Hypothesis and methods of testing properties of rational expectations. In section 3 we employ three different weighting systems to business tendency survey data, and in section 4, we present problems emerging from non-response and propose a balance statistic that accounts for sample structure variability. In sections 5 and 6, we present and compare balance statistics, taking into account non-response and various weighting schemes, and describe quantification procedures. In section 7, we present results of unbiasedness and orthogonality tests with respect to production expectations expressed by Polish firms in business tendency surveys, taking a range of weighting schemes as well as and non-response into consideration. Section 8 concludes.

2. Rational Expectations Hypothesis and its testing

Tests of rationality constitute major part of the economic research on expectations; popularity of this branch of analysis is well documented in both theoretical and empirical literature. Numerous publications focus on the Rational Expectations Hypothesis (REH), introduced in 1961 by J. F. Muth. He defines expectations as rational if, being educated forecasts of future values of economic variables, they are equal to expected values of these variables as reflected in predictions formed on the basis of the relevant economic theory. REH postulates that economic agents make use of all available (and pertinent) information in timely and effective manner, and that they understand their environment well enough to correctly predict its future behaviour. While many criticisms have been aimed at REH since its introduction, majority of them citing information asymmetries, information selection and processing costs, and influence of learning processes, its importance as an empirical hypothesis subject to empirical verification has not been questioned.

Muth's formulation of REH is very general. Several specific tests of rationality of expectations have been proposed for the purpose of its empirical verification, the most common being tests of unbiasedness and orthogonality.

Expectations are considered unbiased if they do not systematically overestimate or underestimate values of an economic variable. Let ${}_t x_{t+s}$ stand for relative change in variable x between t and $t + s$, as noted in official statistics, and

${}_t x^e_{t+s}$ – expected relative change in the same period, as derived from survey data.[1] The standard unbiasedness test of expectations is defined by the hypothesis

$$\mathrm{H}_0\colon\ \alpha_0 = 0\,,\ \ \alpha_1 = 1\,, \tag{2.1}$$

where α_0 and α_1 are parameters of the regression equation

$${}_t x_{t+s} = \alpha_0 + \alpha_1 \cdot {}_t x^e_{t+s} + \xi_t\,, \tag{2.2}$$

and error term ξ_t is assumed to be a white noise process.[2]

Property of orthogonality is defined through expectations error, that is, difference between observed and expected values of a variable. Expectations are termed orthogonal if expectations error is uncorrelated with information available at the moment when expectations were formed; that is, all relevant information has already been incorporated into the forecast. Test of orthogonality of expectations error in relation to information set **Π** is described by the hypothesis

$$\mathrm{H}_0\colon\ \alpha_i = 0\,,\quad i = s,\, s+1,\, ...,\, T, \tag{2.3}$$

where α_i are parameters of the regression equation

$$\left({}_t x_{t+s} - {}_t x^e_{t+s}\right) = \alpha_0 + \sum_{i=t+s}^{T} \alpha_i z_{t-i} + \xi_t\,, \tag{2.4}$$

$z_{t-i} \in \mathbf{\Pi}$, and error term ξ_t is a white noise process. Orthogonality tests require that elements of information set faced by economic agents (that is, variables z_{t-i}) be specified. In empirical setting, they include arbitrarily selected set consisting of series that are likely to have been considered relevant by economic agents. If expectation errors are not orthogonal to freely available and relevant information, then forecasting process may be interpreted as inefficient because expectations could be improved by incorporating information provided by variables z_{t-i}.

Standard approach to evaluating REH is based on tests if observed expectations series fulfill conditions for unbiasedness (2.1) and orthogonality (2.3). Before tests of these properties can be undertaken, expectations series ${}_t x^e_{t+s}$ must be obtained – typically on the basis of survey data or, rarely, controlled experiments. Rapid development of questionnaires as source of information on expectations dates back to the second half of the 20th century and coincides with development of business conditions surveys. They combine assessment of current situation with expectations (forecasts) and constitute a promising source of data

[1] Methods of derivation of quantitative expectations series on the basis of qualitative survey data are briefly described in section 5.

[2] The RC (restricted cointegration) unbiasedness test (Liu, Maddala [1992]) is not considered here due to small sample available for analysis.

for tests of rationality (as well as data necessary for quantification procedures; see section 6). It must be stressed that in surveys expectations are declared and not directly observed.[1]

Since the 1970s extensive literature on rationality of expectations has been published. Empirical results have generally been inconclusive, and highly dependent on time period considered, variables selected for analysis, methods of data aggregation, forecast horizon, specification of econometric tests, and other factors. In Poland, subject of rationality of economic agents emerged along with transformation of the Polish economic system from centrally planned to market economy in the early 1990s. Tests of properties of REH carried out on Polish data on expectations provided results similar to those obtained from research conducted in the United States and Western Europe, that is, sensitive to several factors and not leading to unambiguous results (see Osińska [2000], Łyziak [2003], Tomczyk [2004, 2008]).

In this paper, we aim to re-address the issue and to contribute to the still relatively new field of tests of rationality for Polish economic agents. In addition to testing properties exhibited by expectations rational in the sense introduced by J. F. Muth, we consider two additional dimensions: problem of non-response in business tendency surveys that supply expectations data, and weighting systems employed to adjust the original data for differences in respondent size.

Non-response is widely cited as the main source of non-sampling errors in survey data. Errors resulting from non-response are probably non-random; it cannot be safely assumed that non-response has no systematic effect on results of tests of rationality because respondents who refuse to answer may also exhibit unique characteristics that influence their answers and are correlated with their degree of rationality. Non-response also introduces variability of sample structure in time; we address this problem in section 4.

As far as we are aware, weighting systems have not been analyzed from the point of view of properties of expectations formation processes. In case of industrial enterprises, weight is assigned according to size reflected in employment numbers or value of production. We propose to test if various weighting schemes influence the results of rationality tests. Since weighting amounts to granting more influence to data supplied by a large enterprise than a small one, expectations count 'more' if expressed by a large firm. We propose to verify if the results of rationality tests differ when every respondent is treated equally, or according to several reasonable weighting systems, described in the next section.

[1] In macroeconomic modeling, indirect measurement of expectations is also considered; for summary, see Sheffrin [1996].

3. Weighting systems

The unweighted balance statistics are defined for current situation evaluated by survey respondents as

$$ {}_{t}BA_{t+k} = {}_{t}A^{1}_{t+k} - {}_{t}A^{3}_{t+k}, \tag{3.1} $$

and for expectations as

$$ {}_{t}BP_{t+k} = {}_{t}P^{1}_{t+k} - {}_{t}P^{3}_{t+k}, \tag{3.2} $$

where

${}_{t}A^{1}_{t+k}$ – percentage of respondents reporting improvement between t and $t + k$,

${}_{t}A^{2}_{t+k}$ – percentage of respondents reporting no change between t and $t + k$,

${}_{t}A^{3}_{t+k}$ – percentage of respondents reporting decline between t and $t + k$,

${}_{t}P^{1}_{t+k}$ – percentage of respondents expecting improvement between t and $t + k$,

${}_{t}P^{2}_{t+k}$ – percentage of respondents expecting no change between t and $t + k$,

${}_{t}P^{3}_{t+k}$ – percentage of respondents expecting decline between t and $t + k$.

In case of industrial enterprises, size of employment or value of turnover are typically used as weights. Data on expectations of Polish industrial enterprises have been collected since 1986 by the Research Institute for Economic Development (RIED) at the Warsaw School of Economics through business tendency surveys. Launched for manufacturing industry, currently they also cover households, farming sector, exporters, construction industry, and banking sector. Empirical part of this paper is based on the monthly survey addressed to industrial enterprises. Each survey question asks respondents to evaluate both current situation (as compared to last month) and expectations for the next 3 – 4 months by assigning them to one of three categories: increase / improvement, no change, or decrease / decline (see Appendix 1). Aggregated survey results are regularly published and commented on in RIED bulletins: each month, a number of respondents is announced, along with a percentage of respondents who observed increase / no change / decline and who expect increase / no change / decline in a given area of economic activity, along with a balance statistic calculated as a difference between percentage of 'optimists' (those who judge current situation favorably or predict improvement) and 'pessimists' (those who evaluate present situation unfavorably or predict decline), according to formulas (3.1) and (3.2). Because of ambiguous wording of the questionnaire ("expectations for the next 3 – 4 months"), clarification of the forecast horizon is necessary. On the basis of previous analysis of the RIED data (see Tomczyk [2004]) we are able to limit our attention to the three-month horizon. That is, for the remaining part of the paper,

expectations horizon $k = 3$. When evaluating the current state, respondents are asked for comparison with previous month, hence for realizations $k = 1$.

In RIED survey, neither size of employment nor turnover are known; only the interval to which employment belongs is given. Five intervals are distinguished:

- up to 50 persons,
- 51 to 250 persons,
- 251 to 500 persons,
- 501 to 2000 persons,
- over 2000 persons.

In the RIED survey, respondents are weighted by 1, 2, 3, 4 and 5 respectively, depending to which employment interval the given enterprise belongs (see RIED [2008]). This arbitrarily chosen system of weights, and consequently results obtained on its basis, may be questioned. As far as we are aware, no comparison of values of balance statistics calculated for different weighting systems, or of expectations series derived on the basis of these systems, has been attempted so far for Polish survey data.

We propose to analyze three different systems of weights:

1. **No weighting.** Respondents are not weighted – that is, every respondent receives the same weight equal to 1.
2. **RIED weighting.** Respondents are weighted by 1, 2, 3, 4 and 5 respectively, depending on which employment interval they belong to.
3. **Weighting by lower limit of the employment interval.** Respondents are weighted by 1, 51, 251, 501 and 2001 respectively, depending on which employment interval they belong to.[1]

To obtain the mathematical form of balance statistics, first we define the variables x_A, x_P and y_A, y_P as:

$$x_A = \begin{cases} 1 & \text{if respondent reported improvement between t and } t+k \\ 0 & \text{in other case} \end{cases}$$

$$y_A = \begin{cases} 1 & \text{if respondent reported detoriation between t and } t+k \\ 0 & \text{in other case} \end{cases}$$

$$x_P = \begin{cases} 1 & \text{if respondent is expecting improvement between t and } t+k \\ 0 & \text{in other case} \end{cases}$$

$$y_P = \begin{cases} 1 & \text{if respondent is expecting detoriation between t and } t+k \\ 0 & \text{in other case} \end{cases}$$

[1] The lower limit has been chosen because the upper limit is not attained in case of the largest enterprises, and the lower limit seems to adequately account for differences in sizes of enterprises.

Weighted balance statistics defined for current situation evaluated by survey respondents are then obtained as[1]:

$$ {}_t BA_{t+k} = \frac{\sum w_1 x_{Ai} + ... + \sum w_5 x_{Ai}}{n_1 w_1 + ... + n_5 w_5} \cdot 100 - \frac{\sum w_1 y_{Ai} + ... + \sum w_5 y_{Ai}}{n_1 w_1 + ... + n_5 w_5} \cdot 100 \quad (3.3) $$

and for expectations as:

$$ {}_t BP_{t+k} = \frac{\sum w_1 x_{Pi} + ... + \sum w_5 x_{Pi}}{n_1 w_1 + ... + n_5 w_5} \cdot 100 - \frac{\sum w_1 y_{Pi} + ... + \sum w_5 y_{Pi}}{n_1 w_1 + ... + n_5 w_5} \cdot 100 \quad (3.4) $$

It is clear that for $w_i = 1$, i =1, 2, …, 5 formulas (3.3) and (3.4) reduce to (3.1) and (3.2), respectively.

4. Problem of non-response

Non-response is present in almost all surveys, but the extent and the effect of non-response can vary greatly from one type of survey to another. In RIED business tendency survey the problem of non-response is very significant. Table 4.1 shows the planned sample sizes (number of questionnaires sent), and the actual sample sizes (number of questionnaires received) from January to December 2009.[2]

Table 4.1. Sample sizes and non-response rates.

Period	Planned sample size	Actual sample size	Non-response rate
08.01	1275	362	71.6
08.02	1272	339	73.3
08.03	3303	675	79.6
08.04	2218	530	76.1
08.05	1359	533	60.8
08.06	1353	562	58.5
08.07	1358	512	62.3
08.08	1348	484	64.1
08.09	1346	493	63.4
08.10	1486	431	71.0
08.11	1479	475	67.9
08.12	1479	443	70.0

Source: Authors' calculations.

[1] In all expressions, $\sum x_i$ is abbreviated notation for $\sum_{i \in S} x_i$, where *S* stands for appropriate sample strata. When elements are summed over another set, it is clearly stated.

[2] Non-response has been recorded by RIED only since 2008; previous non-response rates are not known.

The unit non-response rates are very high – in 2008, they oscillate from 58.5% to 79.6%. There may be different causes for such a visible non-response in RIED business tendency survey; experience shows that refusal is its most frequent reason. Two main problems related to non-response are the following:

- structure of the sample does not reflect the structure of the population,
- structure of the sample changes in time.

Table 4.2 shows the structure of the population of manufacturing firms with respect to size of employment according to The Central Statistical Office (CSO) of Poland (GUS [2008]).

Table 4.2. Structure of the population, Central Statistical Office

Employment strata	2007 population	2007 population, %
up to 9 persons	333,426	88.91
10-49	32,100	8.56
50-249	7864	2.10
250-999	1406	0.37
1000 and more	219	0.06

Source: Central Statistical Office of Poland.

Combining employment strata available from The Central Statistical Office data and employment strata available from RIED sample, we compute the following structure of the population and the sample[1] and present the results in Table 4.3, with detailed results provided in Appendix 2.

Table 4.3. Comparison of population and RIED sample structures

Employment level	2007 population, %	December 2008 sample, %
To 50 persons	97.47	43.3
51-250	2.10	31.6
Over 250	0.43	25.1

Source: Authors' calculations on the basis of Central Statistical Office data.

From Table 4.3 it is clear that the structure of the population differs significantly from the structure of the sample. Moreover, the structure of the sample also changes considerably in time. The share of industrial enterprises with employment level up to 50 persons oscillates from 36.9% to 57.2%; with employment level 51-250 persons – from 24.7% to 32.3%; and with employment level over 250 persons – from 17.7% to 32.3% (see Appendix 2). Such significant

[1] The few enterprises with undefined employment level are assumed by RIED to belong to the interval "up to 50 persons"; we follow this practice in our calculations.

disproportions should not be neglected, but have not been analyzed so far. More precisely, although the structure of the sample is different from the structure of the population and changes significantly in time (e.g. the largest industrial enterprises represent in one month 17.7% of the sample and in another month 32.3% of the sample, and their answers are additionally weighted by the factor of 5), original RIED balance statistics do not address this issue.

In order to account for variable structure of the sample and its divergence from population structure we propose a balance statistics, which for current situation is evaluated by survey respondents and is given as[1]:

$$
\begin{aligned}
{}_{t}BA_{t+k} = & \frac{\frac{N_1}{n_1}\sum w_1 x_{Ai} + \frac{N_2}{n_2}\sum w_1 x_{Ai} + \frac{N_3}{n_3}\sum w_3 x_{Ai}}{N_1 w_1 + N_2 w_2 + N_3 w_3} \cdot 100 - \\
& - \frac{\frac{N_1}{n_1}\sum w_1 y_{Ai} + \frac{N_2}{n_2}\sum w_1 y_{Ai} + \frac{N_3}{n_3}\sum w_3 y_{Ai}}{N_1 w_1 + N_2 w_2 + N_3 w_3} \cdot 100
\end{aligned}
\tag{4.1}
$$

and for expectations as:

$$
\begin{aligned}
{}_{t}BP_{t+k} = & \frac{\frac{N_1}{n_1}\sum w_1 x_{Pi} + \frac{N_2}{n_2}\sum w_1 x_{Pi} + \frac{N_3}{n_3}\sum w_3 x_{Pi}}{N_1 w_1 + N_2 w_2 + N_3 w_3} \cdot 100 - \\
& - \frac{\frac{N_1}{n_1}\sum w_1 y_{Pi} + \frac{N_2}{n_2}\sum w_1 y_{Pi} + \frac{N_3}{n_3}\sum w_3 y_{Pi}}{N_1 w_1 + N_2 w_2 + N_3 w_3} \cdot 100
\end{aligned}
\tag{4.2}
$$

Combining employment strata available from The Central Statistical Office data and employment strata available in RIED sample we obtain three pooled strata for employment intervals:

- up to 50 persons,
- 51–251 persons,
- over 250 persons,

for which 2007 population sizes are known, and are equal to: [2]

- $N_1 = 365526$,
- $N_2 = 7864$,
- $N_3 = 1625$,

[1] This form of balance statistics corresponds to post-stratification estimator (non-response weighting). Let us note that for proportional sample formulas (4.1) and (4.2) reduce to formulas (3.3) and (3.4) respectively.

[2] Empirical analysis described in sections 6 and 7 is based on data from January 2006 to January 2009; year 2007 was selected as a base year since population data for 2008 is not yet available.

and n_i, $i = 1, 2, 3$ stand for sample sizes in respective strata.

As far as weights w_i are concerned, we have the following alternatives:

- in case of no weighting, $w_i = 1$,
- in case of RIED weighting, we consider two variants of weights: $w_1 = 1$, $w_2 = 2$, $w_3 = 3$, and $w_1 = 1$, $w_2 = 2$, $w_3 = 4$,[1]
- in case of weighting by the lower limit of the employment level, $w_1 = 1$, $w_2 = 51$, and $w_3 = 251$.

5. Comparison of alternative balance statistics

Our empirical analysis focuses on question number 1, industrial production, for two reasons: first, production expectations influence numerous decisions of firms (among them, investment and employment levels); second, it has well-defined counterpart in official statistics which is necessary to employ quantification methods described below. Our dataset covers monthly data from January 2006 to January 2009 ($n = 37$).[2]

Let us introduce the following notation. On the basis of formulas (3.3) and (3.4) (that is, before consequences of non-response are taken into consideration), the following alternatives are analyzed:

A – no weighting (all respondents are weighted by 1),

B – RIED weighting (respondents are weighted by 1, 2, 3, 4 and 5 according to employment level),

C – weighting by lower limit of the employment interval (that is, by 1, 51, 251, 501 and 2001 respectively).

On the basis of formulas (4.1) and (4.2), when consequences of non-response addressed in this paper are taken into account, the following alternatives are analyzed:

D – no weighting (all respondents are weighted by 1),

E – RIED weighting (respondents are weighted by 1, 2, and 3, respectively),

F – RIED weighting (respondents are weighted by 1, 2, and 4, respectively),

G – weighting by lower limit of the employment interval (that is, by 1, 51, 251, 501 and 2001 respectively).

Current state balance statistics calculated on the basis of formulas (3.3) and (4.1) for different systems of weights of course differ one from another, as do balance statistics for expectations obtained from formulas (3.4) and (4.2). In Appendix 3, we present comparison of balances, both for current state and

[1] Intervals defined by RIED do not correspond to those defined by CSO and consequently unambiguous definition of weights in not possible. In order to make our analysis more general we consider two weighting schemes.

[2] For analysis of weighting and non-response patterns, access to individual-level data was necessary. Authors wish to thank employees of the Research Institute for Economic Development (RIED) at the Warsaw School of Economics for data pre-processing to permit empirical analysis without compromising confidentiality of survey information.

expectations. On the basis of the empirical results we arrive at the conclusion that balances calculated on the basis of formulas (4.1) and (4.2), which take into account variable structure of the sample, are less sensitive to the choice of weights as compared to balances calculated on the basis of formulas (3.3) and (3.4) where the structure of the sample is not taken into consideration. As far as evaluation of current state is concerned, mean of absolute values of differences between balances for any two alternatives and 37 time periods oscillate from 2.23 to 6.72 for balances calculated from (3.3), and from 0.21 to 5.28 for balances calculated from (4.1). Range of differences for 37 time periods oscillates from 6.27 to 24.89 for balances calculated from (3.3), and from 0.6 to 15.63 for balances calculated from (4.1). Turning to expectations series, mean of absolute values of differences between balances for any two alternatives and 37 time periods oscillates from 1.54 to 6.56 for balances calculated from (3.4), and from 0.18 to 3.64 for balances calculated from (4.2). Range of differences for 37 time periods oscillates from 4.25 to 16.74 for balances calculated from (3.4) and from 0.7 to 15.08 for balances calculated from (4.2); for details see Appendix 3.

Empirical results are consistent with our intuition that balance statistics which takes into account variable, and different from population, structure of the sample should be less sensitive to choice of weights than balance statistics obtained in case when structure of the sample is ignored.

6. Quantification of survey data

Balance statistic given by (3.2) is a very simple quantitative measure of qualitative expectations expressed in business surveys. More advanced options are offered by probabilistic and regressive quantification methods.[1] In this paper, two versions of regression method are employed: Anderson and Thomas' models. In Anderson's model, introduced in 1952, the following equation is estimated:

$$ {}_t x_{t+1} = \alpha \cdot {}_t A^1_{t+1} + \beta \cdot {}_t A^3_{t+1} + \nu_t , \tag{6.1} $$

where ${}_t x_{t+1}$ describes relative changes in value of variable x noted in official statistic between t and $t + 1$. Assuming that the same relationship holds true for expectations reported in surveys, and that error term in equation (6.1) meets standard OLS assumptions,[2] parameters α and β are estimated, and quantitative measure of expectations is constructed on the basis of the following equation:

$$ {}_t \hat{x}^e_{t+1} = \hat{\alpha} \cdot {}_t P^1_{t+3} + \hat{\beta} \cdot P^3_{t+3} , \tag{6.2} $$

[1] For a concise review of basic quantification methods and their modifications see Pesaran [1989].

[2] In practice, HAC standard errors are used to account for possible serial correlation and/or heteroskedasticity of the error term in (6.1) and (6.3).

where $\hat{\alpha}$ and $\hat{\beta}$ are OLS-estimators of (6.1) and reflect average change in variable ${}_t x_{t+1}$ for respondents expecting, respectively, increase and decrease of this variable. Let us note that expectations balance statistic (3.2) is a special case of Anderson's expectations series (6.2) for $\hat{\alpha} = -\hat{\beta} = 1$.

To tailor the (very general) Anderson's model for the case when normal or typical situation that respondents compare their current situation to include a certain growth rate, making downward corrections more essential than upward, D. G. Thomas, in 1995, offered a modification. To account for the asymmetry he proposes to estimate

$$ {}_t x_{t+1} = \gamma + \beta \cdot {}_t A^3_{t+1} + \xi_t , \tag{6.3}$$

where $\beta < 0$, and constant γ is interpreted as typical growth rate. Model described by (6.3) reflects assumption that behaviour of economic agents is dependent on growth rate of a variable (usually production or prices) that the enterprise typically observes, and limits the degree of multicollinearity which often emerges in Anderson's model. Thomas' quantitative measure of expectations is given by the formula

$$ {}_t \hat{x}^e_{t+1} = \hat{\gamma} + \hat{\beta} \cdot {}_t P^3_{t+3} , \tag{6.4}$$

where $\hat{\gamma}$ and $\hat{\beta}$ are estimates obtained on the basis of (6.3).

Quantification procedures described above apply also in cases when weighted data are used (see formula (3.3)) and non-response is taken into account (see formula (4.1)).

None of the quantification methods proposed in literature proved to be generally superior; their performance depends on several factors, including dynamics of forecasted variables and time horizon considered. Quantified expectations series do not provide a perfect reflection of true expectations because in addition to measurement errors present in all economic data series, they are also burdened with aggregation and quantification errors introduced by imperfect quantification procedures. Even though both models considered in this section are rather simple, they provide encouraging results, and often prove superior to balance statistics. In the next section, both methods are used to quantify expectations concerning industrial production.

7. Results of rationality tests

We use both Anderson (6.1) and Thomas' (6.3) methods to quantify expectations data provided by the RIED survey in cases **A** – **G**. All quantification models are estimated by OLS with HAC standard errors to account for possible serial correlation of the error term (due to inertia often observed in expectations

series) and heteroskedasticity (likely to result from learning patterns imbedded in expectations formation processes). Both methods provide very similar results for all seven expectations series considered; neither weighting scheme nor non-response issues considered in this paper seem to influence results of quantification procedures. All explanatory variables exhibit correct signs and are statistically significant at the 5% significance level; RESET test also does not allow to differentiate between two competing methods (for summary of estimation results, see Appendix 4). On the basis of slightly better measures of fit, Anderson's models are selected for further analysis.

To confirm similarity of results across expectations series, averages and standard deviations of quantitative expectations series derived for **A–G** are compared in Table 7.1.

Table 7.1. Basic descriptive statistics for expectations data series

	A	B	C	D	E	F	G
Average	0.0238	0.0241	0.0243	0.0223	0.0225	0.0225	0.0238
Standard deviation	0.0168	0.0174	0.0232	0.0157	0.0158	0.0158	0.0171

Source: Authors' calculations.

It is clear that Anderson's quantification model (as well as Thomas'; see Appendix 4) is not sensitive to weighting or non-response issues considered, and quantitative series obtained on its basis are likely to have similar properties. This finding seems favorable from practical point of view because properties of expectations series analyzed in this paper appear to be insensitive to modifications of survey weighting schemes and sample structure, and therefore more reliable.

Let us note that expectations series derived for case **C** exhibits noticeably higher variation than other series due to wide range of weights used (from 1 to 2001) combined with significant changes in sample structure that are not corrected by balance statistic formula. After taking non-response into consideration by applying formula (4.2), variation reduces to levels observed for other cases, as can be seen from **F** and **G**.

To verify if results of rationality tests depend on non-response or weighting schemes, all seven series **A–G** are submitted to unbiasedness test described by hypothesis (2.1). In each case, dependent variable $PP3_t$ is defined as currently observed relative change in industrial production as compared to three months ago; independent variable E_t represents expectations series calculated on the basis of expectations expressed three months earlier on the basis of Anderson's method. All models are estimated by OLS with HAC standard errors; estimation results are presented in Appendix 5, table A5.1. They provide evidence that all expectations series, independently from weighting systems and non-response issues considered, remain unbiased estimates of relative changes in production.

The final step in assessing whether results of rationality tests depend on abovementioned factors consists of orthogonality test defined by hypothesis (2.3). We define information set **Π** to include the following variables: *PP3* (relative change in industrial production as compared to three months ago); *AS* (current state balance statistic) and *PS* (expectations balance statistic) because all three variables are available to industrial enterprises at no additional costs. All variables are lagged two and three months to account, on one hand, for delay in availability of the data, and on the other hand for relatively short attention span that may be expected from managers who are not professional forecasters. All models are estimated with HAC standard errors; detailed results are presented in Appendix 5, table A6.2. They show that none of the expectations error series is orthogonal to the variables included in the information set, independently from weighting system and non-response issues considered; specifically, variables $PP3_{t-3}$ and PS_{t-2} are significant in all cases. This result suggests that industrial enterprises do not efficiently use information included in these series, and incorporating them in firms' information set could improve quality of their forecasts.

8. Concluding comments

We conclude that expectations concerning relative changes in industrial production expressed by Polish industrial enterprises in RIED business tendency surveys are unbiased but do not efficiently use all available information, namely, observed relative changes in production (lagged three months) and expectations balance (lagged two months).

These results remain in line with previous research on the subject. Prior tests of rationality of production expectations in Poland have given mixed results, but generally exhibited unbiasedness and lack of orthogonality with respect to lagged expectations balance statistics and observed changes in production (see Tomczyk [2001, 2004]). Review of rationality tests conducted on the basis of RIED data in 1997–2006 proved rationality an exception rather than a rule, and production expectations being relatively rational when compared to other variables (Tomczyk [2008 b]). Results of econometric tests of REH are confirmed by analyses of contingency tables: production expectations proved to be more rational, as measured by frequency with which they meet rationality condition imposed by the Gourieroux–Pradel theorem, than expectations on prices, employment, and general business situation (Tomczyk [2008 a]). Generally, expectations concerning changes in production proved to be either irrational but relatively precise when compared with other economic variables (in tests based on individual data aggregated in contingency tables), or unbiased and not orthogonal to freely available information (in standard econometric tests of REH). The latter result is confirmed in this paper.

The main objective of the paper was to test if properties of expectations series, particularly those pertaining to their rationality, are influenced by weighting

schemes and two major problems introduced by non-response, namely, the fact that structure of the sample does not reflect the structure of the population, and that it changes in time. We found that they are not; for every expectation series considered, results of standard rationality tests remain the same. Therefore, properties of expectations of Polish industrial enterprises are proved to be independent from modifications of survey weighting schemes and sample structure, and therefore more reliable.

Appendix 1. Monthly RIED questionnaire in industry

		Observed within last month	Expected for next 3–4 months
01	Level of production (value or physical units)	up unchanged down	will increase will remain unchanged will decrease
02	Level of orders	up normal down	will increase will remain normal will decrease
03	Level of export orders	up normal down not applicable	will increase will remain normal will decrease not applicable
04	Stocks of finished goods	up unchanged down	will increase will remain unchanged will decrease
05	Prices of goods produced	up unchanged down	will increase will remain unchanged will decrease
06	Level of employment	up unchanged down	will increase will remain unchanged will decrease
07	Financial standing	improved unchanged deteriorated	will improve will remain unchanged will deteriorate
08	General situation of the economy regardless of situation in your sector and enterprise	improved unchanged deteriorated	will improve will remain unchanged will deteriorate

Source: the RIED database.

Appendix 2. Sample structure with respect to combined RIED and CSO employment levels

Employment level	Period	Percentage	Period	Percentage	Period	Percentage
Up to 50 persons	0601	40.5	0701	48.5	0801	44.8
51-250	0601	32.8	0701	29.5	0801	32.3
Over 250	0601	26.7	0701	22.0	0801	22.9
Up to 50 persons	0602	46.5	0702	47.5	0802	47.5
51-250	0602	33.2	0702	31.2	0802	30.7
Over 250	0602	20.3	0702	21.4	0802	21.8
Up to 50 persons	0603	57.2	0703	46.8	0803	36.9
51-250	0603	24.7	0703	31.9	0803	30.8
Over 250	0603	18.1	0703	21.3	0803	32.3
Up to 50 persons	0604	50.3	0704	45.5	0804	51.7
51-250	0604	30.2	0704	31.2	0804	27.4
Over 250	0604	19.5	0704	23.3	0804	20.9
Up to 50 persons	0605	46.6	0705	44.0	0805	38.5
51-250	0605	31.6	0705	32.6	0805	31.3
Over 250	0605	21.8	0705	23.4	0805	30.2
Up to 50 persons	0606	52.2	0706	47.5	0806	40.4
51-250	0606	30.1	0706	31.7	0806	31.0
Over 250	0606	17.7	0706	20.9	0806	28.6
Up to 50 persons	0607	49.7	0707	47.4	0807	44.1
51-250	0607	30.4	0707	30.0	0807	30.3
Over 250	0607	19.9	0707	22.5	0807	25.6
Up to 50 persons	0608	48.6	0708	44.2	0808	44.4
51-250	0608	28.4	0708	33.2	0808	28.7
Over 250	0608	23.0	0708	22.6	0808	26.9
Up to 50 persons	0609	47.9	0709	47.1	0809	44.6
51-250	0609	29.9	0709	31.6	0809	27.6
Over 250	0609	22.2	0709	21.4	0809	27.8
Up to 50 persons	0610	47.1	0710	46.7	0810	42.0
51-250	0610	30.7	0710	30.9	0810	29.5
Over 250	0610	22.2	0710	22.4	0810	28.5
Up to 50 persons	0611	49.0	0711	47.4	0811	44.2
51-250	0611	30.1	0711	30.2	0811	26.5
Over 250	0611	20.9	0711	22.3	0811	29.3
Up to 50 persons	0612	45.2	0712	48.5	0812	43.3
51-250	0612	31.3	0712	30.1	0812	31.6
Over 250	0612	23.5	0712	21.4	0812	25.1

Source: Authors' calculations on the basis of RIED data.

Appendix 3. Comparison of alternative balance statistics

Table A3.1. Comparison of alternative balance statistics – current state[1]

	A-B	A-C	B-C	D-E	D-F	D-G	E-F	E-G	F-G
MAE	2.23	6.72	5.20	0.21	0.24	5.28	0.04	5.09	5.04
min	0.02	0.01	0.37	0.01	0.00	0.14	0.00	0.17	0.16
max	6.30	24.90	21.75	0.61	0.67	15.77	0.14	15.23	15.10
range	6.27	24.89	21.38	0.60	0.67	15.63	0.14	15.07	14.94

Source: Authors' calculations.

Table A3.2. Comparison of alternative balance statistics – expectations

	A-B	A-C	B-C	D-E	D-F	D-G	E-F	E-G	F-G
MAE	1.54	6.56	5.58	0.18	0.19	3.64	0.03	3.47	3.46
min	0.05	0.86	0.04	0.02	0.03	0.01	0.00	0.08	0.05
max	4.30	17.60	14.61	0.72	0.79	15.08	0.07	14.37	14.30
range	4.25	16.74	14.57	0.70	0.76	15.08	0.07	14.29	14.24

Source: Authors' calculations.

[1] MAE – mean of absolute values of differences between two balances for 37 periods; min – minimum difference between two balances for 37 periods; max – maximum difference between two balances for 37 periods; range = max – min.

Appendix 4. Quantification models: estimation results

Table A4.1. Anderson's model (6.1) with HAC standard errors

	A	B	C	D	E	F	G
A	0.0014	0.0014	0.0016	0.0014	0.0015	0.0015	0.0014
B	-0.0016	-0.0016	-0.0019	-0.0015	-0.0015	-0.0015	-0.0016
centered R^2	0.1048	0.1160	0.1784	0.0977	0.0985	0.0988	0.1089
AIC	-85.2921	-85.7601	-88.4678	-85.0023	-85.0302	-85.0428	-85.4593
RESET *p*-value	0.894	0.598	0.171	0.861	0.873	0.877	0.679

Table A4.2. Thomas' model (6.3) with HAC standard errors

	A	B	C	D	E	F	G
Γ	0.0669	0.0665	0.0669	0.0644	0.0648	0.0648	0.0664
B	-0.0027	-0.0027	-0.0028	-0.0024	-0.0025	-0.0025	-0.0027
R^2	0.1119	0.1207	0.1577	0.1100	0.1020	0.1022	0.1169
adjusted R^2	0.0865	0.0957	0.1337	0.0753	0.0764	0.0766	0.0916
AIC	-85.5852	-85.9576	-87.5477	-85.1343	-85.1764	-85.1845	-85.7929
RESET *p*-value	0.978	0.781	0.105	0.671	0.689	0.695	0.857

Appendix 5. Tests of rationality of expectations: estimation results

Table A5.1. Unbiasedness test (2.1) with HAC standard errors

	A	B	C	D	E	F	G
H_0 *p*-value	0.6163	0.5367	0.5684	0.7323	0.7244	0.7226	0.5796
α_0	-0.0385	-0.0413	-0.0300	-0.0320	-0.0325	-0.0326	-0.0398
α_1	2.0340	2.1270	1.6424	1.8826	1.8894	1.8909	2.0873
adjusted R^2	0.0974	0.1206	0.1298	0.0653	0.0671	0.0676	0.1091
Akaike IC	-63.9821	-64.8619	-65.2207	-62.7897	-62.8553	-62.8741	-64.4227
RESET *p*-value	0.408	0.316	0.172	0.391	0.397	0.397	0.344

Values on grey background are not statistically different from zero at 5% significance level.

Table A5.2. Orthogonality test (2.3) with HAC standard errors

	A	B	C	D	E	F	G
H_0 *p*-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
adjusted R^2	0.6774	0.6681	0.5464	0.6454	0.6491	0.6495	0.6828
Akaike IC	-98.048	-97.571	-87.972	-94.095	-94.503	-94.555	-98.863
RESET *p*-value	0.224	0.127	0.288	0.546	0.526	0.526	0.165

REFERENCES

ANDERSON O. [1952] *The business test of the IFO-Institute for Economic Research, Munich, and its theoretical model*, Revue de l'Institut International de Statistique 20:1–17.

GUS [2008] *Zmiany strukturalne grup podmiotów gospodarki narodowej w 2007 r. (Structural Change in National Economy in 2007)*, GUS (Central Statistical Office of Poland).

KOWALCZYK B., TOMCZYK E. [2008] *Rationality of expectations of industrial enterprises – analysis based on business tendency surveys with item non-response*, Bank i Kredyt 8:3–11.

LIU P. C., MADDALA G. S. [1992] *Rationality of survey data and tests for market efficiency in the foreign exchange markets*, Journal of International Money and Finance 11:366–381.

ŁYZIAK T. [2003] *Consumer inflation expectations in Poland*, European Central Bank Working Paper No. 287.

MUTH J. F. (1961) *Rational expectations and the theory of price movement*, Econometrica 29:315–335.

OSIŃSKA M. [2000] *Ekonometryczne modelowanie oczekiwań gospodarczych (Econometric Modeling of Economic Expectations)*, Wydawnictwo Uniwersytetu Mikołaja Kopernika w Toruniu.

PESARAN M. H. [1989] *The Limits to Rational Expectations*, Basil Blackwell, Oxford.

RIED [2008] *Business Survey*, May, Research Institute for Economic Development, Warsaw School of Economics.

SÄRNDAL C., SWENSSON B., WRETMAN J. [1992] *Model Assisted Survey Sampling*, Springer-Verlag, New York.

SHEFFRIN S. M. [1996] *Rational Expectations*, Cambridge University Press, Cambridge.

THOMAS D. G. [1995] *Output expectations within manufacturing industry*, Applied Economics 27:403–408.

TOMCZYK E. [2001] *Racjonalność oczekiwań respondentów testu koniunktury (Rationality of respondents of business tendency survey)*, in: E. Adamowicz (ed.) *Analiza tendencji rozwojowych w polskiej gospodarce na podstawie testu koniunktury (Analysis of Development Tendencies in Polish Economy on the Basis of Business Surveys)*, Prace i Materiały Instytutu Rozwoju Gospodarczego nr 70, Warsaw School of Economics.

TOMCZYK E. [2004] *Racjonalność oczekiwań. Metody i analiza danych jakościowych (Rationality of Expectations. Methods and Analysis of Qualitative Data)*, Monografie i Opracowania nr 529, Warsaw School of Economics.

TOMCZYK E. [2008 a] *Analysis of expectations on the meso scale: uses of contingency tables*, Warsaw School of Economics Research Grant No 03/S/0053/08.

TOMCZYK E. [2008 b] *Racjonalność oczekiwań polskich przedsiębiorców: reguła czy wyjątek? Wnioski z testu koniunktury IRG SGH (Rationality of expectations of Polish entrepreneurs: rule of exception? Summary of results for RIED business conditions tests), in: E. Adamowicz (ed.) Koniunktura gospodarcza – 20 lat doświadczeń Instytutu Rozwoju Gospodarczego SGH (Business Conditions Surveys – 20 Years of the Research Institute for Economic Development, Warsaw School of Economics)*, Warsaw School of Economics.

STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 423—426

REPORT

The 2nd International Workshop on Internet Survey Methods, Daejeon, South Korea, 8–9 September 2010

The 2nd International Workshop on Internet Survey Methods was held in Daejeon, South Korea, between 8–9 September 2010. The organizer and host of the workshop was Statistics Korea. The Workshop created a forum for international experts in the field of web surveys to share their knowledge and experiences on conducting the surveys and methods of improvement in the quality of survey outcomes. The theme of the 2nd International Workshop on Internet Survey Methods Workshop titled "Use of Internet Survey Methods in Official Statistics" was a continuation of issues undertaken on the 1st International Workshop on Internet Survey Methods organized on 9-10 September 2009 by Korea National Statistical Office. The Workshop took place, as the last year, at the International Conference Hall (the 2nd floor) of the Statistical Center located near the Government Complex where Statistics Korea is located (Statistical Complex).

Under the theme "Use of Internet Survey Methods in Official Statistics" the Workshop aimed to explore major concerns and challenges that national statistical offices are facing in the era of the Internet and show how they are coping with the changes in environment of conducting surveys. National case studies and researches on using Internet surveys in official statistics were also presented. In particular, the presentation of applications, best practice examples and innovative approaches were recommended to be discussed. At the same time, the Workshop touched the general issues of Internet surveys such as mixed-mode surveys and its measurement error, estimation for non-probability samples and technical issues of Internet surveys (questionnaire design and visualization).

At the Workshop fourteen statisticians from national statistical offices around the world and methodologists from private organizations and universities were invited as speakers – one from Australia, Germany, India, Poland and Sweden, two from Netherlands, three from South Korea and four from the United States of America. Besides, anyone who had a special interest in statistical or methodological issues could join this Workshop. Altogether there were one hundred thirty participants from different academic, research centers, statistical agencies, official survey institution and private organisations in South Korea and twelve participants from four countries (Iran, Mongolian, Nigeria and New Zealand).

The papers of invited speakers were published in the Workshop Proceedings and tentative discussion topics of the Workshop were addressing the following themes :

- (Special topic) National Case Studies on applying Internet Surveys in Official Statistics
- Mixed-mode Surveys: Coverage, Errors, Response Rates
- Web Survey Design: Questionnaire design and Visualization
- Estimation for Non-Probability Samples: Weighting Scheme for Voluntary Samples, Propensity Score Adjustment.

The workshop was officially inaugurated by its organizer - Insill Yi, the Commissioner of Statistics Korea. The welcoming speech was also given by Ki Jong Rhee, the President of The Korean Association for Survey Research. The workshop consisted of four sessions and one roundtable discussion. In each session, a special lecture addressed important emerging issues as well as general concerns of the each topic for an hour. Each presentation took 30 minutes and was followed by a 10-minute discussion.

The first session titled "Use of Internet Survey Methods in Official Statistics" explored national case studies and researches on how national statistical offices cope with the changes in environment of conducting surveys in the era of the Internet and how they use Internet surveys in their official statistics. The second session titled "Internet survey: relation to other modes and mixed mode studies" dealt with characteristics of Internet survey and the relationship with other survey modes. Mixed mode surveys' methodological issues such as response rates and mode effect were addressed. In the third session titled "Internet survey design", technical aspects of Internet survey methods including sampling, questionnaire design were introduced and discussed. The fourth session titled "Estimation for non-probability samples and its implication" took a close look at how to estimate non-probability samples of Internet surveys and what its implications are. Roundtable discussion touched the future role of Internet survey methods in official statistics and how national statistical offices need to accommodate them in the production of official statistics.

The official Workshop Agenda is given below and at the http://www.kostat.go.kr/iwis/.

Session I – Use of Internet survey Methods in Official Statistics (case studies)

Chair: Ki Jong Rhee (Professor, Kookmin University, South Korea).

Presentation:

1. **Merylin Henden** (Assistant Statistician, Australian Bureau of Statistics), Bruce Fraser – Use of Internet Survey in Official Statistics: A Case Study,
2. **Kwang sup KIM** (Director General of Statistics Korea, Statistics Korea) – Utilization of Internet survey in Statistics,
3. **Deirdre Giesen** (Senior Statistician, Statistics Netherland), Dirkjan Beukenhorst – Internet Use for Data Collection at Statistics Netherland.

Session II – Internet survey: relation to other modes and mixed mode studies

Chair: Marek Fuchs (Professor, Darmstadt University of Technology, Germany).

Special Lecture: **Frederick Conrad** (Professor, University of Michigan, USA) – Improving Measurement in Web Surveys.

Presentation:

1. **John Kennedy** (Professor, Indiana University, USA), Steven Tepper, Amber Lambert – An Analysis of Mode Effects in Three Survey Modes in the Strategic National Arts Alumni Project,
2. **Anders Holmberg** (Senior Methodologist, Statistics Sweden), Boris Lorenc – Using the Internet in Individual and Household Surveys, Summarizing Some Experiences at Statistics Sweden.

Session III – Internet Survey Design

Chair: Frederic Conrad (Professor, University of Michigan, USA).

Special Lecture: **Arie Kapteyn** (Director, Labor and Population, RAND Corporation, USA), Bas Weerman – Probability Based Web Panels.

Presentation:

1. **Marek Fuchs** (Professor, Darmstadt University of Technology, Germany) – Beyond question Wording: The Use of Visual Design and Multi-media Elements in Web Surveys,
2. **Suchismita Roy** (Research Fellow, Indian Statistical Institute, India) – Web Questionnaire and Paper Questionnaire in Sensitive Research: in the Eyes of Respondents,
3. **Elżbieta Getka-Wilczyńska** (Associate Professor, Warsaw School of Economics, Poland) – Modeling of the Sample in the Internet Mediated Research.

Session IV – Estimation for non-probability samples and its implication

Chair: Arie Kapteyn (Director, Labor and Population, RAND Corporation, USA).

Special Lecture: **Sunghee Lee** (Research Scientist, University of Michigan, USA) – Estimation for General Population Web Surveys and Its Implications.

Presentation:

1. **Sung Kyum Cho** (Prfessor, Chungam National University, South Korea) – Statistical Estimation from Internet Volunteer Samples,
2. **Stephanie Steinmetz** (Researcher, Erasmus University Rotterdam, Netherlands) – Volunteer Web Surveys and Propensity Score Adjustment – the Wage Indicator Example.

Roundtable Discussion – The future role of Internet Survey Methods in Official Statistics

Chair: **Merylin Henden** (Assistant Statistician, Australian Bureau of Statistics).

Panellists: **Sung Kyum** Cho (Professor, Chungam National University, South Korea), **Deirdre Giesen** (Senior Statistician, Statistics Netherlands), **Anders Holmberg** (Senior Methodologist, Statistics Sweden).

Moreover, a very interesting social programme, which included a welcoming dinner with the Deputy Commissioner of Statistic Korea and a tour to see the authentic Korean culture, was prepared for the invited participants by the organisers.

To summarise, the Workshop provided a great opportunity for all participants to learn about cutting-edge trends of Internet survey methods and to construct global networks with the professionals in this field. It gave also an excellent opportunity to explore the beauty of South Korea as well as Daejeon.

Prepared by Elżbieta Getka-Wilczyńska[1]

[1] Division of Mathematical Statistics Institute of Econometrics, Warsaw School of Economics, Warsaw, Poland, E-mail: Elzbieta.Getka-Wilczynska@sgh.waw.pl.