STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

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

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 183–185

FROM THE EDITOR

This issue is composed of three sections followed by two types of notes - one on the past conferences, the other on a future meeting of special importance due to its scale and character, as a part of the ongoing celebrations of the Year of Statistics. This is a kind of innovation worthwhile mentioning at the outset in order to let the persons responsible for organizing such a type of scientific meetings - and interested in popularization of it on our column - know about such a possibility, offered from now on. Another new thing to be implemented with the beginning of the next year is to increase the number of issues to four per year, instead of three so far. Much of the success of this planed innovation will depend on our key collaborators - potential authors and peer-reviewers, as this would imply the necessity to shorten an effective time of production of an issue.

This volume starts as usual with the section devoted to sampling methods and estimation. It is opened by **A. C. Onyeka's** paper ***A Class of Product-Type Exponential Estimators of The Population Mean in Simple Random Sampling Scheme***. It presents a class of product-type exponential estimators for estimating the population mean using known values of some population parameters of an auxiliary character under the simple random sampling without replacement (SRSWOR) scheme. A modified exponential estimator is proposed based on both the ratio-type and the product-type exponential estimators. Properties of the proposed estimators are obtained up to first order approximation. The modified exponential estimator under optimum conditions is shown to be more efficient than the simple sample mean and the ratio-type and product-type exponential estimators, as proven also by an empirical illustration.

The article ***The Class of Estimators of Finite Population Mean Using Incomplete Multi-Auxiliary Information*** by **Meenakshi Srivastava** and **Neha Garg** discusses a class of estimators for the mean of the finite population using available incomplete multi-auxiliary information. Some special cases of this class of estimators are considered and theoretical results are numerically supported. All the proposed estimators are more efficient for mean estimation. Authors conclude that the maximum use of available incomplete multi-auxiliary information can increase the efficiency of the estimators.

Turgay Ünalan and **H. Öztaş Ayhan** devote their paper on ***Probability Sample Selection Method in Household Surveys when Current Data on Regional Population is Unavailable*** to the problems with establishing the appropriate probability of selection of surveyed households when the quality sampling frame is not available. Such situations are taking place in developing countries where surveys planned for future periods long after the census date

cannot be representative. This creates under-coverage and bias for estimations. Therefore, population projections and data adjustment methodologies are proposed to provide a representative probability selection of the updated population. The method contains the correction on the differences of the sum of strata and aggregated values. Examples are also provided to demonstrate the potentials of the proposed methodology.

Janusz L. Wywiał's paper ***Sampling Designs Proportionate to Sum of Two Order Statistics of Auxiliary Variable*** discusses the case of a conditional sampling design proportional to the sum of two order statistics. Several strategies including the Horvitz-Thompson estimator and ratio-type estimators are checked for their accuracy and compared on the basis of computer simulation which allows one to expect the estimation strategies with the sampling design. In general, the accuracies of the considered ratio type strategies showed to be the best among all the strategies considered in the analysis. Some reservation to these conclusion is however advised given that the employed computer simulation was based on special data set and that a simulation analyses on a larger scale is still needed to establish such a type of generality.

Two research papers make up the *research articles* section. **Nicholas T. Longford** and **Ioana C. Salagean** in their paper ***The Effect of Unemployment Benefits on Labour Market Behaviour in Luxembourg*** estimate the effect of awarding unemployment benefits on gaining long-term employment after an unemployment spell and on the time it takes to achieve it. To this aim they apply the potential outcomes framework, and they conclude that such awards, regarded as a treatment, are associated with poorer labour force outcomes than no awards. The authors show that their conclusions are unequivocally negative about the benefit. Unemployment spells associated with benefits tend to be longer, even after matching on background.

In article ***Households' Saving Mobility in Poland*** **Barbara Liberda** and **Marek Pęczkowski** examine the mobility of Polish households with regard to saving rates during the years 2007-2010, and compare it with the households' saving mobility during the years 1997-2000 using panel data constructed from the Household Budget Surveys. The Markov mobility matrix was employed to estimate the long-term (ergodic) distribution of households according to the saving rates. The long-term households' distribution reveals a tendency towards polarization of households in terms of saving rates. However, the polarization showed to be asymmetrical towards the highest saving rate groups what allows the authors to conclude that it explains why Polish households could maintain rising savings during the highly uncertain period of the financial crisis in 2007-2010.

The main part of the *other articles* section constitutes **Andrzej Młodak's** paper ***Coherence and Comparability as Criteria of Quality Assessment in Business Statistics***. The problems of coherence and comparability exceed the classical notion of analysis of survey errors due to including the question of how results of two or more surveys can be used together and how the relevant data can

effectively be compared to obtain a better picture of social and economic phenomena over various aspects (e.g. space or time). This paper discusses characteristics of the main concepts of coherence and comparability as well as a description of differences and similarities between these two notions. Types of coherence and various aspects of perception of these notions in business statistics are analysed. Main sources of lack of coherence and comparability, factors affecting them (e.g. methodology, time, region, etc.) and methods of their measurement in context of information obtained from businesses are also presented.

The next paper, ***Statistics As a Profession - Statistician As an Occupation: Observations and Comments From a Panel of Experts*** by **Wlodzimierz Okrasa** and **Beata Witek** has a bit different character as being based on the results (observations and comments) of a specially organized panel to discuss the issue of the status and the role of statistics as a discipline along with the challenging issue of how to ensure that its followers, statisticians, meet the quality standards in the more and more demanding professional environment. Major panel's contributors were leading representatives of academia: Prof. Prof. Cz. Domanski, J. Dziechciarz, M. Krzyśko, M. Rocki and J. Wywiał. Contributions from highly competent audience are included as well.

Three notes on the past conferences, followed by above mentioned announcement on the upcoming conference devoted to the International Year of Statistics, conclude this volume.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 187

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, Summer 2013
Vol. 14, No. 2, pp. 189–200

A CLASS OF PRODUCT-TYPE EXPONENTIAL ESTIMATORS OF THE POPULATION MEAN IN SIMPLE RANDOM SAMPLING SCHEME

A. C. Onyeka[1]

ABSTRACT

The present study proposes a class of product-type exponential estimators for estimating the population mean of the study variable, using known values of some population parameters of an auxiliary character, under the simple random sampling without replacement (SRSWOR) scheme. Furthermore, the study also proposes a modified exponential estimator based on both the ratio-type and the product-type exponential estimators. Properties of the proposed estimators, under the SRSWOR scheme, are obtained up to first order approximation. The modified exponential estimator under optimum conditions is shown to be more efficient than the simple sample mean and the ratio-type and product-type exponential estimators. The theoretical results are supported by an empirical illustration.

Key words: ratio-type and product-type exponential estimators, auxiliary character, simple random sampling, mean square error. 2010 AMS Classification: 62D05.

1. Introduction

The use of information on auxiliary character to improve estimates of population parameters of the study variable is a common practice in sample surveys, especially when there is a strong linear relationship between the study and auxiliary variables. Several authors have made contributions in this regard, including Sukhatme and Sukhatme (1970) and Cochran (1977). Singh and Tailor (2005) suggested the use of known correlation coefficient between auxiliary characters for the estimation of finite population mean. Khoshnevisan et al. (2007) suggested a family of estimators of the population mean using known values of population parameters in simple random sampling. Tailor and Sharma (2009) used known coefficient of variation and coefficient of kurtosis of an

1 Department of Statistics, Federal University of Technology, PMB 1526, Owerri, Nigeria. E-mail: aloyonyeka@yahoo.com.

auxiliary character in estimating finite population mean of the study variable. Sharma and Tailor (2010) suggested a modified ratio-cum-dual to ratio estimator using known population mean of an auxiliary character. Onyeka (2012) used known values of population parameters of an auxiliary character to improve estimates of population mean in post-stratified random sampling. In using auxiliary information, different types of estimators have been considered, including the usual ratio-type, product-type and regression-type estimators. Recently, some authors have introduced dual and exponential estimators. Let y and x respectively denote the study and auxiliary characters in a finite population of N units; and let (y_i, x_i), i = 1, 2, … , n denote sample values of y and x in a random sample of n units drawn by simple random sampling without replacement (SRSWOR) method. In using known population mean $(\overline{X})$ of an auxiliary character x to improve estimates of the population mean $(\overline{Y})$ of the study variable y, Bahl and Tuteja (1991) introduced ratio and product type exponential estimators of the forms:

$$t_R = \overline{y}\exp\left[\frac{\overline{X} - \overline{x}}{\overline{X} + \overline{x}}\right] \tag{1.1}$$

and

$$t_P = \overline{y}\exp\left[\frac{\overline{x} - \overline{X}}{\overline{x} + \overline{X}}\right] \tag{1.2}$$

where $\overline{y}$ $(\overline{x})$ is the sample mean of the study (auxiliary) variable.

Singh and Vishwakama (2007) in their study used some modified exponential ratio-type and product-type estimators in estimating population mean in double sampling scheme. Singh et al. (2009a), following Kadilar and Cingi (2006) and Khoshnevisan et al. (2007), proposed a class of ratio-type exponential estimators of the population mean in simple random sampling, using known values of population parameters of an auxiliary character, of the form:

$$t_1 = \overline{y}\exp\left[\frac{(a\overline{X} + b) - (a\overline{x} + b)}{(a\overline{X} + b) + (a\overline{x} + b)}\right] = \overline{y}\exp\left[\frac{a(\overline{X} - \overline{x})}{a(\overline{X} + \overline{x}) + 2b}\right] \tag{1.3}$$

where $a\ (\neq 0)$, b are either real numbers or functions of known parameters of the auxiliary variable x such as coefficient of variation (C_x), coefficient of skewness $(\beta_1(x))$, coefficient of kurtosis $(\beta_2(x))$, standard deviation (σ) and correlation coefficient (ρ). It is worth mentioning that the choice of the values of a and b, in

practice, depends largely on the availability of known population parameters, which, admittedly, are not frequently known. The estimator in (1.3) is only useful when such population parameters are known, often from previous surveys. In sampling theory, preference, in terms of the known parameter to use, is usually given to those parameters that yield smaller variance or mean squared error when used to construct estimators. Again, notice that some of the population parameters like ρ and C_x are unitless. If such unitless quantities are used for the quantity b in (1.3), it implies that they are associated with the measurement unit of x and, in fact, they assume, temporarily, the same measurement unit of x. Ordinarily, this assumption would not cause any serious distortion in the value and measurement unit of the expression $a(\overline{X}+\overline{x})+2b$ in (1.3), if the value of b does not exceed unity. The estimator t_1 is biased for $\overline{Y}$ with mean square error, obtained up to first order approximation as:

$$\mathrm{MSE}(t_1)=\left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2+\tfrac{1}{4}\theta C_x^2(\theta-4K)\right] \tag{1.4}$$

where

$$\theta=\frac{a\overline{X}}{a\overline{X}+b}\ ,\ K=\frac{\rho C_y}{C_x}\ ,\ f=\frac{n}{N} \tag{1.5}$$

and $C_y(C_x)$ and ρ are the coefficient of variation of y(x), and the correlation coefficient, respectively. Singh et al. (2009b) considered some ratio-type exponential estimators in stratified random sampling. Shabbir and Gupta (2010) suggested exponential estimator for finite population mean in two phase sampling scheme when auxiliary variables are attributes. Grover and Kaur (2011) improved on the work carried out by Abd-Elfattah et al. (2010) by proposing ratio-type exponential estimators of finite population mean in simple random sampling scheme using an auxiliary attribute. Singh et al. (2011) suggested exponential ratio and exponential product type estimators for estimating unknown population variance, using information on two auxiliary characters.

In the present study, we have extended the work carried out by Singh et al. (2009a), by suggesting a class of product-type exponential estimators for estimating the finite population mean of the study variable, using known values of population parameters of an auxiliary variable. The present study also proposes a modified exponential estimator based on both the ratio-type and product-type exponential estimators, and compares the efficiency of the modified exponential estimator with those of the simple sample mean, the ratio-type exponential estimator and the product-type exponential estimator. Properties of the proposed estimators, especially the biases and mean squared errors of the estimators, are obtained up to first order approximations under the SRSWOR scheme.

2. The proposed product-type exponential estimator

Following Singh et al. (2009a), we propose a class of product-type exponential estimators of the population mean $\overline{Y}$, in SRSWOR scheme as

$$t_2 = \overline{y}\exp\left[\frac{(a\overline{x}+b)-(a\overline{X}+b)}{(a\overline{x}+b)+(a\overline{X}+b)}\right] = \overline{y}\exp\left[\frac{a(\overline{x}-\overline{X})}{a(\overline{x}+\overline{X})+2b}\right] \tag{2.1}$$

Let,

$$e_0 = \frac{\overline{y}-\overline{Y}}{\overline{Y}} \text{ and } e_1 = \frac{\overline{x}-\overline{X}}{\overline{X}}. \tag{2.2}$$

so that

$$E(e_0) = E(e_1) = 0 \tag{2.3}$$

$$E(e_0^2) = \frac{V(\overline{y})}{\overline{Y}^2} = \left(\frac{1-f}{n}\right)C_y^2 \tag{2.4}$$

$$E(e_1^2) = \frac{V(\overline{x})}{\overline{X}^2} = \left(\frac{1-f}{n}\right)C_x^2 \tag{2.5}$$

$$E(e_0e_1) = \frac{Cov(\overline{y},\overline{x})}{\overline{Y}\overline{X}} = \left(\frac{1-f}{n}\right)KC_x^2 \tag{2.6}$$

Rewriting (2.1) in terms of e_0 and e_1, and expanding up to first order approximation in expected value gives:

$$(t_2 - \overline{Y}) = \overline{Y}\,[e_0 + \tfrac{1}{2}\theta e_1 + \tfrac{1}{2}\theta e_0e_1 - \tfrac{1}{8}\theta^2 e_1^2\,] \tag{2.7}$$

and

$$(t_2 - \overline{Y})^2 = \overline{Y}^2\,[e_0^2 + \tfrac{1}{4}\theta^2 e_1^2 + \theta e_0e_1\,] \tag{2.8}$$

Taking expectation of (2.7) and (2.8), and using (2.3) – (2.6) to make the necessary substitutions gives the bias and mean square error of the proposed product-type exponential estimator t_2, respectively as:

$$B(t_2) = \left(\frac{1-f}{n}\right)\left(\frac{\overline{Y}}{8}\right)\theta(4K-\theta)C_x^2 \tag{2.9}$$

and

$$\mathrm{MSE}(t_2) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + \tfrac{1}{4}\theta C_x^2(\theta + 4K)\right] \tag{2.10}$$

3. Modified exponential estimators

Following authors like Cochran (1977) and Singh et al. (2009a), we propose a modified class of exponential estimators (t_3) as a linear function of the exponential estimators t_1 and t_2. We give the modified exponential estimator as:

$$t_3 = \alpha_1 t_1 + \alpha_2 t_2 \tag{3.1}$$

where α_1 and α_2 are weighting fractions such that $\alpha_1 + \alpha_2 = 1$. In practice, α_1 and α_2 should be chosen so as to minimize the mean square error of the estimator t_3. However, the value of α_1 is expected to be greater than $\frac{1}{2}$, that is, greater than the value of α_2 when there is a strong positive linear relationship between the study and auxiliary variates, since most ratio-type estimators are known to perform well under such a condition. Conversely, the value of α_2 is expected to be greater than $\frac{1}{2}$, or greater than the value of α_1 when there is a strong negative correlation between the study and auxiliary variates, since most product-type estimators are known to perform well under such a condition. Following Cochran (1977), the best (optimal) choices of α_1 and α_2 are respectively obtained as:

$$\alpha_1 = \frac{V_{22} - V_{12}}{V_{11} + V_{22} - 2V_{12}} \tag{3.2}$$

and

$$\alpha_2 = \frac{V_{11} - V_{12}}{V_{11} + V_{22} - 2V_{12}} \tag{3.3}$$

with the resultant optimum mean square error of t_3 given as:

$$\mathrm{MSE}_{\mathrm{opt}}(t_3) = \frac{V_{11}V_{22} - V_{12}^2}{V_{11} + V_{22} - 2V_{12}} \tag{3.4}$$

where V_{ii} is the mean square error t_i $(i = 1, 2)$ and V_{12} is the covariance term of the estimators t_1 and t_2. Rewriting (1.3) in terms of e_0 and e_1, and expanding up to first order approximation in expected value gives:

$$(t_1 - \overline{Y}) = \overline{Y}\,[e_0 - \tfrac{1}{2}\theta e_1 - \tfrac{1}{2}\theta e_0 e_1 + \tfrac{3}{8}\theta^2 e_1^2\,] \tag{3.5}$$

so that using (3.5) and (2.7), we obtain up to first order approximation in expected value:

$$(t_1 - \overline{Y})(t_2 - \overline{Y}) = \overline{Y}^2\,[e_0^2 - \tfrac{1}{4}\theta^2 e_1^2\,] \tag{3.6}$$

Taking expectation of (3.6) gives the covariance term of the estimators t_1 and t_2 as

$$\mathrm{Cov}(t_1, t_2) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 - \tfrac{1}{4}\theta^2 C_x^2\right) \tag{3.7}$$

Accordingly, it follows from (1.4), (2.10) and (3.7), that

$$V_{11} = \mathrm{MSE}(t_1) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + \tfrac{1}{4}\theta C_x^2(\theta - 4K)\right] \tag{3.8}$$

$$V_{22} = \mathrm{MSE}(t_2) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + \tfrac{1}{4}\theta C_x^2(\theta + 4K)\right] \tag{3.9}$$

$$V_{12} = \mathrm{Cov}(t_1, t_2) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 - \tfrac{1}{4}\theta^2 C_x^2\right) \tag{3.10}$$

Using (3.8) – (3.10) to make the necessary substitutions in (3.2) – (3.4) gives the best (optimal) choices of α_1 and α_2, together with the associated minimum mean square error of the modified exponential estimator t_3, as:

$$\alpha_1 = \frac{\theta + 2K}{2\theta} \tag{3.11}$$

$$\alpha_2 = \frac{\theta - 2K}{2\theta} \tag{3.12}$$

$$MSE_{opt}(t_3) = \left(\frac{1-f}{n}\right)\overline{Y}^2(1-\rho^2)C_y^2 \tag{3.13}$$

In practice, the weighting fractions α_1 and α_2 should be chosen very close to the expressions given in (3.11) and (3.12), respectively. We note that (3.13) is the same as the variance of the customary regression estimator. This indicates that the efficiency of the proposed modified exponential estimators may not be improved beyond that of the usual regression-type estimator.

4. Efficiency comparison

The variance of the simple sample mean $\bar{y}$ in SRSWOR scheme is given by

$$V(\bar{y}) = \left(\frac{1-f}{n}\right)\overline{Y}^2C_y^2 \tag{4.1}$$

Comparing (4.1) and (3.8), it follows that the ratio-type exponential estimator t_1, proposed by Singh et al. (2009a) would perform better than the sample mean in terms of having a smaller mean square error if

$$\frac{K}{\theta} = \left(\frac{\rho C_y}{C_x}\right)\left(\frac{a\overline{X}+b}{a\overline{X}}\right) > \frac{1}{4} \tag{4.2}$$

Notice that there is a slight mathematical error in the efficiency condition given by Singh et al. (2009a) with respect to the sign of the inequality of (5.2) of Singh et al. (2009a).

Comparing the product-type exponential estimator t_2 proposed in the present study with the sample mean $\bar{y}$, it follows from (4.1) and (3.9) that the proposed product-type exponential estimator t_2 would perform better than the sample mean in terms of having a smaller mean square error if

$$\frac{K}{\theta} = \left(\frac{\rho C_y}{C_x}\right)\left(\frac{a\overline{X}+b}{a\overline{X}}\right) < -\frac{1}{4} \tag{4.3}$$

To compare the efficiency of the proposed modified exponential estimator, t_3, with those of the estimators, $\bar{y}$, t_1 and t_2, we observe from (4.1), (3.8), (3.9) and (3.13), that:

Lemma I: The proposed modified exponential estimator t_3, under optimum conditions, (3.11) and (3.12), is more efficient than the sample mean $\bar{y}$ in terms of having a smaller mean square error, if

$$\rho^2 > 0 \text{ (which is always true)} \tag{4.4}$$

Lemma II: The proposed modified exponential estimator t_3, under optimum conditions, (3.11) and (3.12), is more efficient than the ratio-type exponential estimator t_1 proposed by Singh et al. (2009a) if

$$(\rho C_y - \tfrac{1}{2}\theta C_x)^2 > 0 \text{ (which is always true)} \tag{4.5}$$

Lemma III: The proposed modified exponential estimator t_3, under optimum conditions, (3.11) and (3.12), is more efficient than the product-type exponential estimator t_2 proposed in the present study if

$$(\rho C_y + \tfrac{1}{2}\theta C_x)^2 > 0 \text{ (which is always true)} \tag{4.6}$$

The implication of the above three lemmas is that once the weighting fractions α_1 and α_2 are chosen very close to their optimal values given in (3.11) and (3.12), respectively, then the proposed modified exponential estimator t_3 would be more efficient than the sample mean $\bar{y}$, the ratio-type exponential estimator t_1 proposed by Singh et al. (2009a), and the product-type exponential estimator t_2 proposed in the present study.

5. Numerical Illustration

For the purpose of empirical illustration of our theoretical results, consider the following five (5) members $t_j(i)$, $(j=1,2;\ i=1,2,3,4,5)$, each of the estimators t_1 and t_2.

Esti-mator (i)	$t_1(i)$	$t_2(i)$	a	b
1	$t_1(1) = \bar{y}\exp\left[\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right]$	$t_2(1) = \bar{y}\exp\left[\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right]$	1	1
2	$t_1(2) = \bar{y}\exp\left[\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}+2C_x}\right]$	$t_2(2) = \bar{y}\exp\left[\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}+2C_x}\right]$	1	C_x

Esti-mator (i)	$t_1(i)$	$t_2(i)$	a	b
3	$t_1(3) = \bar{y}\exp\left[\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x} + 2\rho}\right]$	$t_2(3) = \bar{y}\exp\left[\frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X} + 2\rho}\right]$	1	ρ
4	$t_1(4) = \bar{y}\exp\left[\frac{C_x(\bar{X} - \bar{x})}{C_x(\bar{X} + \bar{x}) + 2\rho}\right]$	$t_2(4) = \bar{y}\exp\left[\frac{C_x(\bar{x} - \bar{X})}{C_x(\bar{x} + \bar{X}) + 2\rho}\right]$	C_x	ρ
5	$t_1(5) = \bar{y}\exp\left[\frac{\rho(\bar{X} - \bar{x})}{\rho(\bar{X} + \bar{x}) + 2C_x}\right]$	$t_2(5) = \bar{y}\exp\left[\frac{\rho(\bar{x} - \bar{X})}{\rho(\bar{x} + \bar{X}) + 2C_x}\right]$	ρ	C_x

The mean square errors of the estimators $t_j(i)$, $(j = 1, 2;\ i = 1, 2, 3, 4, 5)$, are easily obtained using (3.8), (3.9) and (1.5). To illustrate the efficiency of the estimators, consider the following two sets of data given in Tailor et al. (2012).

Data I: Johnston, page 171

y = Percentage of hives affected by disease

x = Date of flowering of a particular summer species (number of days from January 1)

$N = 10$, $n = 4$, $\bar{Y} = 52$, $\bar{X} = 200$, $C_y = 0.1562$, $C_x = 0.04583$, $\rho = -0.94$

Using Data I, the computed percentage relative efficiencies of the estimators over the simple sample mean estimator $\bar{y}$ are displayed in Table 1.

Table 1. Percentage Relative Efficiency (PRE) over the Sample Mean $\bar{y}$

Estimators	$\bar{y}$	i					t_3
		1	2	3	4	5	
$t_1(i)$	100	77.18	77.09	76.99	74.96	77.08	859.11
$t_2(i)$	100	133.89	134.09	134.30	139.00	134.11	859.11

Data II: Johnston, page 171

y = Percentage of hives affected by disease

x = Mean January Temperature (^{0}C)

$N = 10$, $n = 4$, $\bar{Y} = 52$, $\bar{X} = 42$, $C_y = 0.1562$, $C_x = 0.13038$, $\rho = 0.80$

Using Data II, the computed percentage relative efficiencies of the estimators over the simple sample mean estimator $\bar{y}$ are displayed in Table 2.

Table 2. Percentage Relative Efficiency (PRE) over the Sample Mean $\bar{y}$

Estimators	$\bar{y}$	i					t_3
		1	2	3	4	5	
$t_1(i)$	100	194.57	197.08	195.14	181.83	196.98	277.78
$t_2(i)$	100	54.99	54.38	54.85	58.30	54.41	277.78

Tables 1 and 2 reveal that the ratio-type exponential estimator t_1 proposed by Singh et al. (2009a), and the product-type exponential estimator t_2 proposed in the present study, are not always or uniformly more efficient than the simple sample mean $\bar{y}$, except when the efficiency conditions (4.2) and (4.3) are respectively satisfied. Again, the numerical results in Tables 1 and 2 confirmed, as expected, that the proposed product-type exponential estimator t_2 is preferred over the ratio-type exponential estimator t_1 proposed by Singh et al. (2009a), when there is a strong negative correlation between the study and auxiliary characters, while the estimator t_1 is preferred over the estimator t_2 when there is a strong positive correlation between the study and auxiliary variables. The numerical results also confirmed that under the optimum conditions (3.11) and (3.12), the modified exponential estimator t_3, which is a linear function of the estimators t_1 and t_2, is more efficient than the sample mean $\bar{y}$ and the exponential estimators t_1 and t_2.

6. Conclusion

We have extended the work carried out by Singh et al. (2009a) by developing a class of product-type exponential estimators of the population mean in SRSWOR scheme, using known population parameters of an auxiliary character. The proposed class of product-type exponential estimators, under certain efficiency conditions, is shown to be more efficient than the usual sample mean estimator $\bar{y}$, in terms of having a mean square error smaller than the variance of $\bar{y}$. Also, numerical illustrations confirmed that when there is a strong negative correlation between the study and auxiliary variables, the proposed product-type exponential estimator is preferred over the ratio-type exponential estimator proposed by Singh et al. (2009a). Furthermore, in the present study we have developed a modified exponential estimator, which is a linear function or combination of both the ratio-type and product-type exponential estimators. By

using the optimal weighting fractions in the proposed modified exponential estimators, the modified exponential estimator is found to be more efficient than the usual sample mean estimator $\bar{y}$, the ratio-type exponential estimators t_1 proposed by Singh et al. (2009a), and the product-type exponential estimator t_2 proposed in the present study. In practice, therefore, we suggest that the weighting fractions in the modified exponential estimator be chosen very close to their optimal choices in (3.11) and (3.12) in order to realize the full benefits of using the proposed modified or improved exponential estimators of $\overline{Y}$ in SRSWOR scheme.

REFERENCES

ABD-ELFATTAH, A. M., EL-SHERPIENY, E. A., MOHAMED, S. M., ABDOU, O. F., (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. Applied Mathematics and Computation, Vol. 215, 4198–4202.

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

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

GROVER, L. K., KAUR, P., (2011). An improved exponential estimator of finite population mean in simple random sampling using an auxiliary attribute. Applied Mathematics and Computation, Vol. 218, No.7, 3093–3097.

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

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

ONYEKA, A. C., (2012). Estimation of population mean in post-stratified sampling using known value of some population parameter(s). Statistics in Transition-new series, 13(1), 65–78.

SHABBIR, J., GUPTA, S., (2010). Estimation of Finite Population Mean in Two Phase Sampling When Auxiliary Variables Are Attributes. Hacettepe Journal of Mathematics and Statistics, Vol. 39, No. 1, 121–129.

SHARMA, B., TAILOR, R., (2010). A New Ratio-Cum-Dual to Ratio Estimator of Finite Population Mean in Simple Random Sampling. Global Journal of Science Frontier Research, Vol. 10, Issue 1 (Ver. 1.0), 27–31.

SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F., (2009a). Improvement in estimating the population mean using exponential estimator in simple random sampling. Bulletin of Statistics & Economics, Vol. 3, A09, 13–18.

SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F., (2011). Improved exponential estimator for population variance using two auxiliary variables. Italian Journal of Pure and Applied Mathematics, No. 28, 101–108.

SINGH, R., KUMAR, M., CHAUDHARY, M. K., KADILAR, C., (2009b). Improved exponential estimator in stratified random sampling. Pakistan Journal of Statistics and Operations Research, Vol. V, No. 2, 67–82.

SINGH, H. P., TAILOR, R., (2005). Estimation of finite population mean using known correlation coefficient between auxiliary characters. Statistica, Anno LXV, 4, 407–418.

SINGH, H. P., VISHWAKAMA, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. Austrian Journal of Statistics, Vol. 36, No. 3, 217–225.

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

TAILOR, R., SHARMA, B. K., (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, Jul-09, Vol. 10, No. 1, 15–24.

TAILOR, R., TAILOR, R., PARMAR, R., KUMAR, M., (2012). Dual to ratio-cum-product estimator using known parameters of auxiliary variables. Journal of Reliability and Statistical Studies, Vol. 5, Issue 1, 65–71.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 201–216

THE CLASS OF ESTIMATORS OF FINITE POPULATION MEAN USING INCOMPLETE MULTI-AUXILIARY INFORMATION

Meenakshi Srivastava[1], **Neha Garg**[2]

ABSTRACT

In this paper, a class of estimators is considered for estimating the mean of the finite population utilizing available incomplete multi-auxiliary information. Some special cases of this class of estimators are considered. The approximate expressions for bias and mean square error of the suggested estimators have also been derived and theoretical results are numerically supported.

Key words: bias, mean square error, multi-auxiliary information.

1. Introduction

The use of auxiliary information for improving the precision of the estimators is well known when the variable under study Y and the auxiliary variable X are correlated. In large scale surveys, we often collect data on more than one auxiliary variable and some of these may be correlated with Y. Olkin (1958), Raj (1965), Rao and Mudholkar (1967), Srivastava (1971), Singh (1982), Agrawal and Panda (1993, 1994), Dalabehara and Sahoo (1997), Sahoo and Bala (2000), Abu-Dayyeh et. al. (2003), Ahmed (2004), Singh et. al. (2004), Kadilar and Cingi (2005), Perril (2007), Singh et. al. (2007), etc., considered some estimators which utilize information on several auxiliary variables which are correlated with the variable under study. In many situations, we may have information on several auxiliary variables but each variable may not be known for each population unit. Singh (1977) considered the concept of stratification for weighting the given incomplete auxiliary information. Srivastava and Garg (2009) suggested a class of estimators for estimating the finite population mean utilizing available incomplete multi-auxiliary information when frame is unknown for each stratum.

1 Department of Statistics, Institute of Social Sciences, Dr. B. R. Ambedkar University, Agra (U.P.), India. E-mail: msrivastava_iss@hotmail.com.
2 School of Sciences, Indira Gandhi National Open University, New Delhi, India. E-mail: neha1garg@gmail.com.

The aim of the present paper is to develop a general class of weighted estimators based on incomplete multi-auxiliary information as to show that it is always better to use additional auxiliary variables which are correlated with Y. The general class of weighted estimators has been constructed when frame is known for each stratum.

Generally, the stratification is done on the basis of heterogeneity in the population with respect to the study variable Y. But we view stratification in the other way. In our case, the heterogeneity of the population is considered with respect to the unequal number of auxiliary variables. We stratify the population in terms of the information provided by the p auxiliary variables. Thus, there will be a stratum for which no auxiliary information is available, p strata for which only one auxiliary variable out of p auxiliary variables is known. Similarly, there will be ${}^{p}C_{2}$ strata for which the two auxiliary variables are known, ${}^{p}C_{3}$ strata for which three auxiliary variables are known, and so on. Ultimately, we will have a stratum for which all the p auxiliary variables are known.

It is seen that the given method, considering stratification of population on the basis of unequal number of auxiliary variable is capable of giving more precise results than simple sample mean per unit.

2. Notations

Let us consider a finite population $U = (U_1, U_2, ..., U_N)$ of N identifiable units taking values on a study variable Y and p auxiliary variables $X_1, X_2, \ldots, X_p$, which are correlated with Y. Auxiliary variables $X_1, X_2, ..., X_p$ are known for total $M_1, M_2, \ldots, M_p$ units of the population, respectively. For the maximum utilization of available incomplete auxiliary information, the population is divided into different strata according to the known number of auxiliary variables and a random sample of n units is drawn from these groups with simple random sampling without replacement.

${}^{p}C_{j}$: Number of strata for which j auxiliary variables are known; $j = 0, 1, 2, \ldots, p$.

N : Population size

n : Sample size

N_0 : Size of the stratum for which no auxiliary variable is known

N_{ii} : Size of the stratum for which 1 auxiliary variable X_i is known; $i = 1,2, \ldots, p$

N_{ij} : Size of the stratum for which 2 auxiliary variable X_i and X_j are known; $i < j$; $i, j = 1, 2, \ldots, p$

N_{ijk} : Size of the stratum for which 3 auxiliary variable X_i, X_j and X_k are known; $i < j < k$; $i, j, k = 1,2, \ldots, p$

$N_{1,2,\ldots,p}$: Size of the strata for which all p auxiliary variable $X_1, X_2, \ldots, X_p$ is known; $i < j < k$; $i, j, k = 1,2,\ldots, p$

where $N_{ii} + N_{ij} + N_{ijk} + \ldots + N_{1,2,\ldots i\ldots,p} = M_i$

2^p : Total number of strata, i.e. $\sum_{i=0}^{p} {}^{p}C_i = 2^p$

N_i : Population size of the i^{th} stratum; i = 1, 2, ..., 2^p such that $\sum_{i=1}^{2^p} N_i = N$

n_i : Sample size of the i^{th} stratum; i = 1, 2, …, 2^p such that $\sum_{i=1}^{2^p} n_i = n$

Y_{ik} : Value of the k^{th} observation on variable under study in i^{th} stratum; i = 1, 2, …, ${}^{p}C_j$; j = 0, 1, 2, …, p; k =1, 2, ..., N_i

X_{ijk} : Value of the k^{th} observation on j^{th} auxiliary variable in i^{th} stratum i = 1, 2, …, ${}^{p}C_j$; j = 0, 1, 2, …, p; k =1, 2, ..., N_i

$\overline{Y}_i$: Population mean of the Y variable in i^{th} stratum

$\overline{y}_i$: Sample mean of the Y variable in i^{th} stratum

$\overline{X}_{ij}$: Population mean of the j^{th} auxiliary variable in i^{th} stratum

$\overline{x}_{ij}$: Sample mean of the j^{th} auxiliary variable in i^{th} stratum

S_i^2 : Population mean square error of *Y* variable in i^{th} stratum

S_{ij}^2 : Population mean square error of the j^{th} auxiliary variable in i^{th} stratum

C_i^2 : Coefficient of variation of the variable under study Y in i^{th} stratum, i.e. $C_i^2 = \frac{S_i^2}{\overline{Y}_i^2}$

C_{ij}^2 : Coefficient of variation of the j^{th} auxiliary variable in i^{th} stratum, i.e. $C_{ij}^2 = \frac{S_{ij}^2}{\overline{X}_{ij}^2}$

ρ_{ij} : Correlation coefficient between the variables Y, variable under study and X_{ij}, the j^{th} auxiliary variable in the i^{th} stratum

ρ_{ijh} : Correlation coefficient between the variables X_j and X_h ($j \neq h$) in the i^{th} stratum.

b_{ij} : Regression coefficient of the variables Y, variable under study and X_{ij}, the j^{th} auxiliary variable in the i^{th} stratum

W_i : Proportion of units in the i^{th} stratum, i.e. $W_i = \dfrac{N_i}{N}$

f : Sampling fraction, i.e. $f = \dfrac{n}{N}$

f_i : Sampling fraction in the i^{th} stratum, i.e. $f_i = \dfrac{n_i}{N_i}$

α_{ij} : Weights, attached to the j^{th} auxiliary variable in the i^{th} stratum adding up to

$$K_{ij} = \rho_{ij} \frac{C_i}{C_{ij}}; \quad j=1,2, \ldots, p$$

The following figure shows the construction of strata ($2^3 = 8$) on the basis of available incomplete three-auxiliary information (X_1, X_2 and X_3):

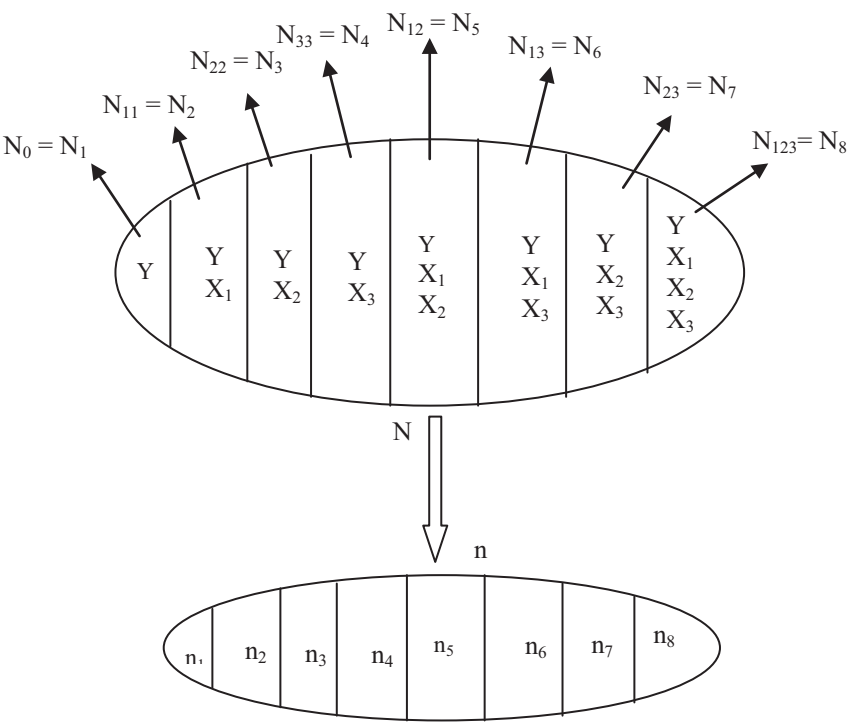


Figure 2.1. Construction of strata.

3. Suggested general class of estimators

Now, we can suggest the following general class of estimators using incomplete multi-auxiliary information.

$$\bar{y}_{pr} = \sum_{i=1}^{2^p} W_i g_i(y_i, x_i) \tag{1}$$

where $g_i(y_i, x_i) = \sum_{j=1}^{p} \alpha_{ij} g_{ij}(y_i, x_{ij})$ and $g_{11}(y_1, x_{11}) = \bar{y}_1$

$g_{ij}(y_i, x_{ij})$ is the function of $y_i = (y_{ik}; k = 1,2,..., N_i)$ and $x_{ij} = (x_{ijk}; k = 1,2,..., N_i)$

The bias and mean squared error of $\bar{y}_p$ are as follows

$$E(\bar{y}_{pr}) = \sum_{i=1}^{2^p} W_i E g_i(y_i, x_i) = \sum_{i=1}^{2^p} W_i \sum_{j=1}^{p} \alpha_{ij} E g_{ij}(y_i, x_{ij}) \tag{2}$$

$$MSE(\bar{y}_{pr}) = \sum_{i=1}^{2^p} W_i^2 MSE(g_i(y_i, x_i)) \tag{3}$$

$MSE(g_i(y_i, x_i))$ can easily be obtained for different values of the function g by generalizing the procedure used by Olkin (1958). Thus,

$$MSE(g_i(y_i, x_i)) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \sum_{j=1}^{p} \sum_{h=1}^{p} \alpha_{ij} \alpha_{ih} v_{ijh} \tag{4}$$

where $\left(\frac{1}{n_i} - \frac{1}{N_i}\right) v_{ijh} = Cov(g_{ij}, g_{ih})$

In matrix notation,

$$MSE(g_i(y_i, x_i)) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \underset{\sim}{\alpha_i} V_i \underset{\sim}{\alpha_i}'$$

where the matrix $V_i=(v_{ijh})$ and $\underset{\sim}{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip})$, $\underset{\sim}{\alpha_i}'$ being the transpose of $\underset{\sim}{\alpha_i}$.

3.1. Optimum values of α_{ij} for j = 1, 2, …, p

It is fairly simple to establish that the optimum α_{ij} is given by

$$\alpha_{ij} = \frac{\text{Sum of the elements of the } j^{th} \text{ column of } V_i^{-1}}{\text{Sum of all the } p^2 \text{ elements in } V_i^{-1}}$$

where V_i^{-1} is the matrix inverse to V_i. using the optimum weights, the mean square error is found to be

$$\text{MSE}\left(g_i\left(y_i, x_i\right)\right) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \Big/ \text{Sum of all the } p^2 \text{ elements in } V_i^{-1}$$

4. Some special cases of the suggested class of estimators

Case I. When each X_j (j = 1, 2, …, p) is positively correlated with Y in each stratum, our estimator will convert into weighted ratio estimator given as

$$\bar{y}_{pr.rat} = \sum_{i=1}^{2^p} W_i \sum_{j=1}^{p} \alpha_{ij} g_{ij.rat}\left(y_i, x_{ij}\right) \tag{5}$$

where $g_{ij.rat}(y_i, x_{ij}) = \dfrac{\bar{y}_i}{\bar{x}_{ij}} \bar{X}_{ij}$

$$\text{Bias}\left(\bar{y}_{pr.rat}\right) = E\left(\bar{y}_{pr.rat}\right) - \bar{Y}$$

$$= \sum_{i=1}^{2^p} W_i \frac{1-f_i}{n_i} \left[\bar{Y}_i \sum_{j=1}^{p} \alpha_{ij} \left\{C_{ij}^2\left(1 - K_{ij}\right)\right\}\right] \tag{6}$$

$$\text{MSE}\left(\bar{y}_{pr.rat}\right) = \sum_{i=1}^{2^p} W_i^2 \bar{Y}_i^2 \frac{1-f_i}{n_i} \left[C_i^2 + \sum_{j=1}^{p}\sum_{h=1}^{p} \alpha_{ij}\alpha_{ih}\left(\rho_{ijh} C_{ij} C_{ih} - \rho_{ij} C_i C_{ij} - \rho_{ih} C_i C_{ih}\right)\right] \tag{7}$$

Case II. When each X_j (j = 1, 2,……, p) is positively correlated with Y in each stratum, we can also use weighted regression estimator given as

$$\bar{y}_{pr.reg} = \sum_{i=1}^{2^p} W_i \sum_{j=1}^{p} \alpha_{ij} g_{ij.reg}\left(y_i, x_{ij}\right) \tag{8}$$

where $g_{ij.reg}(y_i, x_{ij}) = \bar{y}_i + b_{ij}\left(\bar{X}_{ij} - \bar{x}_{ij}\right)$

$$\because \text{Bias}\left(\bar{y}_{\text{pr.reg}}\right) = E\left(\bar{y}_{\text{pr.reg}}\right) - \bar{Y} = 0 \tag{9}$$

$$\text{MSE}\left(\bar{y}_{\text{pr.reg}}\right) = \sum_{i=1}^{2^p} W_i^2 \frac{1-f_i}{n_i}\left[S_i^2 + \sum_{j=1}^{p}\sum_{h=1}^{p}\alpha_{ij}\alpha_{ih}\left(b_{ij}b_{ih}\rho_{ijh}S_{ij}S_{ih} - b_{ij}\rho_{ij}S_iS_{ij} - b_{ih}\rho_{ih}S_iS_{ih}\right)\right] \tag{10}$$

Case III. If each X_j (j = 1, 2,……, p) is negatively correlated with Y in each stratum then our estimator will convert into weighted product estimator given as

$$\bar{y}_{\text{pr.prod}} = \sum_{i=1}^{2^p} W_i \sum_{j=1}^{p}\alpha_{ij} g_{ij.\text{prod}}\left(y_i, x_{ij}\right) \tag{11}$$

where $g_{ij.\text{prod}}\left(y_i, x_{ij}\right) = \frac{\bar{y}_i}{\bar{X}_{ij}}\bar{x}_{ij}$

$$\text{Bias}\left(\bar{y}_{\text{pr.prod}}\right) = \sum_{i=1}^{2^p} W_i \frac{1-f_i}{n_i}\left[\bar{Y}_i \sum_{j=1}^{p}\alpha_{ij}C_{ij}^2 K_{ij}\right] \tag{12}$$

$$\text{MSE}\left(\bar{y}_{\text{pr.prod}}\right) = \sum_{i=1}^{2^p} W_i^2 \bar{Y}_i^2 \frac{1-f_i}{n_i}\left[C_i^2 + \sum_{j=1}^{p}\sum_{h=1}^{p}\alpha_{ij}\alpha_{ih}\left(\rho_{ijh}C_{ij}C_{ih} + \rho_{ij}C_iC_{ij} + \rho_{ih}C_iC_{ih}\right)\right] \tag{13}$$

5. Empirical study

The above theoretical developments are applied on two artificially constructed populations. In data set one, the population consists of N = 70 observations. For example, the yield rates of the crop can be taken as a study variable Y and no of shoots / canes, average height of shoots / cane and average width of the 3rd leaf can be considered as three auxiliary variables X_1, X_2 and X_3, respectively. For the purpose of computation, a sample of size n = 25 has been drawn by SRSWOR from the first population and a sample of n = 43 units has been drawn by SRSWOR from the second data set of size N=90. Both the populations are given in the Appendix.

In both the data sets, the information on all the three auxiliary variables was missing for different population units and hence there will be 8 strata. The computed sample size for each stratum by proportional allocation and the

observed statistics about the population in case of incomplete auxiliary information are given in the following tables.

Data Set I

Table 1. Population parameters

Without Stratification	$\overline{Y}$ = 33.95903, S_y^2 = 4.543188
Stratum	
I	$N_1 = 5$, $n_1 = 2$, $\overline{Y}_1$=27.90000 S_1^2=46.70625
II	$N_2 = 9$, $n_2 = 3$ $\overline{Y}_2$=36.27778 $\overline{X}_{21}$=63.66667 K_{21}=0.77911 S_2^2=124.50694 ρ_{21}=0.79522 b_{21}=0.44394 S_{21}^2=399.50000
III	$N_3 = 13$, $n_3 = 5$ $\overline{Y}_3$=37.63846 $\overline{X}_{32}$=1.13538 K_{32}=1.25426 S_3^2=206.57090 ρ_{32}=0.70908 b_{32}=41.57907 S_{32}^2=0.06008
IV	$N_4 = 6$, $n_4 = 2$ $\overline{Y}_4$=43.25000 $\overline{X}_{43}$=3.70000 K_{43}=2.25222 S_4^2=181.87500 ρ_{43}=0.96175 b_{43}=26.32663 S_{43}^2=0.24272
V	$N_5 = 11$, $n_5 = 4$ $\overline{Y}_5$=26.49382 $\overline{X}_{51}$=55.00000 $\overline{X}_{52}$=1.11455 S_5^2=88.52406 ρ_{51}=0.85455 ρ_{512}=0.65101 S_{51}^2=602.00000 ρ_{52}=0.60681 K_{51}=0.68028 S_{52}^2=0.12349 K_{52}=0.68348 b_{51}=0.32770 b_{52}=16.24694
VI	$N_6 = 6$, $n_6 = 2$ $\overline{Y}_6$=45.83333 $\overline{X}_{61}$=87.83333 $\overline{X}_{63}$=3.75333

	S_6^2=32.96667	ρ_{61}=0.79500	ρ_{613}=0.63680
	S_{61}^2=236.96667	ρ_{63}=0.83236	K_{61}=0.56825
	S_{63}^2=0.05499	K_{63}=1.66899	b_{61}=0.29653
	b_{63}=20.38070		
VII	N_7 = 3, n_7 = 1		
	$\overline{Y}_7$=20.50000	$\overline{X}_{72}$=0.97000	$\overline{X}_{73}$=3.14667
	S_7^2=12.25000	ρ_{72}=0.53995	ρ_{723}=0.96961
	S_{72}^2=0.03430	ρ_{73}=0.31761	K_{72}=0.48283
	S_{73}^2=0.27693	K_{73}=0.32425	b_{72}=10.20408
	b_{73}=2.11242		
VIII	N_8=17, n_8 = 6		
	$\overline{Y}_8$=33.99412	$\overline{X}_{81}$=60.82353	$\overline{X}_{82}$=1.18529
	$\overline{X}_{83}$=3.65176	S_8^2=238.57309	ρ_{81}= 0.82364
	ρ_{812}=0.63381	ρ_{813}=0.14220	S_{81}^2=644.40441
	ρ_{82}=0.54901	ρ_{832}=0.52914	ρ_{83}=0.22148
	S_{82}^2=0.04304	K_{81}=0.89667	K_{82}=1.42522
	K_{83}=0.87872	S_{83}^2=0.17490	b_{81}=0.50115
	b_{82}=40.87506	b_{83}=8.17996	

Data Set II

Table 2. The Population parameters

Without Stratification	$\overline{Y}$ = 48.14444,		
	S_y^2 = 228.01261,		
I Stratum	N_1 = 16, n_1 = 8,		
	$\overline{Y}_1$= 45.81250,	S_1^2=207.22917	
II Stratum	N_2 = 13, n_2 = 6		
	$\overline{Y}_2$=48.53846	$\overline{X}_{21}$=51.69231	K_{21}=0.57466
	S_2^2=284.26923	ρ_{21}=0.85959	b_{21}=0.53960
	S_{21}^2 = 721.39744		

III Stratum	$N_3 = 7,\ n_3 = 3$ $\overline{Y}_3 = 47.85714$ $\overline{X}_{32} = 53.00000$ $K_{32}=0.63427$ $S_3^2 = 285.14286$ $\rho_{32} = 0.83425$ $b_{32}=0.57273$ $S_{32}^2 = 605.00000$
IV Stratum	$N_4 = 10,\ n_4 = 5$ $\overline{Y}_4 = 50.30000$ $\overline{X}_{43} = 46.60000$ $K_{43}=0.64864$ $S_4^2 = 312.23333$ $\rho_{43} = 0.93639$ $b_{43}=0.70014$ $S_{43}^2 = 558.48889$
V Stratum	$N_5 = 12,\ n_5 = 6$ $\overline{Y}_5 = 50.16667$ $\overline{X}_{51} = 56.50000$ $\overline{X}_{52} = 46.41667$ $S_5^2 = 191.60606$ $\rho_{51} = 0.85239$ $\rho_{512} = 0.54232$ $S_{51}^2 = 619.72727$ $\rho_{52} = 0.65961$ $K_{51}=0.53380$ $S_{52}^2 = 247.53788$ $K_{52}=0.53694$ $b_{51}=0.47396$ $b_{52}= 0.58032$
VI Stratum	$N_6 = 8,\ n_6 = 4$ $\overline{Y}_6 = 41.00000$ $\overline{X}_{61} = 46.62500$ $\overline{X}_{63} = 39.87500$ $S_6^2 = 208.00000$ $\rho_{61} = 0.77808$ $\rho_{613} = 0.71440$ $S_{61}^2 = 399.98214$ $\rho_{63} = 0.91753$ $K_{61}=0.63808$ $S_{63}^2 = 496.98214$ $K_{63}=0.57730$ $b_{61}=0.56110$ $b_{63}=0.59358$
VII Stratum	$N_7 = 9,\ n_7 = 4$ $\overline{Y}_7 = 47.88889$ $\overline{X}_{72} = 50.77778$ $\overline{X}_{73} = 44.00000$ $S_7^2 = 252.61111$ $\rho_{72} = 0.86535$ $\rho_{723} = 0.87401$ $S_{72}^2 = 647.94444$ $\rho_{73} = 0.92570$ $K_{72}=0.57291$ $S_{73}^2 = 384.50000$ $K_{73}=0.68939$ $b_{72}=0.54032$ $b_{73}=0.75033$
VIII Stratum	$N_8 = 15,\ n_8 = 7$ $\overline{Y}_8 = 51.33333$ $\overline{X}_{81} = 50.66667$ $\overline{X}_{82} = 51.33333$ $\overline{X}_{83} = 27.40000$ $S_8^2 = 208.66667$ $\rho_{81} = 0.89878$ $\rho_{812} = 0.58511$ $\rho_{813} = 0.56521$ $S_{81}^2 = 619.66667$

	$\rho_{82} = 0.71264$	$\rho_{832} = 0.60341$	$\rho_{83} = 0.56918$
	$S_{82}^2 = 361.80952$	K_{81}=0.51478	K_{82}=0.54120
	K_{83}=0.53627	$S_{83}^2 = 66.97143$	b_{81}=0.52156
	b_{82}=0.54120	b_{83}=1.00469	

The above data set is used to compute the biases and mean square errors of the estimators as discussed in section 4 and 5 and these estimators are compared (i) with each other in accordance with their MSE values (ii) with respect to mean per unit.

Table 3. The Biases, Mean Square Errors and the Relative Efficiencies of data set I

Estimators	*Bias*	*MSE*	*Relative Efficiency*
$\bar{y}$	-	4.54319	100
$\bar{y}_{pr.rat}$	0.6732	1.49930	303.02
$\bar{y}_{pr.reg}$	-	1.31585	345.27

Table 4. The Biases, Mean Square Errors and the Relative Efficiencies of data set II

Estimators	*Bias*	*MSE*	*Relative Efficiency*
$\bar{y}$	-	2.76915	100
$\bar{y}_{pr.rat}$	0.36221	1.48509	186.46
$\bar{y}_{pr.reg}$	-	0.90499	305.99

6. Discussion and conclusion

The proposed class of estimators using incomplete multi-auxiliary information has been compared with simple mean per unit estimator in which auxiliary information has not been used. It is seen that all the proposed estimators are more efficient for mean estimation for both the data set taken.

(i) A critical review of table 3 and 4 reveals that, though the ratio estimator of suggested class is biased ($\bar{y}_{pr.reg}$ is unbiased because we have considered it for pre-assigned value of b_{ij}), the amount of bias is almost negligible for $\bar{y}_{pr.rat}$.

(ii) When we compare the V ($\bar{y}$) with both of the proposed estimators, we find that V($\bar{y}$) = 4.54319, which is considerably higher than the MSE ($\bar{y}_{pr.rat}$) = 1.49930 and MSE ($\bar{y}_{pr.reg}$) = 1.31585 for data set I and V($\bar{y}$) = 2.76915, which is also higher than the MSE ($\bar{y}_{pr.rat}$) = 1.48509 and MSE ($\bar{y}_{pr.reg}$) = 0.90499 for data set II.

(iii) The trend becomes more clear when we compare the % relative efficiency from table 3 and 4, i.e. the gain in % relative efficiency of the estimators of suggested class is substantially higher as compared to usual per unit estimator.

It is evident from the above results that the proposed estimators establish the supremacy over $\bar{y}$ in the estimation of mean using stratification on the basis of available incomplete auxiliary information. Thus, we see that the maximum use of available incomplete multi-auxiliary information can increase the efficiency of the estimators.

REFERENCES

ABU-DAYYEH, W. A., AHMED, M. S., AHMED, R. A., MUTTLAK, H. A., (2003). Some Estimators of a Finite Population Mean Using Auxiliary Information, Appl. Math. Comput., 139, 287–298.

AGRAWAL, M. C., PANDA, K. B., (1993). Multivariate Product Estimators, Jour. Ind. Soc. Agri. Stat., 45 (3), 359–371.

AGRAWAL, M. C., PANDA, K. B., (1994). On Multivariate Ratio Estimation, Jour. Ind. Stat. Assoc., 32, 103–110.

AHMED, M. S., (2004). Some Estimators for a Finite Population Mean Under Two-Stage Sampling Using Multivariate Auxiliary Information, Appl. Math. Comput., 153 (2), 505–511.

COCHRAN, W. G., (1977). Sampling Technique, 3rd edition, Wiley and Sons, New York.

DALABEHARA, M., SAHOO, L. N., (1997). A Class of Estimators in Stratified Sampling with Two Auxiliary Variables, Jour. Ind. Soc. Agri. Stat., 50 (2), 144–149.

KADILAR, C., CINGI, H., (2005). A New Estimator Using Two Auxiliary Variables, Appl. Math. Comput., 162, 901–908.

OLKIN, I., (1958). Multivariate Ratio Estimation for Finite Population, Biometrika, 45, 154–165.

PERRIL, P. F., (2007). Improved Ratio-Cum-Product Type Estimators, Statistics in Transition, 8 (1), 51–69.

RAJ, D., (1965). On a Method of Using Multi-Auxiliary Information in Sample Survey, Jour. Amer. Stat. Assoc., 60, 270–277.

RAO, P. S. R. S., MUDHOLKAR, G. S., (1967). Generalized Multivariate Estimator for the Mean of Finite Population, Jour. Amer. Stat. Assoc., 62, 1009–1012.

SAHOO, J., BALA, M. K., (2000). A Note on the Estimation of the Population Mean in Stratified Random Sampling Using Two Auxiliary Variables, Biom. J., 42 (1), 87–92.

SUKHATME, P. V., SUKHATME, B. V., (1997). Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, Iowa, USA.

SINGH, H. P., UPADHYAYA, L. N., CHANDRA, P., (2004). A General Family of Estimators for Estimating Population Mean Using Two Auxiliary Variables in Two-Phase Sampling, Statistics in Transition, Dec., 6 (7), 1055–1077.

SINGH, R., (1977). A Note on the Use of Incomplete Multi-Auxiliary Information in Sample Surveys, Aust. J. Stat., 19 (2), 105–107.

SINGH, R., CHAUHAN, P., SAWAN, N., (2007). On the Bias Reduction in Linear Variety of Alternative to Ratio-Cum-Product Estimator, Statistics in Transition, Aug., 8 (2), 293–300.

SINGH, M., (1982). A Note on the Use of Multiauxiliary Information, Commun. Statist. Theo-Meth., 11 (8), 933–939.

SRIVASTAVA, M., GARG, N., (2009). A Class of Estimators of Finite Population Mean Using Incomplete Multi-Auxiliary Information when Frame is Unknown. Aligarh J. Statist., 29, 51–73.

SRIVASTAVA, S. K., (1971). A Generalized Estimator for the Mean of a Finite Population Using Multi-Auxiliary Information, Jour. Amer. Stat. Assoc., June, 66 (334), 404–407.

APPENDIX

Data Set I

Stratum I
Y 28.8 32.3 34.0 28.0 16.5

Stratum II
Y 50 29 26.5 46 25.5 38 45 46.5 20
X_1 87 56 50 56 36 60 96 82 50

Stratum III
Y 39 26.5 13 66 53.5 58 28.5 24.7 36.3 36.4 34.5 34.4 38.5
X_2 1.01 1.02 0.61 1.35 1.55 1.29 1.21 1.04 0.87 1.12 1.10 1.43
1.16

Stratum IV
Y 16.5 43.5 48 52 47.5 52
X_3 2.78 3.78 3.64 4.14 3.76 4.1

Stratum V
Y 20 27 26 28.5 34.1 35.1 23.5 19 39.5 33 5.7
X_1 42 48 46 64 104 69 39 32 86 56 19
X_2 0.85 0.75 0.87 0.74 1.57 1.64 0.89 1.01 1.45 1.51 0.98

Stratum VI
Y 35.5 48.5 43.5 48 52 47.5
X_1 69 110 77 89 101 81
X_3 3.42 3.78 3.78 3.64 4.14 3.76

Stratum VII
Y 18 24.5 19
X_2 1.04 1.11 0.76
X_3 3.48 3.42 2.54

Stratum VIII
Y 64.5 10 37.3 36.8 33.5 31 16.5 12.8 43.5 32.5 25 62 32.5 51 38.5
32.5 18
X_1 87 22 67 72 68 59 34 26 60 59 44 76 56 129 72 67 36
X_2 1.18 0.71.25 1.24 1.35 1.47 0.86 0.94 0.98 1.36 1.22 1.36
1.13 1.33 1.37 1.27 1.14
X_3 3.74 3.83.98 3.98 4.08 3.92.92 2.66 3.34 4.12 3.83.78 3.46 3.12
3.72 3.98 3.7

Data Set II

Stratum I
Y 36 72 54 35 47 67 38 45 57 30 26 44 49 67 27 39

Stratum II
Y 30 26 44 72 40 67 72 43 36 72 54 35 40
X_1 18 16 39 95 30 69 80 50 53 97 28 48 49

Stratum III
Y 43 36 72 54 35 28 67
X_2 25 55 97 53 42 30 69

Stratum IV
Y 47 67 38 30 26 44 72 40 67 72
X_3 29 77 28 20 19 54 65 30 69 75

Stratum V
Y 72 43 36 72 54 35 40 34 48 52 49 67
X_1 75 19 37 97 53 26 49 44 53 56 72 92
X_2 80 41 46 47 27 31 33 27 55 55 57 58

Stratum VI
Y 27 39 57 28 36 67 44 30
X_1 18 53 60 34 65 69 54 20
X_3 23 29 60 12 46 82 34 33

Stratum VII
Y 28 57 39 27 67 49 52 40 72
X_2 25 60 28 15 55 72 56 49 97
X_3 12 62 30 23 58 57 55 33 66

Stratum VIII
Y 47 67 38 29 57 44 72 40 63 69 43 36 72 54 39
X_1 48 51 24 19 42 54 95 38 69 75 31 25 97 63 29
X_2 47 79 53 31 29 34 75 31 79 75 41 46 66 31 53
X_3 29 29 16 18 18 20 34 18 31 39 32 21 35 31 40

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 217–230

PROBABILITY SAMPLE SELECTION METHOD IN HOUSEHOLD SURVEYS WHEN CURRENT DATA ON REGIONAL POPULATION IS UNAVAILABLE

Turgay Ünalan[1], **H. Öztaş Ayhan**[2]

ABSTRACT

Availability of the perfect sampling frame only exists in developed countries, which covers a very small proportion of the world countries. On the other hand, in developing countries lists of the latest population census counts are generally used as the sampling frame for sample surveys. Therefore, in developing countries surveys which are planned for future periods long after the census date, cannot be representative of the related time period if the same census counts are utilized. Instead, population projections and data adjustment methodologies must be used to provide a representative probability selection of the updated population. This article proposes a population projection and adjustment methodology in order to establish the ideal selection probability for household surveys. The method contains the correction on the differences of the sum of strata and aggregated values. Comparative examples are also provided to clarify the proposed methodology.

Key words: data adjustment, household surveys, population projection, projection methodology, sample selection, selection probability.

1. Introduction

The techniques used to make population projections can be classified as trend extrapolation models or curve fitting techniques and cohort component projection models. These models are useful when we only need to project the total and domain populations. Therefore, they use total population figures from the past to project future population levels. Among the possible alternative curve fitting methods like *linear, geometric, exponential* and *logistic*, the most widely used model for populations is the exponential model.

1 Statistics and Monitoring Section, UNICEF, New York, USA, E-mail: tunalan@unicef.org.
2 Department of Statistics, Middle East Technical University, Ankara 06800, Turkey. E-mail: oayhan@metu.edu.tr.

There are two types of population estimation techniques, namely intercensal estimate (*between two censuses*) and postcensal estimate (*immediate*). Intercensal estimates are known as *interpolation* and postcensal estimates are referred to as *extrapolation*. Estimates and projections can be on the basis of either *de jure* (usual resident) or *de facto* (physically present) enumeration basis for populations. In most cases, they are based on *de facto* population basis.

The paper concerns the problem of limited importance in developed countries where updates of household population projections are relatively frequent between censuses and the proposed method can be unnecessary. However, the problem can be relatively frequent and worth studying in countries with poor population statistics. Therefore, the goal of this paper is to provide a population projection method and adjustment procedures in order to determine representative probability sample selection for household surveys in countries where the updated sampling frame does not exist. On the other hand, it can also be a useful example of the solution of the problem of conducting surveys in case of lack of recent data on population (the number of households by regions).

The paper has a review of the currently known population projection techniques. An overview of the population projection techniques is illustrated in the following section. An exponential model and adjustment procedures for the population projections are also proposed. A case study illustrated details of the adjustment methodologies which are supported by numerical applications.

2. Overview of population projection techniques

The techniques used to make population projections can be classified under two categories: curve fitting or trend models and cohort component projection models.

Curve Fitting Models are useful when we only need to project the total population. Therefore, they use total population figures from the past to project future population levels.

Cohort Component Models are more data intensive because they disaggregate total population figures into age/gender *cohorts*. Further, the different *components* of population change (births, deaths and migration) are taken into account and past figures for these components are applied to the current age/sex cohorts. The cohort component method is grounded in the *basic demographic equation* because there are three ways that a population can change in size and composition: births, deaths, and migration. Basically, the *cohort component method* survives each age-sex population subgroup at a certain time 1 forward to time 2, adds in the new births and in-migrants, and subtracts out-migrants. Clearly, to obtain more precise and more detailed information about the current and future population, the inclusion of age in the projection techniques is a necessary addition.

Information on the cohort component method has been given here because the method is an alternative to the curve fitting or trend models. Due to available data requirements, the cohort-component method is not finally used in this study.

Several population projection techniques and computer programs are available for general purposes in the literature (Davis 1995; Lutz, Vaupel and Ahlburg 1998; Shorter, Sendek and Bayoumy 1995; Stover 1990; United Nations 1989; van Imhoff and Keilman 1991; van Imhoff 1994; Groenwold and Navaneetham 1998).

Among these, early books on population projection methods and models are covered by Cox 1976; Hinde 1997; Newell 1988; Pollard, Yusuf and Pollard 1974; Shryock, Siegel and associates, 1976. Recent books on population projections covers the classical methodologies (Preston, Heuveline and Guillot 2001; Bongaarts and Bulatao 2000).

Many research projects have also been completed on population projections by various researchers (Groenwold and Navaneetham 1998; Davis 1995; Lutz, Vaupel and Ahlburg 1998; State Institute of Statistics 1995; United Nations 1989 and 1990).

Computer package programs are also developed and widely used on the population projections and some selected examples are covered by Leete 1990; Shorter, Sendek and Bayoumy 1995; Stover 1990; Van Imhoff 1994.

Projections based on growth rates assume constant arithmetic, geometric, or exponential growth to estimate populations between dates or to project numbers for a few years ahead. However, as the constant growth rate is seldom case in real life, projections based on growth rates are either limited to short time spans, or the growth rates are varied from one interval to another (Alho and Spencer 2005; Groenewold and Navaneetham 1998; Hinde 1997; Rowland 2003). All of the trend extrapolation models can be used to prepare intercensal and postcensal estimates.

3. Exponential model for population projection

There are many different techniques for the population projections. Among these, we would like to illustrate the use of the exponential model for the population projections. For this methodology, it is essential to obtain and use annual population growth rates. In practice, the recent past population growth rates are determined from the two latest population censuses (if available) for the related domains. For simplicity, we can call these as *data source* (*k*) *and* (*k+1*). The previous annual population growth rates for the corresponding domains can be obtained by using the population counts of the available data sources. Using the same domain dimensions, the future projected population can be labelled as *data source* (*k+2*), for convenience. Generally, the related population domains (or strata) will correspond to the *segregated class boundaries* in survey sampling methodology. Figure 1 illustrates the relationship among domains.

Data source (k): Previous census			r_h $\Rightarrow$ r	**Data source (k+1):** Latest census			$\hat{r}_h$ $\Rightarrow$ $\tilde{r}$	**Data source (k+2):** Planned survey		**Adjusted by**	
$N_1^{(k)}$				$N_1^{(k+1)}$				$\hat{N}_1^{(k+2)}$			
	$N_h^{(k)}$				$N_h^{(k+1)}$				$\hat{N}_h^{(k+2)}$	W_h	δ_h
Total		$N^{(k)}$		Total		$N^{(k+1)}$		Total		$\hat{N}^{(k+2)}$	$\tilde{N}^{(k+2)}$

Figure 1. Graphical illustration of data sources and population structures.

The population domains are evaluated as statistical regions which are based on EUROSTAT's "*Nomenclature of Statistical Territorial Units*" (*NUTS*) classification. The overall population size (N) is equal to the sum of domain populations.

$$\sum_{h=1}^{H} N_h^{(k)} = N^{(k)} \quad \text{and} \quad \sum_{h=1}^{H} N_h^{(k+1)} = N^{(k+1)}$$

The following form of the exponential growth model is used for population projection of domains. For some designs, population domains may correspond to population strata.

$$N_h^{(k+1)} = N_h^{(k)} \; e^{r_h^{(k,\,k+1)}\; t_h^{(k,\,k+1)}} \quad \text{and}$$

$$\frac{N_h^{(k+1)}}{N_h^{(k)}} = \exp\left[r_h^{(k,\,k+1)}\; t_h^{(k,\,k+1)} \right]$$

The estimated annual growth rate of past population can be obtained by

$$r_h^{(k,\,k+1)} = \ln\left[\frac{N_h^{(k+1)}}{N_h^{(k)}} \right] \Big/ t_h^{(k,\,k+1)}$$

Several alternative scenarios can also be used for determining the future growth rate of the population. Unless there is no special reason (like *migration* or *mortality*) of population change for the future period, the past annual growth rate of the population may also be used as a future population growth rate. That is,

$$r_h^{(k,\,k+1)} \Rightarrow \quad \hat{r}_h^{(k+1,\,k+2)}$$

Taking the future domain growth rate as $\hat{r}_h^{(k+1,\,k+2)}$, the future population growth of the domains can be projected by the use of the following model. The total population projection can be obtained as:

$$\hat{N}_h^{(k+2)} = N_h^{(k+1)} \exp\left[\, \hat{r}_h^{(k+1,\,k+2)}\; t_h^{(k+1,\,k+2)} \right]$$

and finally $\sum_{h=1}^{H} \hat{N}_h^{(k+2)} = \hat{N}^{(k+2)}$.

4. Proposed population adjustment methodology

Total population refers to the household population (*members of household*), institutional population (*armed forces, dormitory, hospital, prison, etc.*) and mobile population (*homeless, nomadic tribes, etc*). For a representative household based sample survey, the target population is considered to be equal to the household population.

When the information is available on the amount of institutional population [$N_h^{(INS)}$] and mobile population [$N_h^{(MOB)}$], these amounts has to be subtracted from the adjusted total population [$\widetilde{N}_h^{(k+2)}$] in order to obtain the household population [$\hat{N}_h^{(HH)}$] for each domain (Ayhan and Ekni 2003). This simple relation is given below:

$$\hat{N}_h^{(HH)} = \widetilde{N}_h^{(k+2)} - \left[N_h^{(INS)} + N_h^{(MOB)} \right]$$

Ayhan and Ekni (2003) also provided real numerical count data on the "institutional population" which was based on 5 regional estimates of the 1990 General Population Census of Turkey.

If these information is available, then it has to be used, otherwise the component can be neglected and taken as zero for practical purposes. Recent developments towards the collection, release and use of special information created some drawbacks. Nowadays, there may also be some political and/or

strategic reasons for not stating/releasing some of the institutional population information components for small domains. Under such circumstances, the *de jure* based household population enumeration may not be possible.

For the illustration of the proposed methodology, the information provided for the *data sources* (*k*) and (*k+1*) corresponds to the two previous population census results of the related domains. Therefore, domain totals and overall total information will be identical. On the other hand, information which is based on population projections for *data source* (*k+2*), will not have the same desired properties. In other words, the sum of domains will be equal to the existing domain total, but this will not be equal to the total projection estimate of the overall population.

In order to balance the relationship between the domain totals and the overall total, an adjustment is required. This can be called as the *total population projection adjustment*. This adjustment will correct the difference between the *sum of the projected domain populations* and *total population projections*.

The sum of domain populations will be equal to the existing domain total of $\sum_{h=1}^{H} \hat{N}_h^{(k+2)} = \hat{N}^{(k+2)}$. But this will not be equal to the total projection estimate of the overall population. That is,

$$\sum_{h=1}^{H} \hat{N}_h^{(k+2)} = \hat{N}^{(k+2)} \neq \tilde{N}^{(k+2)} = \mathrm{N}^{(\mathrm{k+1})} \exp\left[\mathrm{r}^{(\mathrm{k+1,\ k+2})}\ t^{(k+1,\ k+2)}\right]$$

assuming that, $r^{(k,\ k+1)} \Rightarrow \tilde{r}^{(k+1,\ k+2)}$ as before.

In order to balance the relationship between the domain totals and the overall total, an adjustment is required. This can be called as the "total population projection adjustment". This adjustment will correct the difference between the "sum of the projected domain populations" and "total population projections". That is, $\hat{N}^{(k+2)} \Rightarrow \tilde{N}^{(k+2)}$.

Let us show the difference between the above total projections as:

$$\Delta = \tilde{N}^{(k+2)} - \hat{N}^{(k+2)} \quad \text{where} \quad \hat{N}^{(k+2)} = \sum_{h=1}^{H} \hat{N}_h^{(k+2)}$$

$$\delta_h = \left(\frac{\hat{N}_h^{(k+2)}}{\hat{N}^{(k+2)}} \right) \Delta = W_h\ \Delta \quad \text{where} \quad W_h = \hat{N}_h^{(k+2)} \Big/ \hat{N}^{(k+2)}$$

$$\tilde{N}_h^{(k+2)} = \hat{N}_h^{(k+2)} + \delta_h$$

Then, the sum of the adjusted projection totals for domains will be equal to

$$\sum_{h=1}^{H} \widetilde{N}_h^{(k+2)} = \widetilde{N}^{(k+2)}.$$

5. A case study

The latest Population Census of Turkey (October 2000) is used as a reference time location of population projections for an intended sample survey which is planned for October 2012.

The sample design which is based on the Classification of Statistical Regions NUTS1–Level refers to 12 regions of Turkey and is taken as the "reporting domains" in the survey literature. The regions are not planned to represent any geographical or socio-economical breakdown of the country, in this case. The population projections for the year 2012 are achieved on the basis of the information provided from two previous population censuses of Turkey, which is given in Table 1.

Table 1. Recent population censuses of Turkey and population projections for domains

Domains h	Previous population census *October 1990* $N_h^{(k)}$	Latest population census *October 2000* $N_h^{(k+1)}$	Annual growth rate of population *2000-2012* $\hat{r}_h^{(k+1,\,k+2)}$	Projected population for planned survey *October 2012* $\hat{N}_h^{(k+2)}$	Domain weights of projected population for 2012 W_h	Amount of population adjustments for domains δ_h	Adjusted projection for planned survey *October 2012* $\widetilde{N}_h^{(k+2)}$
1	7195773	10018735	0.0330963	14903745	0.174663	- 155094	14748651
2	2589490	2895980	0.0111863	3312021	0.038815	- 34466	3277555
3	7594977	8938781	0.0162912	10868772	0.127376	- 113104	10755668
4	4688514	5741241	0.020256	7321001	0.085798	- 76184,9	7244816
5	5204217	6443236	0.0213562	8325348	0.097568	- 86636,5	8238712
6	7026489	8706005	0.0214326	11259408	0.131954	- 117169	11142239
7	3818444	4189268	0.0092683	4682095	0.054871	- 48723,5	4633371

Table 1. Recent population censuses of Turkey and population projections for domains (cont.)

Domains h	Previous population census *October 1990* $N_h^{(k)}$	Latest population census *October 2000* $N_h^{(k+1)}$	Annual growth rate of population *2000-2012* $\hat{r}_h^{(k+1,\,k+2)}$	Projected population for planned survey *October 2012* $\hat{N}_h^{(k+2)}$	Domain weights of projected population for 2012 W_h	Amount of population adjustments for domains δ_h	Adjusted projection for planned survey *October 2012* $\tilde{N}_h^{(k+2)}$
8	4889323	4895744	0.0001312	4903460	0.057466	- 51027,1	4852433
9	2852806	3131546	0.0093224	3502214	0.041044	- 36445,3	3465769
10	2354030	2507738	0.0063252	2705493	0.031707	- 28154,3	2677338
11	3101812	3727034	0.0183626	4645803	0.054446	- 48345,8	4597457
12	5157160	6608619	0.0247989	8899198	0.104293	- 92608,1	8806589
Total	**56473035**	**67803927**	**0.0182857**	**85328559** **84440600**	**1.010516**	**- 887959**	**84440600**

Source: Partly based on T.S.I. (2003).

The difference between the *overall total projection* and the *sum of the domain projections* simply arises from the fact that in (*k+1*) the fast growing components (*h*) have become bigger than they were in (*k*); hence, assuming (r_h) values to be constant from (*k+1*) to (*k+2*) will always give larger sum of the domain projections compared to assuming (*r*) to be constant (which gives overall total projection). This mathematical necessity hardly needs numerical illustration of this type, unless the discrepancy is analysed e.g. as a function of the variation in r_h (if r_h = constant = r, then the two total projections will be identical).

5.1. Comparison of overall selection probabilities for households

Overall selection probabilities for households can be determined from the household based information of the population information for each domain. This information can be obtained by dividing the total population of each domain by the corresponding average household size for this domain.

Here $\frac{\hat{N}_h^{(k+2)}}{\overline{H}_h} = \hat{M}_h \ \forall\ h$ where average household size is $\overline{H}_h = S^{-1}\sum_{i=1}^{S}\overline{H}_{hi}$ where S is the number of subdomains. The overall selection probability of the domains will be equal to the overall sampling fractions. That is

$$Pr\left(F_h^{(DU)}\right)^{-1} = f_h^{(DU)} = \frac{m_h}{\hat{M}_h} \quad \forall\ h$$

Overall sample selection probabilities for households can be determined by obtaining household based information from the population information for each domain. This information can be obtained by dividing the total population of each domain by the corresponding average household size for this domain, which is given in Table 2. The overall selection probability of the domains will be equal to the overall sampling fractions. In this study, overall selection probabilities are compared for projected and non-projected populations. The results have indicated under-representation in the probabilities of selection for non-projected populations, and consequently the selected sample estimates will also be biased.

Table 2. Comparative estimates of several population sizes

Domains H	Latest census population 2000 $N_h^{(k+1)}$	Household survey population 2012 $\hat{N}_h^{(k+2)}$	Average size of households $\overline{H}_h$	No. of Population households 2012 $\hat{M}_h$
1	10018735	14903745	3.85	3871103
2	2895980	3312021	3.56	930343
3	8938781	10868772	3.81	2852696
4	5741241	7321001	4.03	1816626
5	6443236	8325348	4.17	1996486
6	8706005	11259408	4.60	2447697
7	4189268	4682095	4.97	942071
8	4895744	4903460	4.78	1025828
9	3131546	3502214	5.11	685365
10	2507738	2705493	6.01	450165
11	3727034	4645803	6.44	721398
12	6608619	8899198	6.55	1358656
Total	**67803927**	**85328559**	**4.50**	**18961902**

When information for institutional population and mobile populations are not available, then unfortunately the projected total survey population and household survey population has to be taken as equal. That is,

$$\hat{N}_h^{(k+2)} = \widetilde{N}_h^{(k+2)} - \left[\, N_h^{(INS)} + N_h^{(MOB)} \right] = \widetilde{N}_h^{(k+2)} - 0 = \hat{N}_h^{(k+2)}$$

Computed sample sizes of domains for fixed sampling fractions are given in Table 3.

Table 3. Computed sample size of domains for a fixed sampling fraction of $f_h = 0.001$

Domains h	Sample size based on latest census of 2000 $n_h^{(k+1)}$	Sample size of household survey population for 2012 $n_h^{(k+2)}$	Sample size of households for 2012 m_h
1	10019	14904	3871
2	2896	3312	930
3	8939	10869	2853
4	5741	7321	1817
5	6443	8325	1996
6	8706	11259	2448
7	4189	4682	942
8	4896	4903	1026
9	3132	3502	685
10	2508	2704	450
11	3727	4646	721
12	6609	8899	1359
Total	**67805**	**85329**	**18962**

5.2. Alternative sampling fractions

A. For a fixed overall sampling fraction

(1). Based on the population from the latest (2000) population census results.

$$f_h^{(k+1)} = \frac{n_h^{(k+1)}}{N_h^{(k+1)}} = \frac{1}{F_h^{(k+1)}} \quad \text{and} \quad f^{(k+1)} = \frac{67805}{67803927} = \frac{1}{1000} = 0.001$$

(2). Based on the population from the adjusted population projection (2012) results.

$$f_h^{(k+2)} = \frac{n_h^{(k+2)}}{\hat{N}_h^{(k+2)}} = \frac{1}{F_h^{(k+2)}} \quad \text{and similarly}$$

$$f^{(k+2)} = \frac{85329}{85328559} = \frac{1}{1000} = 0.001$$

When there are no values for institutional and mobile population, the resulting values will be the same estimates for all strata.

(3). Based on the number of dwelling units (or households) from the estimated household population (2012) results.

$$f_h^{(DU)} = \frac{m_h}{\hat{M}_h} = \frac{1}{F_h^{(DU)}} \qquad \text{and} \qquad f^{(DU)} = \frac{18962}{18961902} = \frac{1}{1000} = 0.001$$

The *gain* and *relative gain* from alternative sample sizes can be examined by using the following formulation.

The gain in sample size:

$$G(n) = \frac{n^{(k+1)}}{n^{(k+2)}} = \frac{67805}{85329} = 0.795$$

The relative gain in sample size:

$$RG(n) = \left[n^{(k+2)} - n^{(k+1)} \right] / n^{(k+2)} = \left[85329 - 67805\right] / \ 85329$$
$$= \frac{17542}{85329} = 0.205$$

This reflects 20.5% under-representation of sample selection probability over the twelve year period. As an alternative, it is possible to compare alternative selection probabilities for the case of a fixed sample size.

All this merely says that if a sample is selected (only) from past lists of elementary units, a nominal (f) would give smaller (n) or to get a given (n) would need larger (f) – compared to if the same procedure were applied to updated lists. The elaborated and often repeated computations add little to this basic point. Projected population (if accurate) tells us by how much:

$$\frac{n_1}{n_2} = \frac{f_2}{f_1} = \frac{N^{(k+1)}}{N^{(k+2)}} \ .$$

B. For a fixed sample size

(1). Fixed sample size is used from the latest (2000) population census results.

$n_h^{(k+1)} = f_h^{(k+1)} N_h^{(k+1)}$ and $n^{(k+1)} = f^{(k+1)} N^{(k+1)} = (0.001)\ (67803927) = 67805$

For this case, the same fixed sample size $n^{(k+1)} = 67805$ is used in future sample selections. The *gain* and *relative gain* from fixed sample sizes can be examined by using the following relation. The gain in sampling fraction is:

$$G(f) = \frac{f^{(k+1)}}{f^{(k+2)}} = \frac{0.001}{0.000795} = 1.2579$$

The relative gain in sampling fraction is $RG(f) = \left[f^{(HH)} - f^{(k+1)} \right] / f^{(HH)}$

$$RG(f) = [0.000795 - 0.001] / 0.000795 = \frac{0.000205}{0.000795} = \frac{1}{3.878} = 0.2579$$

where $f^{(k+1)} = \frac{n_h^{(k+1)}}{N_h^{(k+1)}} = \frac{67805}{67803927} = \frac{1}{1000} = 0.001$

$$f^{(HH)} = \frac{n_h^{(k+1)}}{\hat{N}_h^{(k+2)}} = \frac{67\,805}{85328559} = \frac{1}{1258} = 0.000795$$

(2). Expansion factor can also be used for comparison.

The relative gain in expansion factors;

$$RG(F) = \left[F^{(k+2)} - F^{(k+1)} \right] / F^{(k+2)}$$

$$RG(F) = [1258 - 1000] / 1258 = 0.205$$

Again, the comparison reflects 20.5% under-representation for sample selection probability over the specified period.

6. Conclusions

This study reflected the results of the importance of the updated population counts as the basis for representative sample selection. The earlier computations based on census data reflects the population under-coverage error when compared to results which are based on latest adjusted methodologies. Comparison of relative gains can be interpreted differently for a fixed sample size and for a fixed sampling fraction. In both cases, the latest and more refined information provides better representation of the population coverage.

Rather than using updated population projections, the use of the latest population counts are very common in developing countries. These figures are creating under-coverage error and serious bias for population representation.

Sample allocation into sub-domains can also be made by taking the urban/rural proportions of the latest population census as weights for the proportional allocation within the stated domains (*population size of 20000 was taken as boundary value for urban/rural domains*). For the sample design there may be other requests to have independent urban and rural domains for inference objectives. In this case, separate population projections have to base on urban and

rural breakdowns. In addition, sample allocations in sub-domains will also be taking care of the "size groups of settlements".

The findings from comparative results of the case study indicated that for a fixed overall sampling fraction the gain in sample size was 0.795, and the relative gain in sample size reflects 20.5% under-representative sample selection probability over the twelve-year period. The gain and the relative gain in sampling fraction are found to be 0.2579. On the other hand, the gain due to the comparison of the expansion factors was 0.205. These results clearly show that the use of new probabilities (*latest overall sampling fractions*) will reflect better representation of the population when compared to old selection probabilities.

Acknowledgements

The authors are grateful to Professor Vijay Verma, Department of Methodology, University of Siena, Italy for his valuable comments on the earlier draft of this paper. We also would like to thank the editor, associate editor and the referees of this journal for very constructive comments, which improved the paper in many ways.

REFERENCES

ALHO, J. M., SPENCER, B. D., (2005). Statistical Demography and Forecasting. Springer, New York.

AYHAN, H. Ö., EKNI, S., (2003). Coverage Error in Population Censuses: The Case of Turkey. Survey Methodology 29 (2), 155–165.

BONGAARTS, J., BULATAO, R. A. (editors), (2000). Beyond Six Billion:

Forecasting the World's Population. Panel on Population Projections, National Academy Press, Washington D. C.

COX, P., (1976). Demography. Cambridge University Press, Cambridge.

DAVIS, H. C., (1995). Demographic Projection Techniques for Regions and Smaller Areas: A Primer, UBC Press, Vancouver.

GROENEWOLD, G., NAVANEETHAM, K., (1998). The Projection of Populations: Data Appraisal, Basic Methods, and Applications. Centre for Development Studies and UN Population Fund, Kerala.

HINDE, A., (1997). Demographic Methods. Arnold Publishers, London.

LEETE, R., (1990). People: User's Manual. Overseas Development Administration, United Kingdom and Economic Planning Unit, Kuala Lumpur. *PEOPLE Version 2.0*.

LUTZ, W., VAUPEL, J. W., AHLBURG, D. A. (editors), (1998). Frontiers of Population Forecasting, A Supplement to Vol. 24, 1998, Population and Development Review, Population Council, New York.

NEWELL, C., (1988). Methods and Models in Demography. Belhaven Press, London.

POLLARD, A. H., YUSUF, F., POLLARD, G. N., (1974). Demographic Techniques. Pergamon Press, Sydney.

PRESTON, S. H., HEUVELINE, P., GUILLOT, M., (2001). Demography: Measuring and Modeling Population Processes. Blackwell Publishers, Oxford.

ROWLAND, D. T., (2003). Demographic Methods and Concepts. Oxford University Press, Oxford.

SHORTER, F., SENDEK, R., BAYOUMY, Y., (1995). Computational Methods for Population Projections: With Particular Reference to Development Planning, Second Edition, The Population Council, New York. *FIVFIV Version 11.0*.

SHRYOCK, H. S., SIEGEL, J. S. and associates, (1976). The Methods and Materials in Demography. Academic Press, Inc., Florida.

STATE INSTITUTE OF STATISTICS, (1995). The Population of Turkey, 1923–1994, Demographic Structure and Development: With Projections to the Mid-21st Century. State Institute of Statistics, Publication No. 1716, Ankara, Turkey.

STOVER, J., (1990). Demproj: A Demographic Projection Model for Development Planning, The Futures Group, Glastonbury. *DEMPROJ Version 3.0*.

T. S. I., (2003). Census of Population 2000. Social and Economic Characteristics of Population, Turkey. Turkish Statistical Institute Publication Number 2759. Ankara, Turkey.

UNITED NATIONS, (1989). Projection Methods for Integrating Population Variables into Development Planning, Volume I: Methods for Comprehensive Planning, Module One: Conceptual Issues and Methods for Preparing Demographic Methods, United Nations Publications, ST/ESA/SER.R/90/Add.1, New York.

UNITED NATIONS, (1990). Projection Methods for Integrating Population Variables into Development Planning, Volume I: Methods for Comprehensive Planning, Module Two: Methods for Preparing School Enrolment, Labour Force and Employment Projections. United Nations Publications, ST/ESA/SER.R/90/Add.1, New York.

VAN IMHOFF, E., KEILMAN, N., (1991). LIPRO 2.0: An Application of a Dynamic Demographic Projection Model to Household Structure in the Netherlands. Netherlands Interdisciplinary Demographic Institute. The Hague, Netherlands.

VAN IMHOFF, E., (1994). Lipro User's Guide, Netherlands Interdisciplinary Demographic Institute, The Hague. *LIPRO Version 3.0*.

STATISTICS IN TRANSITION new series, Summer 2013
Vol.14, No. 2, pp. 231-248

SAMPLING DESIGNS PROPORTIONATE TO SUM OF TWO ORDER STATISTICS OF AUXILIARY VARIABLE

Janusz L. Wywiał[1]

ABSTRACT

In this paper the case of a conditional sampling design proportional to the sum of two order statistics is considered. Several strategies including the Horvitz-Thompson estimator and ratio-type estimators are discussed. The accuracy of these estimators is analyzed on the basis of computer simulation experiments.

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

1. Introduction

Sampling designs are usually constructed on the basis of an auxiliary variable observations in order to improve the accuracy of estimation of population parameters. For instance, the sampling design of Lahiri (1951), Midzuno (1952) and Sen (1953) proportional to the sample mean of the positive valued auxiliary variable leads to unbiasednees of the well known ordinary ratio estimator. Wywial (2008, 2009) proposed sampling designs dependent on order statistics of the auxiliary variable. Here that approach is continued because the sampling design proportional to the sum of two auxiliary variable order statistics is proposed. The conditional version of this sampling design is considered in order to improve the estimation effects. The review of the sampling designs or schemes dependent on auxiliary variables and their conditional versions is considered by Tillé (2006).

[1]Katowice University of Economics, Poland, janusz.wywial@ue.katowice.pl.

2. Sampling design

Let U be a fixed population of the size N. The observation of a variable under study and of an auxiliary variable are denoted by y_i and $x_i, i = 1, \ldots, N$, respectively. Moreover, let $0 < x_i \leq x_{i+1}, i = 1, \ldots, N-1$. Our problem is the estimation of the population average $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$. Let us consider the sample space $\mathbf{S}$ of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$, so $P(s) \geq 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$.

Let $(X_{(j)})$ be the sequence of the order statistics of observations of auxiliary variable in the sample s. It is well known that the sample quantile of the order $\alpha \in (0;1)$ is defined as follows: $Q_{s,\alpha} = X_{(r)}$ where $r = [n\alpha] + 1$, the function $[n\alpha]$ means the integer part of the value $n\alpha$, $r = 1, 2, ..., n$. Let us note that $X_{(r)} = Q_{s,\alpha}$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$. In this paper it will be more convenient to consider the order statistic rather than the quantile. Let $G(r,u,i,j) = \{s : X_{(r)} = x_i, X_{(u)} = x_j\}$, $r = 1, .., n-1$; $u = 2, ..., n$, $r < u$ be the set of all samples whose r -th and u -th order statistics of the auxiliary variable are equal to x_i and x_j, respectively, where $r \leq i < j \leq N - n + u$.

$$\bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j) = \mathbf{S}. \tag{1}$$

The size of the set $G(r,u,i,j)$ is denoted by $g(r,u,i,j) = Card(G(r,u,i,j))$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}, \tag{2}$$

$$\binom{N}{n} = Card(\mathbf{S}) = Card\left(\bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j)\right) =$$

$$= \bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} Card(G(r,u,i,j)) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} g(r,u,i,j),$$

$$P\left(X_{(r)} = x_i, X_{(u)} = x_j\right) = \frac{g(r,u,i,j)}{\binom{N}{n}}. \tag{3}$$

Let $h(x_j, x_i)$ be a non-negative function of values x_j, x_i of the order statistics $X_{(u)}$ and $X_{(r)}$, respectively. Moreover let

$$f(x_j, x_i, c) = \begin{cases} h(x_j, x_i) & \text{if} \quad h(x_j, x_i) \geq c, \\ 0 \quad for & h(x_j, x_i) < c. \end{cases} \tag{4}$$

and

$$z(r,u,c) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) g(r,u,i,j). \tag{5}$$

The straightfoward generalization of the Wywiał's (2009) sampling design is as follows.

Definition 1.1. The conditional sampling design proportional to the non-negative functions of the order statistics $X_{(u)}$, $X_{(r)}$ is as follows:

$$P_{r,u}(s|c) = \frac{f(x_j, x_i, c)}{z(r,u,c)} \qquad \text{for} \quad s \in G(r,u,i,j) \tag{6}$$

where $i < j, \quad r \leq i \leq N-n+r \quad and \quad r < u \leq j \leq N-n+u$, $0 \leq c \leq c_0$.

As it is well known, the inclusion probability of the first order is determined by the equations: $\pi_k(r,u,c) = P_{r,u}(s : k \in s|c) = \sum_{\{s:k\in s\}} P_{r,u}(s|c), \quad k = 1, ..., N$.

Now, the upper possible value c_0 of the constant c should be stated in such a way that $\pi_k(r,u,c) > 0$ for all $k = 1, ..., N$. It is important because under the just defined condition the well known Horvitz-Thompson estimator (considered in the sections 3) is unbiased for the a population mean.

The above defined sampling design is treated as conditional (unconditional) when $c > 0$ $(c = 0)$, see the definition of the conditional sampling design considered by Tillé (1999, 2006).

Particularly, let $h(x_j, x_i) = x_j + x_i$. Thus

$$f(x_j, x_i, c) = \begin{cases} x_j + x_i & for \quad x_j + x_i \geq c, \\ 0 \quad for & x_j + x_i < c. \end{cases} \tag{7}$$

We have to assume that $0 \leq c \leq c_0 = x_1 + x_N$. Thus, under this assumption the inclusion probabilities $\pi_N(r,u,c) > 0$ for all $i = 1, ..., N$. When $c > c_0$, $x_1 + x_i < c$ for all $i = 2, ..., N$ and $\pi_1(r,u,c) = 0$ and $\pi_k(r,u,c) \geq 0$ for $k = 2, ..., N$. In this case, as it is well known, the Horvitz-Thompson's statistic is a biased estimator of the population mean.

Let $\delta(x)$ be such a function that if $x \leq 0$ then $\delta(x) = 0$ otherwise $\delta(x) = 1$. Moreover, $\delta(x)\delta(x-1) = \delta(x-1)$. The following two theorems are the straightforward generalizations of those ones proved by Wywiał(2009).

Theorem 1. *The inclusion probabilities of the first order for the conditional sampling design $P_{r,u}(s|c)$ are as follows:*

$$\pi_k(r,u,c) =$$

$$= \frac{1}{z(r,u,c)}\Bigg(\delta(r-1)\sum_{i=r\delta(r-k)+(k-1)\delta(k+1-r)\delta(N-n+r-k)}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-2}{r-2}$$

$$\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_i,c)+\delta(k-r)\delta(N-n+u-k)\delta(u-r-1)$$

$$\sum_{i=r}^{min(k-1,N-n+r)}\sum_{j=max(i+u-r,k+1)}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j}{n-u}f(x_j,x_i,c)+$$

$$+\delta(k-u)\delta(n-u)\delta(N-n+u-k+1)$$

$$\sum_{i=r}^{k-u+r-1}\sum_{j=i+u-r}^{k-1}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+\delta(n-u)$$

$$\delta(k-N+n-u)\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+$$

$$+\delta(k+1-r)\delta(N-n+r-k+1)\binom{k-1}{r-1}$$

$$\sum_{j=k+u-r}^{N-n+u}\binom{j-k-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_k,c)+\delta(k-u+1)\delta(N-n+u-k+1)$$

$$\binom{N-k}{n-u}\sum_{i=r}^{k-u+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f(x_k,x_i,c)\Bigg) \quad (8)$$

Theorem 2. *The inclusion probabilities of the second order for the conditional sampling design $P_{r,u}(s|c)$ are as follows:*

$$\pi_{k,t}(r,u,c) = P\left(k,t\in s_1\right)+P\left(k\in s_1,X_{(r)}=x_t\right)+P\left(k\in s_1,t\in s_2\right)+$$
$$+P\left(k\in s_1,X_{(u)}=x_t\right)+P\left(k\in s_1,t\in s_3\right)+P\left(X_{(r)}=x_k,t\in s_2\right)+$$
$$+P\left(X_{(r)}=x_k,X_{(u)}=x_t\right)+P\left(X_{(r)}=x_k,t\in s_3\right)+P\left(k,t\in s_2\right)+$$
$$+P\left(k\in s_2,X_{(u)}=x_t\right)+P\left(k\in s_2,t\in s_3\right)+P\left(X_{(u)}=x_k,t\in s_3\right)+$$
$$+P\left(k,t\in s_3\right) \quad (9)$$

where

$$P\left(k, t \in s_1\right) = \frac{\delta(r-2)\delta(N-n+r-t)}{z(r,u)} \cdot$$
$$\cdot \sum_{i=max(r,t+1)}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-3}{r-3}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c).$$

$$P\left(k \in s_1, X_{(r)} = x_t\right) =$$
$$= \frac{\delta(r-1)\delta(N-n+r-k)\delta(N-n+r+1-t)\delta(t+1-r)}{z(r,u)} \cdot$$
$$\cdot \binom{t-2}{r-2} \sum_{j=t+u-r}^{N-n+u} \binom{j-t-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_t, c),$$

$$P\left(k \in s_1, t \in s_2\right) = \delta(N-n+r-k) \cdot$$
$$\cdot \frac{\delta(t-r)\delta(r-1)\delta(u-r-1)\delta(N-n+u-t)\delta(t-k-1)}{z(r,u)} \cdot$$
$$\cdot \sum_{i=max(r,k+1)}^{min(t-1,N-n+r)} \sum_{j=max(t+1,i+u-r)}^{N-n+u} \binom{i-2}{r-2}\binom{j-i-2}{u-r-2}\binom{N-j}{n-u} f(x_j, x_i, c),$$

$$P\left(k \in s_1, X_{(u)} = x_t\right) = \delta(N-n+r-k) \cdot$$
$$\cdot \frac{\delta(t+1-u)\delta(N-n+u+1-t)\delta(t-k-u+r)\delta(r-1)}{z(r,u)} \cdot$$
$$\cdot \binom{N-t}{n-u} \sum_{i=max(r,k+1)}^{t-u+r} \binom{i-2}{r-2}\binom{t-i-1}{u-r-1} f(x_t, x_i, c),$$

$$P\left(k \in s_1, t \in s_3\right) =$$
$$= \frac{\delta(N-n+r-k)\delta(t-u)\delta(r-1)\delta(n-u)\delta(t-k-u+r-1)}{z(r,u)} \cdot$$
$$\cdot \sum_{i=max(r,k+1)}^{min(t-u+r-1,N-n+r)} \sum_{j=i+u-r}^{min(t-1,N-n+u)} \binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1} \cdot$$
$$\cdot f(x_j, x_i, c),$$

$$P\left(X_{(r)}=x_k, t\in s_2\right)=$$
$$=\frac{\delta(k+1-r)\delta(N-n+r+1-k)\delta(t-r)\delta(N-n+u-t)\delta(u-r-1)}{z(r,u)}\cdot$$
$$\cdot\binom{k-1}{r-1}\sum_{j=max(t+1,k+u-r)}^{N-n+u}\binom{j-k-2}{u-r-2}\binom{N-j}{n-u}f(x_j,x_k,c),$$

$$P\left(X_{(r)}=x_k, X_{(u)}=x_t\right)=$$
$$=\frac{\delta(k+1-r)\delta(N-n+r+1-k)\delta(t+1-u)\delta(N-n+u-t+1)}{z(r,u)}\cdot$$
$$\cdot\delta(t+1-k-u+r)\binom{k-1}{r-1}\binom{t-k-1}{u-r-1}\binom{N-t}{n-u}f(x_t,x_k,c),$$

$$P\left(X_{(r)}=x_k, t\in s_3\right)=$$
$$=\frac{\delta(k+1-n)\delta(N-n+r+1-k)\delta(t-u)\delta(n-u)\delta(t-k-u+r)}{z(r,u)}\cdot$$
$$\cdot\binom{k-1}{r-1}\sum_{j=k+u-r}^{min(t-1,N-n+u)}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_k,c),$$

$$P\left(k,t\in s_2\right)=\frac{\delta(k-r)\delta(N-n+u-t)\delta(u-r-2)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=r}^{min(k-1,N-n+r)}\sum_{j=max(t+1,i+u-r)}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-3}{u-r-3}\binom{N-j}{n-u}f(x_j,x_i,c),$$

$$P\left(k\in s_2, X_{(u)}=x_t\right)=$$
$$=\frac{\delta(k-r)\delta(N-n+u-k)\delta(t+1-u)\delta(N-n+u+1-t)\delta(u-r-1)}{z(r,u)}\cdot$$
$$\cdot\binom{N-t}{n-u}\sum_{i=r}^{min(k-1,t-u+r)}\binom{i-1}{r-1}\binom{t-i-2}{u-r-2}f(x_t,x_i,c),$$

$$P\left(k \in s_2, t \in s_3\right) =$$
$$= \frac{\delta(k-r)\delta(N-n+u-k)\delta(t-u)\delta(t-k-1)\delta(u-r-1)\delta(n-u)}{z(r,u)} \cdot$$
$$\cdot \sum_{i=r}^{min(k-1,t-u+r-1,N-n+r)} \sum_{j=max(i+u-r,k+1)}^{min(t-1,N-n+u)} \binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1} \cdot$$
$$\cdot f(x_j, x_i, c),$$

$$P\left(X_{(u)} = x_k, t \in s_3\right) =$$
$$= \frac{\delta(k+1-u)\delta(N-n+u+1-k)\delta(t-u)\delta(n-u)}{z(r,u)} \cdot$$
$$\cdot \binom{N-k-1}{n-u-1} \sum_{i=r}^{k-n+r} \binom{i-1}{r-1}\binom{k-i-1}{u-r-1} f(x_k, x_i, c),$$

$$P\left(k, t \in s_3\right) = \frac{\delta(k-u)\delta(n-u-1)}{z(r,u)} \cdot$$
$$\cdot \sum_{i=r}^{min(N-n+r,k-1-u+r)} \sum_{j=i+u-r}^{min(k-1,N-n+u)} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2} \cdot$$
$$\cdot f(x_j, x_i, c).$$

3. Sampling scheme

The sampling scheme implementing the sampling design $P_{r,u}(s|c)$ is as follows. Firstly, population elements are ordered according to increasing values of the auxiliary variable. Let $s = s_1 \cup \{i\} \cup s_3 \cup \{j\} \cup s_3$ where $s_1 = \{k : k \in U, x_k < x_i\}$ is the simple random sample of the size $r-1$ drawn without replacement from the subpopulation $U(1, i-1) = (1, ..., i-1)$, $s_2 = \{k : k \in U, x_j > x_k > x_i\}$ is the simple random sample of the size $u-r-1$ drawn without replacement from $U(i+1, j-1) = (i+1, ..., j-1)$ and $s_3 = \{k : k \in U, x_k > x_j\}$ s the simple random sample of the size $n-u$ drawn without replacement from $U(j+1, N) = (j+1, ..., N)$. Let us note that $U = U(1, i-1) \cup \{i\} \cup U(i+1, j-1) \cup \{j\} \cup U(j+1, N)$. Let $\mathbf{S}\left(U(1, i-1); s\right)$ be sample space of the sample s_1, let $\mathbf{S}\left(U(i+1, j-1); s\right)$ be sample space of the sample s_2 and let $\mathbf{S}\left(U(j+1, N); s\right)$ be sample space of the sample s_3. Moreover, $\mathbf{S} = \mathbf{S}\left(U, s\right)$).

The sampling scheme is given by the following probabilities:

$$P_{r,u}(s|c) = P_1(s_1|i)p_{r,u}(i|c)P_2(s_2|i,j)p_{r,u}(j|i,c)P_3(s_3|j) \tag{10}$$

where

$$P_1(s_1|i) = \binom{i-1}{r-1}^{-1}, \quad P_2(s_2|i,j) = \binom{j-i-1}{u-r-1}^{-1},$$

$$P_3(s_3|j) = \binom{N-j}{n-u}^{-1},$$

$$p_{r,u}(j|i,c) = \frac{p_{r,u}(i,j|c)}{p_{r,u}(i|c)}, \tag{11}$$

$$p_{r,u}(i,j|c) = \sum_{s\in G(r,u,i,j)} P_{r,u}(s) = \frac{f(x_j,x_i,c)g(r,u,i,j)}{z(r,u,c)}, \tag{12}$$

$$p_{r,u}(i|c) = \frac{1}{z(r,u,c)} \sum_{j=i+u-r}^{N-n+u} f(x_j,x_i,c)g(r,u,i,j). \tag{13}$$

In order to select the sample s, firstly the i-th element of the population should be selected according to the probability function $p_{r,u}(i|c)$. Next, the j-the element of the population should be drawn according to the probability function $p_{r,u}(j|i,c)$. Finally, the samples s_1, s_2 and s_3 should be selected according to the sampling designs $P_1(s_1)$, $P_2(s_2)$ and $P_3(s_3)$, respectively.

4. Some sampling strategies

The well known Horvitz-Thompson estimator (1952) is as follows:

$$\bar{y}_{HT,s} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}. \tag{14}$$

The statistic is unbiased estimator of the population mean value if $\pi_k > 0$ for $k = 1, ..., N$. The variance and its estimator are determined by the expressions (20) and (22), respectively.

The particular case of the above estimator is the well known sampling design of the simple random sample drawn without replacement whose sampling design is: $P_0(s) = \binom{N}{n}^{-1}$. The variance of the mean from the simple random sample $\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k$ drawn without replacement is $D^2(\bar{y}_s, P_0(s)) = \frac{N-n}{nN}v(y)$ where $v(y) = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

Let us construct the following ratio sampling strategy for the population mean $\bar{y} = \frac{1}{N}\sum_{i\in U} y_i$. We assume that $y_i = bx_i + e_i$ for all $i \in U$, $\sum_{i\in U} e_i =$

0 and the residuals of that linear regression function are not correlated with the auxiliary variable. Thus, it has been assumed that the intercept of the linear regression function is equal to zero. The linear correlation coefficient between the variables y and x will be denoted by ρ. Let $\left(X_{(r)}, Y_r\right)$ be two dimensional random variables where $X_{(r)}$ is the r-th order statistic of an auxiliary variable and Y_r is the variable under study. Let us define the following ratio type estimator:

$$\bar{y}_{r,u,s} = \bar{y}_s \frac{E\left(X_{(r)} + X_{(u)}|c\right)}{X_{(r)} + X_{(u)}} \tag{15}$$

where on the basis of the expressions (3) and (5) we have

$$\begin{aligned} &E\left(X_{(r)} + X_{(u)}|c\right) = \\ &= \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) P\left(X_{(r)} = x_i, X_{(u)} = x_j | X_{(r)} + X_{(u)} \geq c\right) = \\ &\qquad \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) \frac{P\left(X_{(r)} = x_i, X_{(u)} = x_j\right)}{P\left(X_{(r)} + X_{(u)} \geq c\right)} = \\ &\qquad = \frac{\sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) g(r, u, i, j)}{\sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \gamma(x_j, x_i, c) g(r, u, i, j)} = \frac{z(r, u, c)}{\alpha(r, u, c)}, \end{aligned} \tag{16}$$

$$\alpha(r, u, c) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \gamma(x_j, x_i, c) g(r, u, i, j) = \binom{N}{n} P\left(X_{(r)} + X_{(u)} \geq c\right),$$

$$\gamma(x_i, x_j, c) = \begin{cases} 1 & \text{for} \quad x_i + x_j \geq c \\ 0 & \text{for} \quad x_i + x_j < c. \end{cases}$$

Let $\mathbf{S}_c = \{s : x_{(r)} + x_{(u)} \geq c\}$ and $\mathbf{S}_{\bar{c}} = \{s : x_{(r)} + x_{(u)} < c\} = \mathbf{S} - \mathbf{S}_c$ where $(x_{(r)}, x_{(u)})$ are values of the order statistics $(X_{(r)}, X_{(u)})$. Moreover, let $\bar{y}_c = \frac{1}{Card(\mathbf{S}_c)} \sum_{s \in \mathbf{S}_c} \bar{y}_s$, $\bar{y}_{\bar{c}} = \frac{1}{Card(\mathbf{S}_{\bar{c}})} \sum_{s \in \mathbf{S}_{\bar{c}}} \bar{y}_s$ where $Card(\mathbf{S}_c) = \alpha(r, u, c)$. Under the stated assumptions we have:

$$\begin{aligned} &E\left(\bar{y}_{r,u,s}, P_{r,u}(s|c)\right) = \sum_{s \in \mathbf{S}_c} \bar{y}_s \frac{E\left(X_{(r)} + X_{(u)}|c\right)}{x_{(r)} + x_{(u)}} P_{r,u}(s|c) = \\ &= \sum_{s \in \mathbf{S}_c} \bar{y}_s \frac{z(r, u, c)}{\alpha(r, u, c)\left(x_{(r)} + x_{(u)}\right)} \frac{x_{(r)} + x_{(u)}}{z(r, u, c)} = \frac{1}{Card(\mathbf{S}_c)} \sum_{s \in \mathbf{S}_c} \bar{y}_s = \bar{y}_c. \end{aligned}$$

Particularly, if $c = 0$ then $P_{r,u}(s|0) = P_{r,u}(s)$ and

$$E\left(\bar{y}_{r,u,s}, P_{r,u}(s)\right) = \sum_{s\in\mathbf{S}} \bar{y}_s \frac{E\left(X_{(r)} + X_{(u)}\right)}{x_{(r)} + x_{(u)}} P_{r,u}(s) = \frac{1}{\binom{N}{n}} \sum_{s\in\mathbf{S}} \bar{y}_s = \bar{y}.$$

These results and the decomposition:

$$\begin{aligned}\bar{y} = \frac{1}{\binom{N}{n}} \sum_{s\in\mathbf{S}} \bar{y}_s &= \frac{1}{\binom{N}{n}} \left(\sum_{s\in\mathbf{S}_c} \bar{y}_s + \sum_{s\in\mathbf{S}_{\bar{c}}} \bar{y}_s \right) = \\ &= \frac{1}{\binom{N}{n}} \left(\alpha(r,u,c)\bar{y}_c + (1-\alpha(r,u,c))\bar{y}_{\bar{c}}\right) = \\ &= \bar{y}_c P\left(X_{(r)} + X_{(u)} \geq c\right) + \bar{y}_{\bar{c}} P\left(X_{(r)} + X_{(u)} < c\right)\end{aligned}$$

lead to the following expression:

$$E\left(\bar{y}_{r,u,s}, P_{r,u}(s|c)\right) = \begin{cases} \bar{y} & \text{for} \quad c = 0 \\ \bar{y} + \left(\bar{y}_c - \bar{y}_{\bar{c}}\right) P\left(X_{(r)} + X_{(u)} < c\right) & \text{for} \quad c > 0. \end{cases} \tag{17}$$

Hence, under the unconditional sampling design $P_{r,u}(s)$ the strategy $(\bar{y}_{r,u,s}, P_{r,u}(s|c))$ is unbiased.

The next ratio type estimator, see e.g. Särndal at. all (1992), is as follows:

$$\tilde{y}_s = \bar{y}_{HT,s} \frac{\bar{x}}{\bar{x}_{HT,s}} \tag{18}$$

The parameters of the strategy $(\tilde{y}_s, P_{r,u}(s|c))$ are approximately as follows:

$$E\left(\tilde{y}_s, P_{r,u}(s|c)\right) \approx \bar{y},$$

$$\begin{aligned}D^2\left(\tilde{y}_s, P_{r,u}(s|c)\right) \approx D^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) - 2hCov\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) + \\ + h^2 D^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right)\end{aligned} \tag{19}$$

where $h = \frac{\bar{y}}{\bar{x}}$ and

$$Cov\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) = \frac{1}{N^2} \left(\sum_{k\in U} \sum_{l\in U} \Delta_{k,l} \frac{y_k}{\pi_k} \frac{x_l}{\pi_l} \right), \tag{20}$$

$\Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l$, $\quad D^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right) = Cov\left(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right)$,
$D^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) = Cov\left(\bar{y}_{HT,s}, \bar{y}_{HT,s}, P_{r,u}(s|c)\right)$.

The variance: $D^2\left(\bar{y}_{r,u,s}, P_{r,u}(s|c)\right)$ can be estimated by the following approximately unbiased estimator:

$$\hat{D}^2\left(\tilde{y}_s, P_{r,u}(s|c)\right) = \hat{D}^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) + \\ - 2h_{r,u,s}\widehat{Cov}\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) + h_{r,u,s}^2\hat{D}^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right) \quad (21)$$

where

$$h_{r,u,s} = \frac{\bar{y}_{HT,s}}{\bar{x}_{HT,s}},$$

$$\widehat{Cov}\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) = \frac{1}{N^2}\left(\sum_{k\in s}\sum_{l\in s}\Delta_{*,k,l}\frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right), \quad (22)$$

$\Delta_{*k,l} = \frac{\Delta_{k,l}}{\pi_{k,l}}$, $\hat{D}^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right) = \widehat{Cov}\left(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right)$,
$\hat{D}^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) = \widehat{Cov}\left(\bar{y}_{HT,s}, \bar{y}_{HT,s}, P_{r,u}(s|c)\right)$.

Let us remind the following ordinary ratio estimator.

$$\hat{y}_s = \bar{y}_s\frac{\bar{x}}{\bar{x}_s}. \quad (23)$$

The approximate value of the variance is as follows:

$$D^2\left(\hat{y}_s, P_0(s)\right) \approx \frac{N(N-n)}{n}\left(v(y) + h^2v(x) - 2hv(x,y)\right)$$

where $v(x,y) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})$ and particularly $v(y) = v(y,y)$, $v(x) = v(x,x)$.

The approximately unbiased estimator of the variance is as follows:

$$\hat{D}^2\left(\hat{y}_s, P_0(s)\right) = \frac{N(N-n)}{n}\left(v_s(y) + h_s^2v_s(x) - 2h_sv_s(x,y)\right),$$

where $h_s = \frac{\bar{y}_s}{\bar{x}_s}$, $v_s(x,y) = \frac{1}{n-1}\sum_{k\in s}(x_k - \bar{x}_s)(y_k - \bar{y}_s)$ and particularly $v_s(y) = v_s(y,y)$, $v_s(x) = v_s(x,x)$. The strategy $(\hat{y}_s, P_0(s))$ is approximately unbiased for the population mean $\bar{y}$.

It is well known that the strategy $(\hat{y}_s, P_1(s))$ is unbiased for $\bar{y}$ where

$$P_1(s) = \frac{\bar{x}_s}{\binom{N}{n}} \quad (24)$$

is the sampling design of Lahiri (1951), Midzumo (1952) and Sen (1953).

5. Simulation analysis of strategies' accuracy

The population taken into account consists of the municipalities in Sweden whose number is $N = 284$. The value x_k, $k = 1, ..., N$, of the auxiliary variable x is equal to the size (in thousands) of people population in the k-th

municipality in 1975. The value y_k, $k = 1, .., N$, of the variable under study y is the taxation revenues (in millions of kronor) from the k-th municipality in 1985. Their observations were published by Särndal, Swenson and Wretman (1992), pp. 652-659. number is $N = 284$.

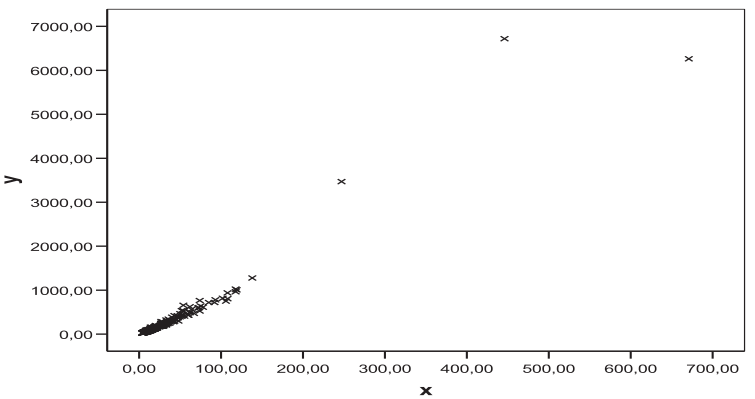


FIGURE 1. Scatterplot of y versus x.

There are three outlier observations of the variable as it is shown in Figure 1. Let σ and β_3 be the standard deviation and the skewness coefficient, respectively, of the auxiliary variable in the population. In case of data without outliers (the number of municipalities is 281) $\bar{x} = 24,263$, $\sigma = 24,153$ and $\beta_3 = 0,043$. In case of data with outliers (the number of municipalities is N=284) $\bar{x} = 28.810$, $\sigma = 52,873$ and $\beta_3 = 8,427$. Thus, the distribution of the variable without the outliers is almost symmetric. But the distribution of the variables with outliers is highly right skewed. The samples according to the preassigned sampling design were drawn from the just presented population. The samples were replicated 1000 times.

The Figure 1 shows that the dependence between the variable under study and the auxiliary variable can be approximated by means of a linear regression with its constance equal to zero. In this case, as it is well-known, the accuracies of regression type estimators of a population mean are similar to the accuracy of the ratio type estimators. Moreover, the ratio estimators are simpler than the regression ones. Thus, that is why we consider only the ratio estimator in the analysis.

Let $MSE(t, P(s))$ be the mean square error of the strategy $(t, P(s))$ used to estimate the population mean $\bar{y}$. The coefficient of the relative efficiency is defined as follows:

$$e(t, P(s)) = \frac{MSE(t, P(s))}{D^2(\bar{y}_s, P_0(s))} 100\%$$

The results of the simulation analysis are presented by Tables and Figures 2, 3. The tables show the relative efficiency coefficients. The distributions of the estimators, generated on the basis of samples of the size $n = 14$ drawn by means of appropriate sampling schemes are presented by means of the well-known box-plots on Figures 2 and 3. Table 1 shows the notation of the strategies.

TABLE 1. The symbols of the strategies.

strategy	symbol	efficiency
$\bar{y}_s, P_0(s)$	$ty0$	$-$
$\hat{y}_s, P_0(s)$	$ty1$	e_1
$\hat{y}_s, P_1(s)$	$ty2$	e_2
$\bar{y}_{HT,s}, P_1(s)$	$ty3$	e_3
$\bar{y}_{HT,s}, P_{n-1,n}(s)$	$ty4$	e_4
$\bar{y}_{HT,s}, P_{n-1,n}(s\|3\bar{x})$	$ty4d3$	e_4
$\tilde{y}_s, P_{n-1,n}(s)$	$ty5$	e_5
$\tilde{y}_s, P_{n-1,n}(s\|3\bar{x})$	$ty5d3$	e_5
$\bar{y}_{n-1,n,s}, P_{n-1,n}(s)$	$ty6$	e_6
$\bar{y}_{n-1,n,s}, P_{n-1,n}(s\|3\bar{x})$	$ty6d3$	e_6

TABLE 2. The relative efficiency coefficients (%) of the strategies.

N :	281			284		
n	$e1$	$e2$	$e3$	$e1$	$e2$	$e3$
2 (0,7%)	2.50	2.24	17.62	1.56	1.15	5.60
3 (1%)	2.88	2.58	28.10	1.88	1.35	8.64
6 (2%)	3.09	2.86	46.62	2.73	1.96	14.87
9 (3%)	3.28	3.09	57.95	3.78	2.66	20.49
14 (5%)	3.42	3.30	73.59	5.04	3.62	29.33
29 (10%)	3.22	3.14	84.51	7.22	5.92	47.44

Firstly, let us consider the strategies under the unconditional sampling designs. Table 2 shows the relative efficiency coefficients of the strategies which

do not depend on the sampling design $P_{r,u}(s|c)$. The ratio estimator under the sampling design $P_1(s)$ is slightly better than the ratio estimator under the simple random sample and they are both significantly more accurate than the Horvtiz-Thompson estimator under the sampling design $P_1(s)$. In general, Tables 2, 3 and 4 let us infer that the ratio type strategies are significantly better than the Horvitz-Thompson ones. This conclusion is strongly confirmed by the box-plots shown by Figures 2 and 3. The strategy $(\hat{y}_s, P_1(s))$ is the best among the six considered strategies except for the case of the population with outliers when the strategy $(\tilde{y}_s, P_{n-1,n}(s))$ is the best for $n = 14$ $n = 29$.

Wywiał(2007) considered the accuracy of estimation on the basis of the sampling design proportional to one order statistic denoted by $P_r(s)$. For some of possible values r of the sampling design $P_r(s)$ the mean squares of the estimators were determined on the basis of the simulation analysis. The results of the analysis lead to the conclusion that the considered strategies dependent on sampling design $P_r(s)$ are the most accurate when $r = n - 1$ or $r = n$. That is why Tables 3 and 4 deal only with the case when $r = n-1$ or $r = n$. In general, all the inclusion probabilities of the first order are expected to be proportional to the appropriate values of a positively valued auxiliary variable. Thus, in the case considered we can suppose that if $r < n - 1$ and $r < u \leq n$, the inclusion probabilities $\pi_k(r, u, c)$ are not so proportional to x_k as in the case when $r = n - 1$ and $u = n$ for all $k = 1, ..., N$.

TABLE 3. The relative efficiency coefficients (%) of the conditional strategies for $P_{n-1,n}(s|k\bar{x})$, $k = 0, 1, 2, 3$. The population with outliers, $N = 284$.

$k\bar{x}$	0			$\bar{x}$			$2\bar{x}$			$3\bar{x}$		
n	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$
2	6.5	1.3	2.0	1.9	8.6	2.2	4.3	0.7	3.2	9.8	5.4	3.2
3	9.2	1.5	1.9	3.5	1.1	2.0	3.9	0.9	3.0	6.3	0.5	3.5
6	18.2	2.3	3.0	11.8	2.0	2.9	5.6	1.5	3.0	5.7	0.9	3.8
9	21.4	2.9	5.5	17.8	2.8	5.4	9.0	2.0	4.4	6.2	1.3	4.8
14	25.6	3.4	13.9	23.2	3.2	13.5	15.3	2.8	10.0	8.4	1.9	7.7
29	34.5	4.5	56.8	34.9	4.6	58.2	29.1	4.4	48.7	19.8	3.6	35.1

The analysis of Tables 3, 4 and Figures 2, 3 leads to the following conclusion. The relative accuracies of the sampling strategies for sampling designs $P_{r,u}(s|c)$ are usually better in case of the population with outliers or extreme values.

The Figures 2 and 3 let us infer that in the case of the population without outliers values the distributions of the estimators are almost symmetric. In the case of the population with outliers values the estimators are distributed with a large number of outliers. In this situation we cannot expect an accurate estimation. But the analysis of the Figures 2 and 3 lets us say that in the case of the population with outliers the conditional sampling design (for $c > 0$) leads to reduction of the number of outliers observations in the distribution of appropriate estimators. The conditional strategies which depend on the sampling design $P_{r,u}(s|c)$ are usually more accurate than their appropriate unconditional versions (for $c = 0$). The estimators of the conditional strategies are unbiased or negligible biased estimators of the population mean. When the conditioning value c increases (the considered levels: $c = 0$, $c = \bar{x}$, $c = 2\bar{x}$, and $c = 3\bar{x}$) the accuracies of the strategies $(\tilde{y}_s, P_{r,u}(s|c))$ and $(\bar{y}_{HT,s}, P_{r,u}(s|c))$ usually increase, too. The accuracy of the strategy $(\tilde{y}_s, P_{r,u}(s|c))$ is the best among the conditional ones. The strategy $(\bar{y}_{r,u,s}, P_{r,u}(s|c))$ is not worse than $(\bar{y}_{HT,s}, P_{r,u}(s|c))$ except for the case of the population with outliers and of the sample size $n = 29$.

TABLE 4. The relative efficiency coefficients (%) of the conditional strategies for $P_{n-1,n}(s|k\bar{x})$, $k = 0, 1, 2, 3$. The population without outliers,$N = 281$.

$k\bar{x}$	0			$\bar{x}$			$2\bar{x}$			$3\bar{x}$		
n	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$
2	17.7	2.3	2.7	3.8	1.6	2.1	17.6	2.5	1.2	32.3	3.0	0.8
3	24.5	2.3	3.1	8.0	1.7	3.0	15.9	2.0	2.2	14.3	2.4	1.9
6	41.8	2.5	12.7	30.3	2.2	11.8	17.1	2.2	2.3	14.8	2.3	1.8
9	51.8	2.8	25.3	50.6	2.8	22.8	26.3	2.3	16.8	26.8	2.3	11.3
14	63.8	3.1	44.6	59.9	3.0	39.6	39.9	2.8	28.7	36.4	2.2	22.1
29	81.7	3.1	84.3	75.0	3.1	82.9	71.8	2.9	69.1	56.9	2.9	52.4

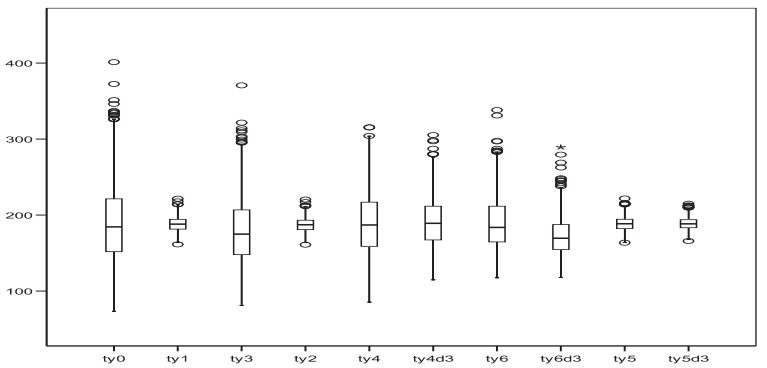


FIGURE 2. Boxplot of the estimator distributions in the population without outliers for n=14

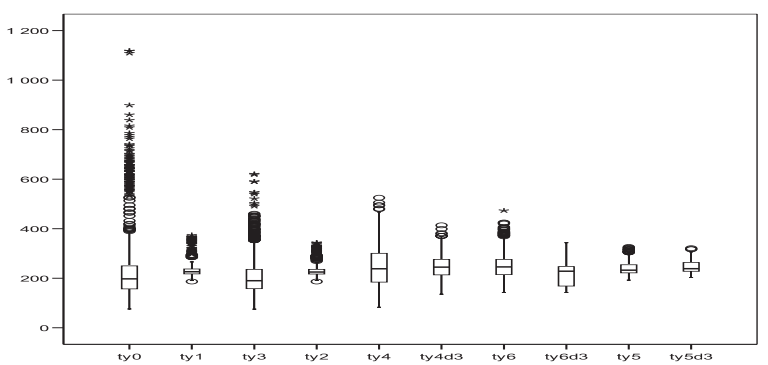


FIGURE 3. Boxplot of the estimator distributions in the population without outliers for n=14

6. Conclusions

The inclusion probabilities of the conditional sampling design proportionate to the sum of two order statistics are presented. They let us determine the variance of the Horvitz-Thompson estimator as well as its estimate.

The simulation analysis lets us expect the estimation strategies with the sampling design $P_{n-1,n}(s|c)$ not to be less accurate than the strategies with sampling design $P_{r,u}(s|c)$.

In general, the accuracies of the considered ratio type strategies: $(\tilde{y}_s, P_{n-1,n}(s|c))$, $(\hat{y}_s, P_1(s))$ or $(\hat{y}_s, P_0(s))$ are the best among the all strategies considered in the analysis. The accuracies of these three strategies are comparable. Moreover, the conditional strategies for $c = k\bar{x}$, $k > 0$ are slightly better than the appropriate unconditional ones.

Let us underline that the above conclusions cannot be treated as sufficiently general because they have been derived on the basis of a partial computer simulation analysis based on special data set taken into account. But it seems that those results can be an inspiration for larger simulation analyses or studies of the formal properties of the sampling design.

Acknowledgement

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

REFERENCES

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

LAHIRI G. W., (1951). A method for sample selection providing unbiased ratio estimator. *Bulletin of the International Statistical Institute,* Vol. 33, pp. 133–140.

MIDZUNO H., (1952). On the sampling system with probability proportional to the sum of sizes. *Annals of the Institute of Statistical Mathematics,* Vol. 3, pp. 99–107.

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

SEN A. R., (1953). On the estimate of variance in sampling with varying probabilities *Journal of the Indian Society of Agicultural Statistics*, **5**, 2, pp. 119–127.

TILLE´ Y., (1999). Estimation in Surveys Using Conditional Inclusion Probabilities: Complex Design. *Survey Methodology*, Vol. 25, No 1, pp. 57–66.

TILLE´ Y., (2006). *Sampling algorithms*. Springer.

WYWIAŁ, J. L., (2007). Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design proportionate to the order statistic of an auxiliary variable. *Statistics in Transition-new series,* Vol. 8, No. 1, pp. 125–137.

WYWIAŁ, J. L., (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers,* Vol. 49, No. 2, pp. 277–289.

WYWIAŁ, J. L., (2009). Performing quantiles in regression sampling strategy. *Model Assisted Statistics and Applications,* Vol. 4, No. 2, pp. 131–142.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 249–272

THE EFFECT OF UNEMPLOYMENT BENEFITS ON LABOUR MARKET BEHAVIOUR IN LUXEMBOURG

Nicholas T. Longford[1], **Ioana C. Salagean**[2]

ABSTRACT

We apply the potential outcomes framework to estimate the effect of awarding unemployment benefits on gaining long-term employment after an unemployment spell and on the time it takes to achieve it. We conclude that such awards, regarded as a treatment, are associated with poorer labour force outcomes than no awards.

Key words: administrative register, labour market, propensity matching, unemployment spells.

1. Introduction

The effect of unemployment benefits is extensively studied in the labour-market and other economics literature (Hunt, 1995; Roed and Zhang, 2003; Lalive and Zweimüller, 2004; and Uusitalo and Verho, 2010). The research is stimulated by high expenditure on these benefits, especially at times of higher unemployment. Such expenditure is regarded by some as an investment; the desirable return on it is stable and long-term employment after a short spell of unemployment, possibly with some training and short-term subsidised employment. The key issue is whether and to what extent the benefits have such an effect, and how the rules for awarding benefits should be adjusted to optimally combine the roles of welfare support while encouraging the resumption of employment and being frugal with public funds (Lalive, van Ours and Zweimüller, 2006). Jobs that are created as part of active labour market policies, intended as short-term stop-gaps, or are subsidised by other means, are not regarded as successful outcomes in this context, although they may be effective conduits for such outcomes.

[1] SNTL and Department d'Economia i Empresa, Universitat Pompeu Fabra, c/ Ramon Trias Fargas 25 – 27, 08005 Barcelona, Spain. E-mail: sntlnick@sntl.co.uk.
[2] CEPS/INSTEAD, Avenue de la Fonte 3, L–4364, Esch-Belval, Luxembourg. E-mail: ioana.salagean@ceps.lu.

The methodological challenge related to estimation of the effect of benefits is the impossibility of directly observing the labour market behaviour of two persons identical in all relevant aspects, including their background, who have just become unemployed – one who receives benefits and the other one who does not. This is compounded by the rule-based nature of awarding benefits. In principle, two persons with identical background at the point of becoming unemployed would be treated identically, and either both or neither of them would receive benefits. Nevertheless, we have a clear conception of what is meant by the effect of the benefits, as the difference in the individual's future position in the labour market (employment status and security, income, and the like) under the two conditions: receiving benefits and not receiving them.

The problem is addressed by models that account for the systematic differences between those who receive benefits (the 'treated', a term motivated by the medical statistics literature) and those who do not. An alternative approach is based on matching (Rosenbaum, 2002) – selecting subsets of recipients and non-recipients that for all purposes appear as if they were assigned to these two treatment groups completely at random. Adjustment by regression uses all the data, exploits a powerful modelling framework aided by methods of model selection, but is associated with numerous caveats related to model assumptions. One of them is the assumption of constant (universal) difference in the outcomes for the treated and untreated units, after a suitable adjustment for covariates and allowance for measurement error. In our case, there is no measurement error but the assumption of treatment homogeneity is not realistic.

Influential observations are another concern. They are extreme or outlying observations in the space of the covariates and their removal may alter the model fit much more than the removal of an observation closer to the centre of gravity of the observations would. However, such observations are often least relevant to the comparison of the treated and not treated units, because these two groups usually differ in the tails of their distributions of the covariates (Crump *et al.*, 2009). In estimation with large datasets, our focus should be on reducing the bias of estimators because their variances are small. Model selection procedures balance the effort to reduce the bias (retain more complex models) and variance (reduce model complexity). In contrast, methods based on matching emphasise bias reduction (Rubin, 2006). For discussion of related issues, see Lechner (2002); Hirano, Imbens and Ridder (2003); and Abadie and Imbens (2006).

Another reason for preferring the analysis by matched pairs is that some of the outcome variables we define are not on scales that can be associated with any common distributions. In some cases, the values of the outcome variable permit only partial ordering. That is, the difference of the outcomes cannot be quantified for some pairs of units, or we do not want to commit ourselves to any particular scale on which such differences could be quantified. We prefer to form matched pairs and tabulate the comparisons of their outcomes. There are three possible elementary results for a pair:

- the treated unit has a superior outcome;

- the treated unit has an inferior outcome;
- there is a tie – the units have outcomes that we regard as identical, or we cannot resolve which outcome is superior.

We say that the treated unit is a winner (and the untreated unit in the pair a loser) in the first case, and a loser in the second. The *balance* in a set of matched pairs is defined as the difference of the number of winners and losers in the treated group. We use the balance as an estimator of the average effect of the treatment. We then decide whether the realised balance could differ from zero purely by chance or it reflects the sign of the average treatment effect among the (treated) units that receive unemployment benefit.

The number of ties in the matched pairs also has a role in the analysis. If there are many ties, a more refined definition of the outcome variable may resolve some (or even most) of these ties, resulting in a substantially altered balance of the winners and losers. Desirable are results in which the number of ties is much smaller than the balance, so that the conclusion would not be altered if all the ties were redefined as winners (or losers). In this way, the ties become the element of a sensitivity analysis. Of course, the ties cannot be resolved in some settings when the outcomes are genuinely identical or not comparable.

For a large population, we might associate these three counts with a trinomial distribution. How- ever, our target is the assessment of the treatment effect for a particular set of units – the units that have been treated. We seek to establish what would happen if those who received benefits were denied them. Simplistic views include the suggestion that receiving benefits reduces the urgency of job search, but increases its quality – the recipient is more selective (discriminating) among the options for new employment and more patient in committing him- or herself to a new job. We test this hypothesis by matched-pairs analysis.

The dataset in our analysis is extracted from the databases of the Luxembourg Unemployment Agency (*ADEM – Agence pour le développement de l'emploi*) and the Register of National Insurance Contributions of Luxembourg (IGSS – *Inspection générale de la sécurité sociale*). The extracts are monthly lists of unemployed registered at ADEM, together with some background information, and labour force states at the end of each month inferred from the IGSS database. The lists cover the period from January 2007 to July 2011 (55 months). The records in these lists are linked by the (unique) national insurance number. The lists are reformatted to a dataset in which a record contains the sequence of monthly labour force states, supplemented by background information about the individual. A record is associated with an unemployment spell, and so an individual may have several records in the dataset. Such records are not exact duplicates, because some information, such as level of education and marital status are temporal, especially for younger people. Also, the age of the person at the time of becoming unemployed is an important variable.

The units of the analysis are unemployment (U) spells qualified by the period of the study (January 2008 – December 2010, 36 months) and the age of the

person at the beginning of the U spell (up to 30 years). The information from 2007 is reserved for defining covariates and the information from 2011 (part) for defining outcome variables. The U spell that defines the unit is called *reference*. Apart from the treatment, the reference spell is also associated with a person, the month when it starts and its length (in months) or the month when it ends. The treatment (award of unemployment benefits) is applied to the reference spell. Other U spells of the same person may be treated differently.

The next section introduces a terminology for sequences and gives more details of the ADEM and IGSS databases. The following section discusses the method of analysis based on the potential outcomes framework (Rubin, 1974; Holland, 1986). Section 4 gives further details specific to the matched-pairs analysis. The application to the resolutions of unemployment spells of young members of the labour force in 2008 – 2010 is discussed in Section 5.

2 Discrete sequences

A record in the original dataset is the sequence of $K = 55$ labour force states of a person. An element of the sequence corresponds to a calendar month, and the sequence to a contiguous set of calendar months, from January 2007 until July 2011. The status has five possible values: employment (E), unemployment (U), economic inactivity (I), transition (T) and absence (A). For example, a person who has been continually employed over a period of K months has the sequence EE ... EE of length K; a person who completed his/her education and found a job three months later, has sequence AA ... AUUUEE ... . Status A in month m results from having a record in neither ADEM nor IGSS at any time prior to and including month m. The complete definitions of the states are given in Appendix.

A sub-sequence is defined as a sequence for a contiguous subset of months. A sequence comprises *spells*. A spell is defined as a sub-sequence that is composed of the same status throughout, and the states at the immediately preceding and succeeding time points are different. A spell is characterised by the status and its length. For example, the sequence of the first person, continually employed, comprises a single spell of E. This spell has length greater than or equal to K, because the person may have been employed in the period immediately preceding the beginning of our records and following the end of the records. The sequence of the second person comprises three spells: A, U and E. The second spell is U and has length 3. Suppose the other two spells of the person have lengths 17 (A) and 35 (E). Then the sequence is completely described as (17A, 3U, 35E). A person may have a more complex description of the sequence, with many spells and several spells in the same state, such as (5A, 4U, 1T, 6E, 2U, 3E, 3I, 1T, 4E, 22E, 4I). The order of the spells is essential, and we do not summarise the sequence by the totals of time points (months) spent in a state, such as 31 in E, 6 in U, and so on. In fact, the longest spell of E, or of U, is of

interest in some analyses. The first and last spells in a sequence are often *censored*. That is, the first spell may have started before the first time point (month) of the sequence. Similarly, the last spell is incomplete if it continues into the future beyond the data horizon (month $K = 55$ in our data). In some instances, the first spell is complete. For example, in the next section we study sequences that start with (the first month of) a U spell.

The *history* of a spell (of a sequence) is defined as a sub-sequence that ends in the month immediately before the beginning of the spell. The history is qualified by its length. For example, suppose a spell of a sequence starts in January 2008. Then the sub-sequence for the 12 months of 2007 is its history of length 12. The future of a spell is defined similarly, although we include the defining spell in the sub-sequence. For example, the future of length 19 of a spell that lasted from January 2010 until February 2011 is the sub-sequence from January 2010 until July 2011.

We distinguish between sets of sequences that have the same delimiting (starting and ending) points (months), such as January 2007 and July 2011, and sets in which the start (or the end) is triggered by an event, such as the beginning of a U spell. In the latter case, some selection may take place, as there may be units (persons) whose sub-sequences do not qualify to the set. In fact, the selection takes place even in our original dataset, because persons with no U spell before their 31st birthday are not included. Further, we draw a distinction between sets of persons (individuals) and sequences. In the former, each person from a given domain is included at most once. In the latter, a person may be included several times. For instance, we consider in the Sections 4 and 5 sets of U spells. A person with several qualifying U spells is included once for each spell. Such spells of a person have different starting months, but differ also in length, as well as in their histories and futures.

2.1. ADEM and IGSS databases

The sequences of labour force states of length 55, covering the period from January 2007 until July 2011, are defined for all the persons who had a U spell in this period (had a case file open in ADEM) and their age was in the range 15–30 years at the beginning of one of their U spells in the period. We focus on the young because many of them who have had a U spell tend to have complex sequences, whereas among older members of the labour force U spells tend to be longer and are often followed by long spells of E or I. Also, the young are substantially over-represented in the U spells; about 43% of all ADEM records are for persons aged 30 or below at the beginning of the U spell.

We combine the information extracted from ADEM records, comprising the beginning and end of each U spell with information from IGSS, from which we infer the states in the other months (E, I, T, A); see Appendix for details. The

sequences are supplemented by background information listed in Table 1. The choice and purpose of the transformed variables and interactions, listed at the bottom part of the table is explained in Section 5. There are 43 825 reference spells (units of analysis) involving 26 835 persons (12 541 women and 14 294 men, with 19 328 and 24 497 spells, respectively). A person has a sequence in the dataset for every U spell in the studied period, subject to the condition of being aged 15 – 30 years at the beginning of the U spell.

The 55-month sequences contain a total of 2.410 million time points; 1.036 million of them are in state E, 351 000 in U, 452 000 in I, 82 000 in T and 489 000 in A. When reduced to unique persons, after discarding duplicate records, the number of time points is $26\,835 \times 55 = 1.476$ million, with 649 000 E's, 169 000 U's, 256 000 I's, 38 000 T's and 365 000 A's.

We discard all the records with reference (U) spells in 2007 and 2011 because the former do not have a history of one year, which we want to use for defining background variables, and the recorded future (up to July 2011) of the latter is too short. Thus, our analysis is reduced to 24 040 reference spells, of 10 346 women and 13 694 men, with 7874 unique women and 9 507 men. Men have higher average number of U spells (1.44) than women (1.31). Of the selected reference spells, 9 094 are associated with benefits (3982 for women and 5112 for men) and 14 946 are not (6364 + 8582).

The top panel of Figure 1 shows the distribution of the reference spells over time (starting months), separately for spells associated with benefits and those without. Every autumn (September and October) there is a peak in the number of failed applications (with no benefits awarded), and a smaller peak in January. The numbers of successful applications are greater in autumn and early winter, with a sharp fall in February. The following three panels display the distribution of ages of the applicants. They show that failed applications dominate among the young, and men in particular, but from about 24 years of age on the success rate it is close to 50%. It increases slowly with age.

There are 4820 persons with multiple reference spells; the highest multiplicity is seven, in three instances, followed by six, in eleven instances. The multiple reference spells are either treated (successful applications) on all occasions, not treated (failures) on any occasions, or have one or several untreated spells followed by one or several treated spells. There are 3424 persons with two reference spells; 1670 of these pairs are both untreated, 1112 are both treated, 642 have an untreated spell followed by a treated spell. Of the 1048 persons with three reference spells, 530 have three untreated spells each, 299 three treated spells each, 93 have a treated spell only in the third U spell and 126 with a sole untreated spell in the first U spell in the designated period. We note that the persons' histories may include U spells that have not been recorded (before 2007).

Table 1. Covariates used in the analysis.

Variable	Notation	Values	Summary
Age (years)	*age*	15 – 30	Mean 23.7; median 24.0, st. dev. 3.73.
Sex	*sex*	0, 1	0: women (43.3%), 1: men (56.7%).
Marital status	*civstat*	1, 2	1: single (82.8%), 2: other (17.2%).
Nationality	*natio_cat*	1, 3, 7	1: Luxemburgois (44.0%), 3: Portuguese (28.1%), 7: other (27.8%).
Level of education	*edu*	1, 2, 3	1: basic (39%), secondary (46.2%) tertiary (14.8%).
Spell No.	*rk*	1, 2, 3, 4	U spell since Jan. 2007; 1: first (54.1%), 2: second (27.1%), 3: third (11.9%), 4: fourth or later (8.9).
Months since completing education	*dfinetu*	0 – 51.5	Mean 44.3; median 45.5; st. dev. 5.8.
Months since the first record in IGSS	*prem_affil*	22 – 51.6	Mean 43.2; median 44.6; st. dev. 6.8.
Unemployment office (Region)	*ccel*	1, 2, 3, 4	1: Luxembourg City (42.5%); 2: Esch-sur-Alzette (36.5%); 3: Diekirch (14.3%); 4: Wiltz (6.8%)
Employment sector	*cemprec*	1, 2, 4, 5, 8, 9, 10, 11, 12	1: arts and technical; 2: management; 4: sales; 5: agriculture and forestry; 8: crafts and manual; 9: food, chemical, machinery; 10: hotels and restaurants; 11: other services; 12: no profession
History	*pstat*	E, U, I, T, A	12 categorical variables.
Future	*astat*	E, U, I, T, A	25 categorical variables.
Start	*start*	13 – 48	Mean 29.5, median 30.0; st. dev. 9.7.
Duration of the U spell	*dur.UN*	0 – 42	Mean 3.12; median 2.00; st. dev. 4.09.
Transformed variables			
Square of dfinetu	*Sq.dfinetu*		$dfinetu^2$
Square of prem_affil	*Sq.prem_affil*		$prem\_affil^2$
Square of age	*Sq.age*		age^2
Bi-months in status E in 6 months of history	*Pstat.SumE*	0, 1, 2, 3	Categorical variable (0 – 49.9%, 1 – 10.5%, 2 – 10.7% ,3 – 28.9%)
Interactions			
Age × pstat 1 = E	*Age.Pstat1E*		Age by status E in month −1
Age × sex	*Age.Sex*		Age by level of education 2
Age × education (2)	*Age.Edu2*		
Nationality × pstat 1 = I	*Nat.pstat1I*		Nationality by status I in month −1
Region × sex	*Ccel.Sex*		
Education × sex	*Edu.Sex*		

Note: All the summaries are for the reference spells.

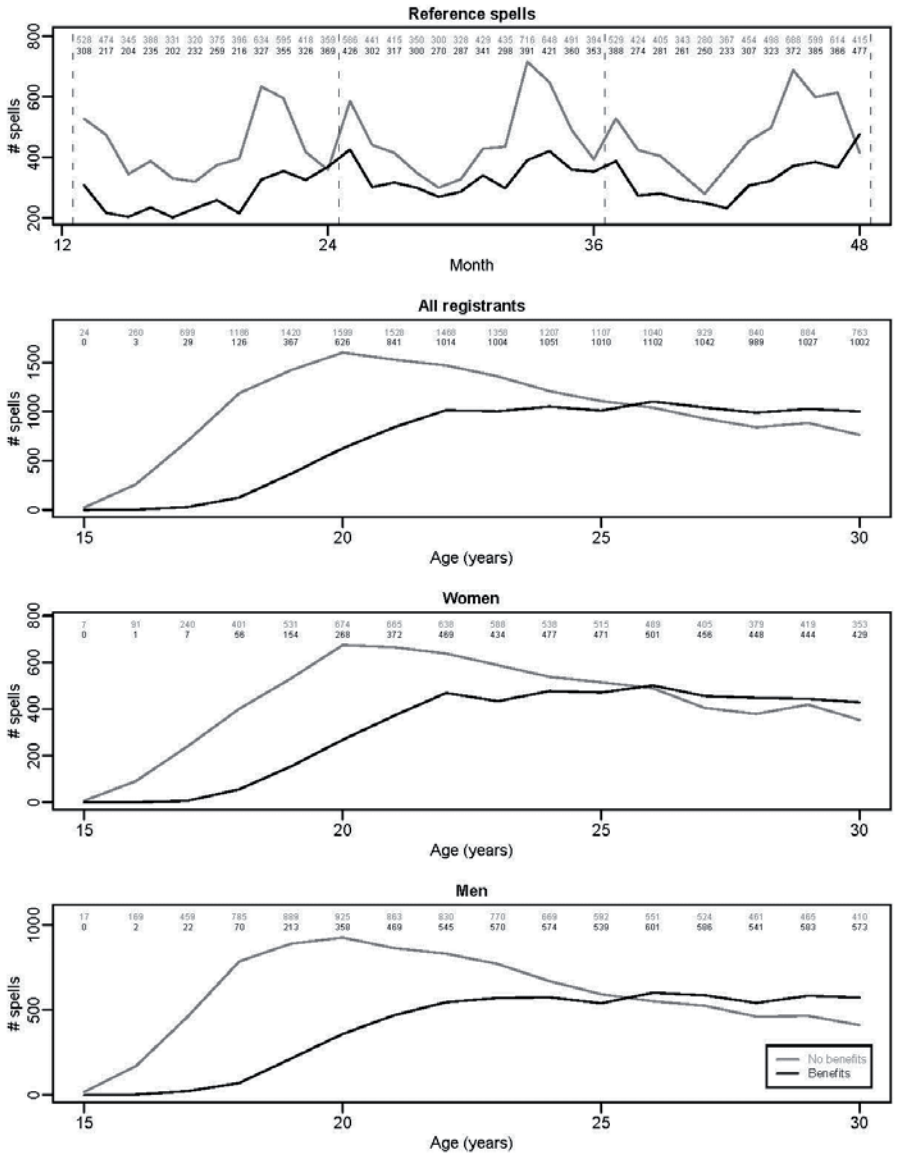


Figure 1. Receipt of unemployment benefits for reference spells, by age and sex.

3. Methods

Our analyses concern the complete enumeration of ADEM registrants in the period from January 2008 to December 2010 (36 months) who were 15–30 years of age at the beginning of a U spell. We have data for an additional year, 2007, but this we reserve for the definition of background variables. The unit of analysis is a U spell, supplemented by its history and future. Thus, the unit is itself a sequence, but a person may be represented in the analysis by several sequences, one for each U spell. Each reference spell is associated with receipt or not of unemployment benefits. We consider the future of the reference spell as the outcome. A person may have U spells (case files in ADEM) additional to the reference, both in its history and future. Two sequences of a person with multiple U spells are linked by histories and futures that are shifted by the distance in time between the two reference spells.

We want to establish how different the futures of the treated reference spells (with receipt of benefits) would be if they were not treated (with no benefits received). Such research questions are regarded as *counterfactual* because the difference considered cannot be observed for any unit. We have no control over the assignment of the benefit status. Hypothetically, if the assignment could have been made at random, applying experimental design, the analysis would be relatively straightforward. Even in that setting we could not establish the difference of the futures of a reference spell under the two conditions – treatment by the benefit, and no treatment – because only one of the conditions could be realised. However, the random assignment, together with a set of assumptions discussed in Section 3.1, would ensure that the average treatment effect is estimated without bias.

In summary, our focal population is the set of all U spells in the designated period that result in a uniquely identified case file (record) and registration in ADEM. The unit of analysis is defined by its reference spell. It has to be distinguished from the other U spells in the history and future of the person's sequence. For the history, we consider the immediately preceding 12 months, and for the future up to 25 months immediately following the start of the reference spell. The futures are censored by the data horizon of July 2011 for all reference spells that started after June 2009.

We distinguish three categories of variables:

- treatment indicators
- outcomes
- covariates

A treatment indicator T is a binary variable with values zero or unity; $T = 1$ for one treatment and $T = 0$ for the other. Every unit is assumed to have been administered one of the two treatments. The only pair of treatments we consider is the receipt of benefits and not receiving them. We elect $T = 1$ for the receipt and $T = 0$ for the complement. A person may have several U spells in the

designated period, between January 2008 and the earlier of July 2011 and the 31st birthday, and each of them is the reference spell of a sequence (the unit of analysis).

We specify several outcome variables that summarise a particular aspect of the future of the reference spell. We regard the resolution (closure) of an ADEM case file that coincides with the reference spell as a success if the future contains an E spell of length at least L. We set $L = 12$, regarding as a landmark uninterrupted employment over a period of one year after a U spell. Insisting on a longer spell, such as $L = 18$, would lead to excessive censoring by the data horizon in July 2011 and too few successes would be recorded. Another variable is the length of the longest E spell in the sequence. Comparisons of the values of this variable for sequences with different starting points are problematic because of the fixed data horizon. However, comparisons for two sequences with the same starting point have face validity. If two sequences have the same maximum length of E spell, then we regard as superior the sequence in which the first E spell of maximum length was achieved earlier. Some other problems (ambiguities) with this definition are discussed in Section 5.

3.1. The potential outcomes framework

An outcome variable Y is associated with two potential outcome variables; $Y^{(t)}$ is the variable defined as the outcome assuming that treatment $t = 0$ or $t = 1$ was applied. Our goal is to compare the values of $Y^{(1)}$ and $Y^{(0)}$ on the set of units (the futures of the reference spells) that received treatment $t = 1$. The variable Y can be expressed as a composition

$$Y^{(T)} = (1 - T)Y^{(0)} + TY^{(1)},$$

and its distribution is a *mixture* of the distributions of $Y^{(0)}$ and $Y^{(1)}$. Inferences about $Y^{(1)} - Y^{(0)}$ are relatively easy to make when T is independent of $Y^{(0)}$ and $Y^{(1)}$, as it would be if the treatment were randomised. Potential versions, associated with the two treatments, can be defined for any variable. Of course, the two versions may be identical. Background variables, defined prior to administering the treatment, which could not have been informed by the value of T, are obvious examples.

We use as covariates an extensive set of variables listed in Table 1. The qualifying attribute for a covariate is that its two potential versions are identical − that its values would not be altered if the treatment assignment (the values of T) were changed. Variables defined prior to the beginning of the reference spell are covariates. We discount the possibility that the history of the reference spell is influenced in any way by the anticipated value of T in any future U spell. The covariates include variables defined on the 12 months of labour market history (labour force states). This is based on a pragmatic decision not to lose in the analysis too many spells for which this history is not completely recorded.

The potential outcomes framework has the assumption of stable unit-treatment variable, referred to by the acronym SUTVA. It can be summarised as follows: in hypothetical replications of the study, with the same set of units, the same sets of values of all variables except the treatment, but a replicate realisation of the treatment assignment process (mechanism), the outcome for unit i is determined entirely by the treatment assigned; $y_i^{(0)}$ if $t_i = 0$ and $y_i^{(1)}$ if $t_i = 1$.

Close scrutiny of this condition reveals that it is far from trivial, and in several aspects contentious in our setting. First, it implies that the units do not interfere with one another. That is, the outcome of one unit is unaffected by the treatment assigned to another unit. In our case, it entails the assumption that the future of one U spell is unrelated to the future of a later U spell of the same person. This is patently false, especially for two short U spells that are separated by a short time. However, the dependence is difficult to describe and relates to a small fraction of the units. Next, the rules for awarding benefits are well known to all parties involved, and so the conduct of the persons threatened by unemployment is affected by the anticipated (possibility of) loss of job. Unemployment often arises in quanta (several persons losing their jobs) and is anticipated. A lot of unemployment arises after the conclusion of fixed-term contracts that cover the same period of time and are awarded to a set of workers at the same time. Every person's behaviour is affected by the experiences of acquaintances, and family members in particular, so the two potential outcomes do not describe the entire range of possible values of the outcome of a person. Further, some unemployed and those about to become unemployed are advised by agents, such as union representatives. As a consequence, the conduct of some unemployed may be coordinated. A principled solution to this problem is to consider all configurations of plausible treatment assignments and define a potential outcome variable for every one of them. However, the number of such variables would proliferate and become unmanageable even in some simple scenarios, so pursuing this approach is not feasible.

3.2. The treatment effect

With two potential outcome variables, the unit-level treatment effect is defined as the difference of the outcomes under the alternative treatments:

$$\Delta Y = Y^{(1)} - Y^{(0)},$$

with value Δy_i for unit i. We impose no conditions on the distribution of this variable. In particular, we do not assume that ΔY is constant. The average treatment effect for a set of units u is defined as the mean of the (unit-level) treatment effects:

$$\Delta \bar{Y}_u = \frac{1}{n_u} \sum_{i \epsilon u} \Delta y_i,$$

where n_u is the number of units in u. The average treatment effect is qualified by the set u. In our analysis, u is the set of U spells for which $T = 1$ was applied.

Another factor relevant to our analysis is the treatment assignment process (mechanism). It is defined by the joint distribution of the treatments assigned to the units in u. In a typical observational study this distribution is not known, and inferences about it are difficult to make because we have only one realisation of this distribution – the realised assignment of the treatments to u. It is more practical to think about treatment assignments in alternative replications of the study. We cannot rule out the possibility that certain units would never receive a particular treatment. For them the unit-level treatment effect ΔY is not defined. It might be constructive to exclude such units from u because the underlying question (What is the difference...?) about them is meaningless.

3.3. The missing-data perspective

If the values of $Y^{(0)}$ and $Y^{(1)}$ were observed for all units in u, the analysis would be straightforward, evaluating the mean of the differences in (1). In the established terminology for missing data, the set of $n_u \times 2$ values of the potential outcomes $Y^{(0)}$ and $Y^{(1)}$ is referred to as the *complete dataset*. The observed dataset, comprising a subset of n_u values (determined by the realised treatment assignment), is called the *incomplete dataset*. The set-difference of the two datasets is the set of *missing values*.

An analyst's first instinct might be to impute a value for each missing item. Such an imputation (data completion) results in a *completed* dataset. The theory of multiple imputation (Rubin, 2002), documents the problematic nature of this approach. If we analyse a completed dataset as if it were the complete dataset, we obtain inferences that project too much confidence because in the analysis we pretend to have more information than was collected. The problem does not arise when the statistic of interest, such as an estimator, is a linear function of the missing values. However, we also want to assess the quality of the estimator by estimating its sampling variance (or standard error), and that is a distinctly nonlinear function of the missing values. Note also that we have to distinguish between the sampling variance of the estimator that relates to complete data and the sampling variance for the combination of the processes of data generation and nonresponse (missingness). The former sampling variance is zero because we study a fixed set of units and assume their potential outcomes to be fixed.

Multiple imputation is a flexible alternative for addressing this problem. We generate several sets of plausible values for the missing items according to a model specified for them. In the imputations, we reflect the uncertainty about the model parameters, as well as the uncertainty that would be present in a model even if all its parameters were known. In our setting, the treatment assignment is the sole source of uncertainty.

4. Matched pairs analysis

Our analysis proceeds by the following steps. First we estimate the propensity of receiving a treatment as a function of the background variables. Next we form pairs of units, one treated and one not treated, that have similar propensities. Then we estimate the average treatment effect by averaging the within- pair contrasts (comparisons). In the final step, we estimate the sampling variance of this estimator by replicating the processes of matching and estimation. The following sections give details of the steps.

4.1. Propensity scoring

In this step, we fit a logistic regression model in which the outcome variable is the treatment applied and the covariates are all the variables listed in Table 1. The propensity is defined as the conditional probability of treatment 1 given the values of the covariates:

$$p(\mathbf{x}) = \mathrm{P}(T = 1 \mid \mathbf{X} = \mathbf{x}).$$

The propensity for unit i, with the vector of covariates $\mathbf{x}_i$, is defined as $p(\mathbf{x}_i)$. The propensities are used for forming pairs of units (reference spells), one with $T = 0$ and the other with $T = 1$, by matching on their values of $p(\mathbf{x})$. That is equivalent to matching on a (strictly) monotone transformation of the propensities.

In a perfectly implemented experiment, the propensities are known. For example, if units are assigned to the two treatment groups completely at random with probabilities equal to 0.5, then $p(\mathbf{x}) = 0.5$ for all $\mathbf{x}$. Problems arise when the protocol of the experiment is not adhered to the letter, as when some units depart from the assigned treatment regimen or drop out from the study.

Our approach can be described as selecting from the observed units a subset that resembles as closely as can be arranged, in all features that can be checked, a dataset generated by a (perfectly executed) experimental design. The main criterion for this is that the two groups in the selected subset are balanced – have near-identical profiles of the covariates. When there are several covariates, such a balance is very difficult to arrange. Rosenbaum and Rubin (1983) showed that arranging this balance can be replaced by the task of matching on the values of the propensity or its monotone transformation. Using the logit (log-odds) scale for the scores is an obvious choice. The logit is defined as $\mathrm{logit}(p) = \log(p) - \log(1 - p)$.

The propensity is a variable defined on the interval [0, 1]. It usually attains many distinct values, most of them with small frequency or even uniquely, and therefore matching on its values exactly would yield too few pairs. In practice, the range [0, 1] is split into a number of intervals and matches are sought within them. Further, the propensity is not established with precision, but merely

estimated. Rubin and Thomas (1996) showed that the uncertainty about the propensity can be ignored, and matching based on the estimated propensity (or its monotone transformation) is sufficient.

Apart from the propensity score, we match also on the date when the U spell starts. This is referred to as *blocking* – matching separately within subsets of the units; in our case, the subsets are defined by the starting date (month). This is essential for some of the outcome variables which are affected by censoring and for which a comparison of two units with different starting dates is problematic.

4.2. Matching and complete-data analysis

We divide the fitted propensities into $H = 10$ intervals so that each interval contains approximately the same number of units. These intervals are further subdivided into groups according to the starting date of the reference spell. There are up to $H \times H'$ matching groups, where $H' = 36$ is the number of distinct starting dates (months). A combination of intervals and dates may contain no units or only units with one value of T. Within every one of the groups in which both treatments are represented, we form matched pairs by the following process. Suppose group h has n_{ht} units with treatment $t = 0, 1$. If $n_{h0} \leq n_{h1}$, then we match each untreated unit ($t = 0$) in the group with a treated unit ($t = 1$), selected at random and without replacement. This is implemented in practice by sorting the n_{h0} untreated units, selecting n_{h0} units sequentially from the set of n_{h1} treated units in the group, and forming pairs in the obvious manner.

Let y_{htj} be the value of the outcome variable for the unit which represents treatment t in pair j of group h. For each pair we evaluate the sign of the contrast $\Delta y_{hj} = y_{h1j} - y_{h0j}$, or the sign (positive, negative or tie) is established without evaluating Δy_{hj} when the difference is not well defined. The results are then tallied across the pairs, to obtain the counts of winners, losers and ties. These tables (triplets), denoted by Δy_h, are further tallied over the matching groups, to obtain a single triplet Δy of counts of pairs in which the representative of the treatment is the winner, loser, or there is a tie.

As an alternative the triplets can be tallied with weights that account for the treated units that have not been matched. If group h has fewer untreated than treated units, $r_h = n_{h1}/n_{h0} > 1$, then $n_{h1} - n_{h0}$ treated units are not matched. We compensate for their loss by increasing the contribution of group h from Δy_h to $r_h \Delta y_{0h}$, where $r_h = n_{h1}/n_{h0}$. For groups in which $n_{h0} \geq n_{h1}$, the contribution to Δy is not changed; r_h is set to unity. There may be one or several groups in which there are no untreated units, $n_{h0} = 0$, so none of the treated units are matched. In every matching exercise, we monitor the number of such units. If it is excessive, then we revise the definition of the matching groups. With fewer groups defined, fewer such failures to match are likely.

The number of treated units involved in such failures can be included in a sensitivity analysis. Consider the following two extremes for the set of all treated units without a match:

1. if they were matched they would all be winners;
2. if they were matched they would all be losers.

If in both of these scenarios we obtain the same inference, then the failures are immaterial. Of course, if there are many such failed matches, then we are threatened with an impasse, when different conclusions are arrived with the two extreme assumptions. The ties between the units in a pair can be dealt with similarly.

4.3. Sampling variation

Regarding the set of focal units (spells) u for which we want to estimate the average effect of the treatment as fixed, the entire process of generating the propensity scores, matching and evaluation entails a single source of uncertainty, namely, the forming of matched pairs. The logistic (or another) regression used for generating the propensity scores would be associated with variation *if* we considered the focal set of units and their assignment to treatment as a realisation of a random process. However, our analysis is conditioned on the treatment assignment and the units included in the analysis, and so each propensity score is without any sampling variation.

We estimate the sampling variation associated with matching by simply replicating it several times. When we do not want to refer to a scale for the outcomes we evaluate the variances of the counts of positive and negative within-pair contrasts, and make inferences about the expectation of the average contrasts, $E(\Delta y)$.

5. Application

We implemented the matched pairs analysis and associated diagnostics in a customised set of R functions. In the first step, we fit a logistic regression of the treatment indicator on all the covariates. The regression parameter estimates or the quality of the fit are of no interest because the sole purpose of the fit is to form matched pairs based on the estimated propensities. We have a lot of observations, 24 040, so we are concerned principally about bias resulting from poor matching. That is why we prefer to err on the side of specifying a richer propensity model; little efficiency is lost by including in the model a few redundant covariates. The key criterion for the appropriateness of the model is that the pairs matched on propensities are also balanced with respect to their distributions of the background variables.

The propensity model with the original variables, involving 71 parameters, 55 of them related to the history of the reference spell, turned out to be unsatisfactory, because the pairs formed were poorly matched on the dispersions of several continuous variables. We supplemented the model with the covariates listed at the bottom of Table 1, involving 16 further parameters. The transformations and interactions were identified principally by trial and error. Note that the addition of a covariate to the model does not always result in an improved balance of the other variables, and may even be detrimental to the balance of some of them.

There are 14 946 units without benefits (untreated) and 9 094 treated units. The distribution of the fitted propensities $\hat{p}$ within the treatment groups is displayed in Figure 2. The diagram shows that a lot of untreated units have very small propensities. Only a small fraction of them will be used in matches with the far fewer treated units that have similar values of the fitted propensity. Most treated units with $\hat{p}$ in the range 0.1–0.5 will be matched because there are sufficiently many untreated units within every narrow range of propensities (the width of a bar in the histogram). For treated units

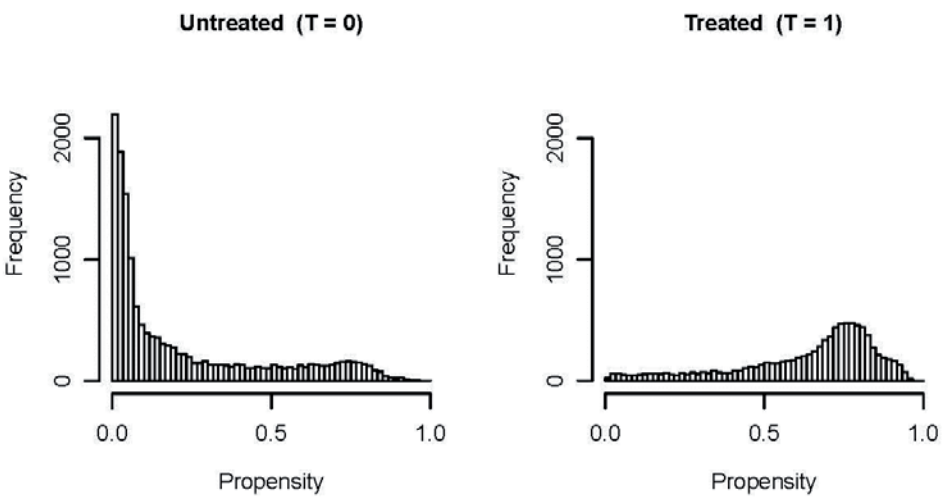


Figure 2. The fitted propensities for the two treatment groups. The vertical dashes mark the deciles of the within-group distributions.

with $\hat{p} > 0.5$, a substantial percentage of treated units will not be matched because they are in a majority.

We match on the fitted propensity and the start (month) of the reference spell. The propensities are split by their deciles into ten groups with approximately equal numbers of units in each. Each of these groups is then split into subgroups

according to the start of the U spell. There are 36 distinct months (13–48). Of the 360 combinations of propensity group and month, 40 groups contain no match (one or both treatment groups are not represented in it), 21 groups have one matched pair each, 19 groups have two pairs, and the largest numbers of pairs matched in a group are 47 and 48, in one instance each. In total, 4631 matched pairs have been formed; they contain 50.9% of the treated units.

The set of matched pairs is satisfactory if they are close to balance, as they would be in an experiment with random treatment assignment. The left-hand panel of Figure 3 displays the contrasts of the means of the continuous variables for the two treatment groups. A horizontal segment is drawn for each variable between the contrast for the unmatched (original) groups and its negative. The value of the contrast is marked by a vertical tick. The contrast for the matched pairs with the original model is marked by a gray disc and the contrast for the extended model by a smaller black disc. The

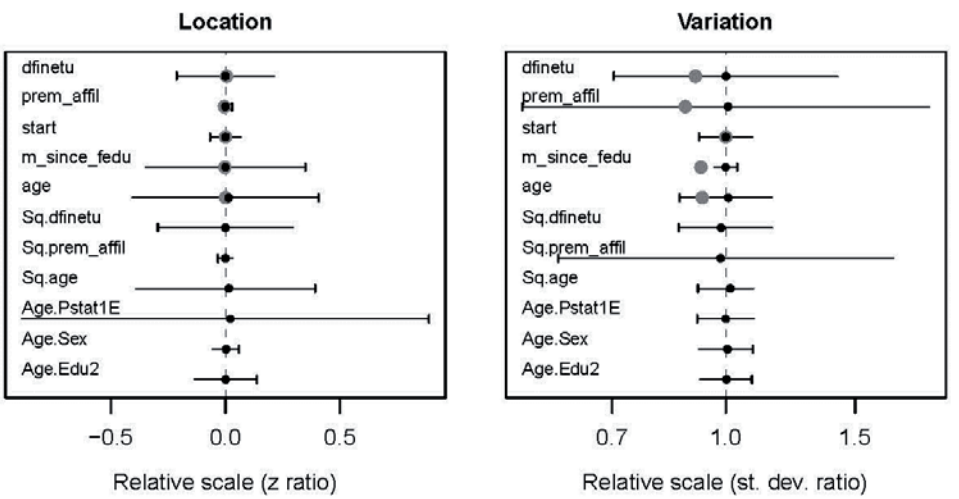


Figure 3. Check on the balance of the continuous covariates for the matched pairs.

gray discs are drawn only for the five variables in the original model. Owing to blocking, the balance for *start* is perfect by design. Note that we broke some rules of invariance, such as not including both variables defined by the interaction of age and the three categories of education. If we followed the rules, a slightly worse balance would be obtained.

The matched pairs should be balanced not only in the within-group means, but in the within-group distributions in general. The right-hand plot displays the ratios of the standard deviations of the continuous covariates for the unmatched and matched groups. The horizontal segments are drawn between s and $1/s$, where s is the ratio for the unmatched groups. The ratios for the matched groups are marked by gray and black discs for the original and extended propensity models,

respectively. The horizontal axis is on the log-scale. The ratios for the matched groups deviate from the ideal of 1.0 only slightly, and the extended model yields a better balance for all four original variables (except for *start*, for which the balance is perfect by design). The balance is nearly perfect for the six added variables.

Figure 4 presents the corresponding plot for the contrasts of the categorical variables. A variable with C categories is represented by $C - 1$ horizontal segments connecting the unmatched balances with their negatives. The black and gray discs mark the matched balances based on the original and extended propensity models. The matched balances are uniformly closer to zero for both the original

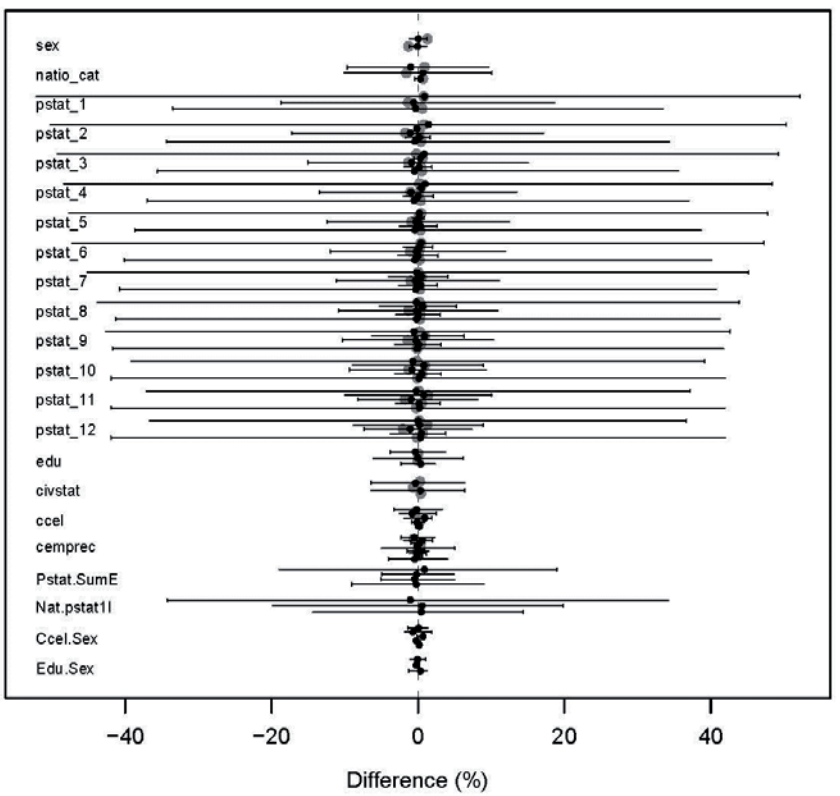


Figure 4. Check on the balance of the categorical covariates for the matched pairs.

model and for the model with four additional variables (15 additional parameters), displayed at the bottom of the diagram.

Having accepted the match as satisfactory, the remainder of the analysis entails comparing the numbers of winners and losers in the within-pair contests. We score the future of each unit as a success when it contains a 12-month (or longer) spell of E, and as a failure otherwise. In a matched pair, the treated

unit is a winner if it is a success and the untreated unit is a failure. The treated unit is a loser if it is a failure and the untreated unit is a success. The result is a tie if the outcome is the same for both units in the pair (successes or failures).

With these rules for scoring, the treated unit is a winner in 831 pairs and a loser in 1169 pairs, and 2631 pairs are tied. Thus, the estimate of the treatment effect is negative, with a balance of 338 units in favour of the untreated group. Some of the ties can be resolved if we compare the lengths of the longest spell within the pairs that are successes. Then the treated group has 1509 losers, 976 winners and there are 2146 ties. The balance, 533, is increased substantially. If we compare the lengths of the longest spells without the qualification of L = 12 months, the treated group has 2434 losers, 1698 winners (positive balance of 736 units) and there are only 499 ties; the treated group would have more losers (237) even if all the ties were conceded to the untreated group.

We can compensate for the failure to match some treated units by weighting the within-pair contrasts. If in matching group h there are more treated units than untreated, $n_{1h} > n_{0h}$, we apply weights $r_h = n_{1h}/n_{0h}$ to each within-pair contrast formed in this group. Otherwise, when all treated units are matched, the weight r_h is set to unity. When an E spell of at least 12 months in the future is regarded as a success, the treated group has score 1582.1 and the untreated group 2563.7, and the ties account for 4948.2 points. The total of these three scores is 9094. It coincides with the number of treated units in the (unmatched) sample, because there happen to be no matching groups in which all the units are treated. For such a group, with n_{0h} = 0, the weight r_h would not be defined. The balance, 2563.7 − 1582.1 = 981.6, is much greater than its counterpart without weights, 338. However, the sampling variation of the balance is increased because some within-pair contrasts have large weights. The largest weights are 13.5 for four pairs and 12.4 for nine pairs each, and further 190 pairs have weights greater than 5.0. In total, 2538 pairs (54.8%) have weights greater than unity; 1046 of them are smaller than 2.0. Weights greater than unity are for pairs from matching groups in which treated units have a majority. For the other two outcome variables that score the future of the reference spell, the treated group has far fewer winners than losers and the balance is much greater than in the analysis without weights. The results are summarized in Table 2. The rows are for the analyses in which only E spells of one year or longer count as successes (analysis A), only such spells count but two such spells are compared by their lengths (analysis B), and when the longest E spell of the future counts (even if shorter than one year), and these maxima are compared by their lengths and recency (analysis C).

By comparing the two treatment groups without matching, we obtain the following results. In analysis A, 17.7% of the treated units have an E spell of length 12 or longer in their futures, compared to 18.9% of the untreated. In analysis B, the mean scores of the treated and untreated groups are 3.40 and 3.59, respectively. In analysis C, the treated group has mean 4.54 (months) and the untreated

Table 2. Matched-pairs analysis of the employment futures

	No weights			Weights		
	Winners	Ties	Losers	Winners	Ties	Losers
A. E spell length ≥ 12 months	831	2631	1169	1582.1	4948.2	2563.7
B. E spell length ≥ 12 months +	976	2146	1509	1925.8	3848.4	3319.8
C. Longer E spell (earlier)	1698	499	2434	3275.0	839.9	4979.1

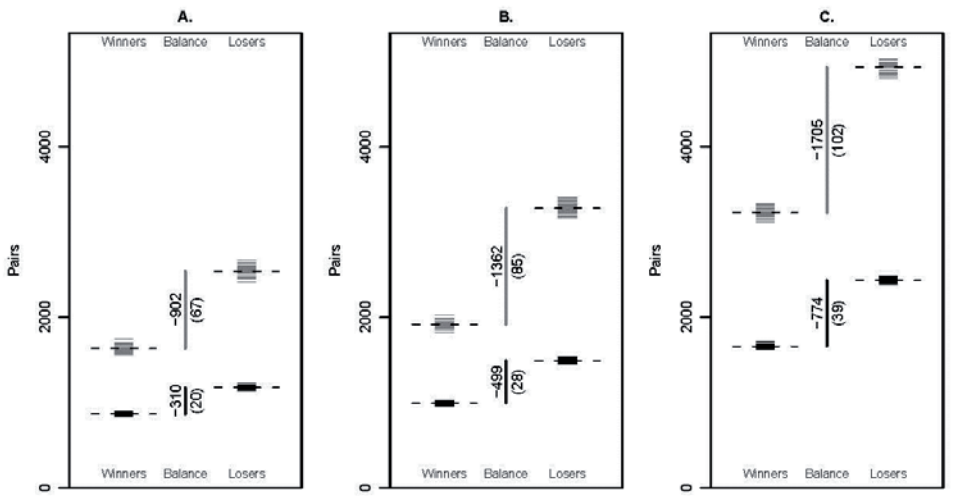


Figure 5. Summary of a set of 50 replicates of the analyses in Table 2.

4.94. Thus, the matched-pairs analysis is in agreement with the 'raw' comparison for all three outcome variables.

The sampling variance of the balances is estimated by replicating the matching process. A set of 50 replications is summarised in Figure 5. The panels A – C correspond to the rows of Table 2. The replicate counts of winners and losers are marked by thin horizontal segments, black for unweighted and gray for weighted analysis. Their means are marked by horizontal dashes and the balances are indicated by vertical segments with their values printed to the left and standard errors to the right in parentheses.

If the estimates of the balance and their standard errors are taken at face value, there is very strong evidence of negative balance for all three outcome variables – benefits are associated with less desirable outcomes of reference spells

(E spells, if any, that are shorter and achieved later). The checks on the balance of the covariates in Figures 3 and 4 are not a complete diagnostic because they cannot indicate that some important covariates have been omitted (not recorded). We can only argue that the list in Table 1 is quite exhaustive.

As another limitation, we highlight a problem in the definition of the outcomes in analysis C, and partly also in B. Suppose both units in a pair have outcome *L*, but one unit is the winner because his or her E spell of length *L* was realised earlier. Suppose further that the qualifying E spell of the losing unit ended at the end of the recorded future. It is therefore likely that the losing unit would have had a superior outcome (longer E spell of maximum length), had we had longer recorded future. If the losing unit had shorter maximum E spell than the winner, but this spell were at the end of the recorded future, there is still a likelihood, albeit smaller, that the losing unit would have had a superior outcome if the records of the future were longer.

We refer to such cases as ambiguities. Formally, a matched pair is said to have an ambiguity of order $M = 0, 1, \ldots$, if the difference of their outcomes is *M*, and the loser's qualifying E spell is at the end of the recorded future. In the 50 replications, the numbers of such ambiguities are in the range 119 – 157, with mean 137. The numbers of ambiguities of order 0 range from 44 to 74, of order 1 from 17 to 42, and they decline rapidly with the order. For example, they range from 3 to 14 for order 5. Also the likelihood that the loser would have a superior outcome if we had extended records (futures) diminishes with the order. The numbers of ambiguities are distributed evenly between treated and untreated units.

We can incorporate the ambiguities in a sensitivity analysis by reclassifying every treated loser in a pair with ambiguity as the winner, and the untreated winner as the loser. Such an analysis can be refined by allowing a realistic percentage of losers to remain losers, but this is not necessary in our case. Even if every ambiguity in the treated group is reclassified, the balance is reduced by approximately the number of ambiguities in the unweighted analysis C. This would not alter the conclusion that the treated units have fewer winners. For the weighted analysis, the ambiguities should be counted with weights; the same conclusion is arrived at.

Other analyses are in accord with the results related to the longest E spell. For example, the mean lengths of the reference spells are 5.17 months in the treated group and 1.88 in the untreated. In the matched-pairs analysis, this difference is reduced only slightly; the respective means are 4.89 and 1.71 months. The counts of winners and losers are in accord with this result; the treated group has 889 winners and 3112 losers, with 630 ties. Some caution is called for in interpreting this analysis, because administrative procedures may be delayed and, in case of short reference spells in particular, the award of an unemployment benefit may be linked with an incorrect U spell. However, we cannot condition the analysis on the length of the reference spell, because it is an outcome of the treatment.

6. Conclusion

We applied the potential outcomes framework to estimate the effect of awarding unemployment benefits on the resolution of the unemployment spell. Our conclusions are unequivocally negative about the benefit. Even after detailed matching on an exhaustive list of background variables, the outcomes are far superior on average for those not awarded benefits, both in terms of the length of the longest E spell in the recorded future and the speed of achieving it. Unemployment spells associated with benefits tend to be longer, even after matching on background.

The principal modelling effort is in the propensity analysis, which does not involve the outcome variable. Therefore, model selection does not introduce any bias. Concerns that one might have about the distribution of the outcome do not arise in our approach; the outcome is defined on a scale that best reflects our purpose, with no regard for the distribution of the values. We avoided defining a scale altogether and based the analysis on the within-pair comparisons. One might argue that some efficiency is lost in the process. However, the sample size in the analysis is so large that variance reduction is of secondary importance to combatting all possible sources of bias, among which selection bias (non-ignorable assignment to the treatment groups) presents the greatest threat. This can be interpreted as insisting on a higher standard for comparing like with like.

The agreement of the results of raw comparisons with matched-pairs analysis is no licence to apply the simple method. The 'raw' comparisons have no credibility for any causal analysis, and the agreement of their results in our analyses is no indication that such an agreement may arise in a similar context.

We studied the effect of unemployment benefits. A more detailed and arguably more relevant issue is the effect of a change in the rules for awarding benefits. Such a study is feasible only if the planned or contemplated changes are implemented, either by an experiment or administratively; see Card and Levine (2000); Carling, Holmlund and Vejsiu (2001); and Lalive, van Ours and Zweimu¨ller (2006) for examples.

REFERENCES

ABADIE, A., IMBENS, G., (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, 235–267.

CARD, D., LEVINE, P. B., (2000). Extended benefits and the duration of UI spells: evidence from the New Jersey extended benefit program. *Journal of Public Economics* 78, 107–138.

CARLING, K., HOLMLUND, B., VEJSIU, A., (2001). Do benefit cuts boost job finding? Swedish evidence from the 1990s. *Economic Journal*, 111, 766–790.

CRUMP, R.K., HOTZ, V. J., IMBENS, G. W., MITNIK, O. A., (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, 187–199.

HIRANO, K., IMBENS, G., RIDDER, G., (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.

HOLLAND, P. W., (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–970.

HUNT, J., (1995). The effect of unemployment compensation on unemployment duration in Germany. *Journal of Labor Economics* 13, 88–120.

LALIVE, R., ZWEIMÜLLER, J., (2004). Benefit entitlement and unemployment duration: The role of policy endogeneity. *Journal of Public Economics* 88, 2587–2616.

LALIVE, R., VAN OURS, J., ZWEIMÜLLER, J., (2006). How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies* 73, 1009–1038.

LECHNER, M., (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society* Series A 165, 59–82.

LONGFORD, N. T., NICODEMO, C., NÚÑEZ, M., NÚÑEZ, E., (2011). Well-being and obesity of rheumatoid arthritis patients. *Health Services and Outcomes Research Methodology* 11, 27–43.

ROED, K., ZHANG, T., (2003). Does unemployment compensation affect unemployment duration? *Economic Journal* 113, 190–206.

ROSENBAUM, P. R., (2002). *Observational Studies*, 2nd ed. Springer-Verlag, New York.

ROSENBAUM, P. R., RUBIN, D. B., (1983). On the central role of the propensity score in matching. *Biometrika* 70, 41–55.

RUBIN, D. B., (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.

RUBIN, D. B., (2002). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons, New York.

RUBIN, D. B., (2006). Matched Sampling for Causal Effects. Wiley and Sons, New York.

RUBIN, D. B., THOMAS, N., (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52, 249–264.

UUSITALO, R., VERHO, J., (2010). The effect of unemployment benefits on re-employment rates: Evidence from the Finnish unemployment insurance reform. *Labour Economics* 17, 643–654.

APPENDIX

Definitions of the labour force states

The labour force status of a resident of Luxembourg is established at the end of each month by the following rules:

- employed (E) – an entry in IGSS, but no entry in ADEM in the month;
- unemployed (U) – an entry in ADEM, but no entry in IGSS in the month;
- in transition (T) – entries in both ADEM and IGSS in the month;
- economically inactive (I) – entry in neither ADEM nor IGSS in the month, but an entry in either database in an earlier month;
- absent (A) – entry in neither ADEM nor IGSS in the month, and no entry in either of them at any time in the past.

The status A is relevant only to persons who appear in a dataset in a later month.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 273–286

HOUSEHOLDS' SAVING MOBILITY IN POLAND

Barbara Liberda[1], **Marek Pęczkowski**[2]

ABSTRACT

In this paper we examine the mobility of Polish households with regard to saving rates during the years 2007-2010 and compare it with the households' saving mobility during the years 1997-2000. The analysis for 2007-2010 is based on the household budget panel data from three panels of 15,000 Polish households selected by authors for the years 2007-2008, 2008-2009 and 2009-2010 from the Household Budget Surveys. We use the Markov mobility matrix and estimate the long-term ergodic distribution of households according to the saving rates. Our results show that the long-term households' distribution reveals a tendency towards polarization of households with regard to saving rates. Comparing the results for 2007-2010 with the authors' previous research on the households' saving mobility for a decade earlier during 1997-2000, we prove that between the years 1997-2000 and 2007-2010 the long-term change in the distribution of households was asymmetrical toward the highest saving rate groups. This helps to explain why Polish households could maintain positive and rising savings during the highly uncertain period of the financial crisis in 2007-2010.

Key words: household, saving, distribution, mobility, Markov matrix, polarization. JEL classification: D12, D19.

1. Introduction

According to the standard theory of life cycle and permanent income hypothesis, savings are determined by the persistence and growth of income as well as by consumers' perception of income uncertainty (Campbell 1987; Carroll 1994, 2009; Deaton 1992). The expectation of an income rise will lower savings but the fear of a cut in future income might increase saving rates. Consumers with greater uncertainty of income would save more. In the theoretical and empirical literature concerning household income and saving there is evidence that savings react mainly to the expected changes of current income. The impact of

[1] Barbara Liberda, Faculty of Economic Sciences, University of Warsaw, Dluga 44/50, 00-241 Warsaw, Poland. E-mail: barbara.liberda@uw.edu.pl.
2 Marek Pęczkowski, Faculty of Economic Sciences, University of Warsaw, Dluga 44/50, 00-241 Warsaw, Poland. E-mail: mpeczkowski@wne.uw.edu.pl.

unexpected changes of current income on saving is highly unpredictable (Hall 1978; Flavin 1981; Baxter, Jermann, 1999; Poterba 1994; Mody *et. al.,* 2012).

In earlier studies on Polish households for the years 1997-2000 it was found that the uncertainty of permanent and transitory income in Poland is similar to the income uncertainty in most of European countries and the variance of permanent income is generally lower than the transitory income variance. As the difference between uncertainty of permanent and transitory income in Poland is small, the shocks to both permanent and transitory income affect saving positively. The distribution of households with regard to saving rates was very uneven in Poland during the first decade after transition into the market economy (Liberda *et al.*, 2003, 2004, 2005).

This article extends the research on the changing structure of household savings in Poland to the decade after the previous research was conducted. We analyse the mobility of Polish households with regard to saving rates during the time of increased income uncertainty caused by financial crisis of 2007-2010.

The analysis is based on a very rich and representative set of household budget panel data from Polish households surveyed during 2007-2010. Three panels of 15,000 households each (for the years 2007-2008, 2008-2009 and 2009-2010) were selected by authors from the sample of 37,000 households surveyed yearly by Polish Central Statistical Office. Every year half of the surveyed group of households is exchanged and each household is surveyed during two consecutive years. Data are collected with the use of a monthly rotating method, e.g. each month one twelfth of the whole sample was surveyed. The selected panels of 15,000 households for the years 2007-2008, 2008-2009 and 2009-2010 meet all the formal requirements regarding data consistency. The chosen panel of households is representative and allows for a generalization of results with regard to the whole population of Polish households within a margin of a random error.

The paper is organized in the following way: in Section 2 we classify households into five groups with regard to saving rates. Section 3 presents the household income growth and financial savings of households in 2007-2010. In Section 4 we investigate changes in the households' distribution with regard to saving rates in three panel data sets for the years 2007-2008, 2008-2009 and 2009-2010 by applying the Markov mobility matrix. Then we estimate the long-term distribution of households for 2007-2010 and compare it with a similar distribution one decade earlier during 1997-2000. In Section 5 we present our conclusions.

Our study meets two goals. First, it fulfills the lack of research on household savings in transition economies of Europe during recent financial crisis. Second, according to our knowledge, this study is the only one that combines the analysis of household savings during financial crisis (2007-2010) with the results of similar research for Poland one decade earlier (1997-2000).

2. Definitions and classification of saving rate categories

We define household income as disposable income registered in the month of the survey (Methodology of Household Budget Surveys, CSO, 2011). Savings are calculated as a difference between disposable income and household expenditures, which is:

s = income - expenditures,

the saving rate *sr* being a ratio of savings and disposable income:

sr = s/income.

We express the saving rate as a percentage of the household disposable income.

Table 1. Classification of the saving rate class

Below -20%	very low saving rate (negative)
From -20% to -5%	low saving rate (negative)
From -5% to 5%	saving rate close to zero
From 5% to 20%	high saving rate (positive)
Above 20%	very high saving rate (positive)

Source: Own calculations.

The saving rate was divided into 5 categories: very low saving rate (negative), low saving rate (negative), saving rate close to zero (negative or positive), high saving rate (positive), very high saving rate (positive).

The distribution of households between groups of different saving rates in the years 2007-2010 is presented in Table 2.

Table 2. Number of households in groups of different saving rates in the years 2007-2010

Saving rates	2007	2008	2009	2010
<-20%	6343	5937	5722	5388
-20% to -5%	4083	3779	3655	3644
-5% to 5%	4488	4362	4325	4191
5% to 20%	7901	7917	7910	7789
20%+	14305	15112	15419	16177
Total	37120	37107	37031	37189

Source: Own calculations based on Household Budget Surveys 2007-2010, Central Statistical Office (GUS), Poland, Warsaw.

During the period of 2007-2010 the number of households in a group with the negative saving rates below minus 20% of the household income was decreasing.

The opposite happened in the group of households with the highest positive saving rate above 20% of the household income. The latter group was increasing in number and in its relative share in the total sample of households. The "middle" groups of households with either positive or negative saving rates were more stable in size during 2007-2010. The most visible changes of the household structure concern thus the two polar ends of the households saving rates classes.

In the next sections we will examine how the savings rates of the households varied within the studied years. Three constructed panel data sets for the years 2007-2008, 2008-2009 and 2009-2010 allow for a comparison throughout the four observed years.

3. Household income and savings in 2007-2010

The level and growth of real disposable income of Polish households grouped by saving rates in 2007-2010 are shown in Table 3.

Table 3. Level and growth of real disposable income of Polish households grouped by saving rates in 2007-2010

Saving rate groups	Monthly income 2007 zlotys*	Income growth		
		2008/2007 %	2009/2008 %	2010/2009 %
<-20%	1975	8.0	3.9	1.3
-20% to -5%	2220	8.4	2.2	3.2
-5% to 5%	2283	8.7	0.2	3.2
5% to 20%	2449	6.1	4.3	2.3
20%+	3298	6.9	2.2	4.1
Total	2650	8.2	3.0	4.1

* The exchange rate of zloty to US dollar and to euro in 2007 was 1US$=2.8 zloty and 1 euro=3.8 zloty (Statistical Yearbook, Poland, 2008, p. 617).

Source: Own calculations based on panel data for 2007-2008, 2008-2009 and 2009-2010 from Household Budget Surveys 2007-2010, Central Statistical Office (GUS), Poland, Warsaw.

The level of household income was positively correlated with the household saving rates. Disposable income of households grew from year to year during 2007-2010 in all households grouped by saving rates. This growth of household disposable income was higher in 2008 than in the following years when financial shocks continued affecting the Polish economy. The average growth of real disposable income of Polish households observed on a micro scale was about 5% per year during 2007-2010.

The increase of disposable income was different in each group of households classified by saving rates with no clear tendency detected. Households with the highest saving rates encountered very low growth of income between years 2008-2009, similar to income growth of households which save at low rates. In 2009-2010 the income of high saving households recovered fast, while the low saving households experienced still declining growth rates of yearly income in the same years.

With varying volatility of real household disposable income and increasing uncertainty caused by the financial crisis, Polish households maintained high and rising saving rates from their current disposable income during 2007-2010 (Table 4).

Table 4. Household saving rates in Poland in 2007-2010 (% of disposable income)

	2007	2008	2009	2010
Median	12.1	13.5	14.2	15.4

Source: Own calculations based on Household Budget Surveys 2007-2010, Central Statistical Office (GUS), Poland, Warsaw.

The median of the saving rate for all households increased in 2007 from a relatively high level of 10.6% in 2006 and was rising during the subsequent years to reach 15.4% in 2010. This increase of households' saving rates took place at the time of the fall of the value of household assets in different financial instruments like shares, investment funds, pension funds, etc.

During the turbulent years of the financial crisis Polish households took, on the whole, rational financial decisions. They did not stop investing in real estates (houses), partly because many of those investments were undertaken earlier and resulted in mortgage payments in the following years. Generally, households maintained repaying mortgage loans but raised less loans and credits. They also decreased the stock of deposits and securities, albeit on a moderate scale (Table 5).

Table 5. Average financial investment and credits of Polish households in 2007-2010 (in zlotys per household per month)

Year	2007	2008	2009	2010
Net purchase of real estates	77	94	99	109
Net increase of deposits and securities	-32	-61	-49	-45
Loans and credits repaid	134	155	176	190
Loans and credits raised	144	152	123	121

Source: Own calculations based on Household Budget Surveys 2007-2010, Central Statistical Office (GUS), Poland, Warsaw.

The strategies of financial investment of Polish households were rather standard and steered households to increase saving from current disposable income in order to protect decreasing value of their total financial stocks and real wealth.

4. Mobility of households with regard to saving rates during the years 2007-2010 and 1997-2000

In this section we investigate the changes in the structure of households with regard to saving rates from year to year during 2007-2010 and compare the results with the similar analysis for a panel of households one decade earlier during 1997-2000.

Having classified the five groups of households with different levels of saving rates (Table 1), we examine the households' mobility between these groups in any of two consecutive years: 2007-2008, 2008-2009 and 2009-2010. Tables 6, 7 and 8 illustrate the distribution of mobility from a particular class/group in the initial year to the same or other groups in the following year (transition matrix). This method of research is used in Markow processes (Podgórska et al., 2000) and concerns many fields, like income mobility, regional convergence and labour markets (Jenkins, 2006; Quah, 1996; Wójcik, 2009; Tyrowicz, Wójcik, 2010).

To conclude about a long-term distribution of households subject to the saving rates, we presume a constant household mobility between the high and low levels of saving rates. Based on the calculated transition matrix for yearly transitions in 2007-2008, 2008-2009 and 2009-2010, such a procedure would allow for a construction of an ergodic distribution. The distribution would illustrate a potential long-term household structure, which reflects the probability of the household to fall into the group of households of higher, lower or the same level of saving rate (under the assumption that mobility in the long run will be similar to the mobility observed in the analysed period).

In order to verify such an assumption, the long-term household distribution subject to the saving rates was calculated, initiating a Markov process from mobility matrix in particular years (2007-2008, 2008-2009 and 2009-2010).

Table 6. Households' mobility matrix subject to saving rates, 2007-2008

		Saving rates 2008					Total
		<-20%	-20% to -5%	-5% to 5%	5% to 20%	20%+	
Saving rates 2007	<-20%	31.2	14.5	14.6	16.4	23.3	100
	-20% to -5%	19.5	15.4	15.5	22.7	26.9	100
	-5% to 5%	15.0	12.8	17.5	25.5	29.3	100
	5% to 20%	12.6	9.0	13.5	27.0	37.8	100
	20%+	8.6	6.3	7.3	18.7	59.1	100
Total		15.4	10.1	12.0	21.4	41.1	100

Source: Own calculations based on panel data for 2007-2008 from Household Budget Surveys 2007-2008, Central Statistical Office (GUS), Poland, Warsaw.

Among the households characterized by very low negative saving rates (below minus 20% of the household income) in 2007, about one third remained in the same group in the following year. Three groups of ca. 15% of the households each moved to the 'low saving rate' class (minus 20% to minus 5% of the household income), 'close to zero saving rates' class (minus 5% to plus 5% of

the household income) and to a 'high saving rates' class (5% to 20% of the household income).

The most spectacular move was observed by almost one fourth of the households that shifted up to the 'very high saving rate' class (above 20% of the household income) in 2008 from the 'very low negative saving rates' class (below minus 20% of the household income) in 2007. Moreover, in the 'very high saving rate' group three fifths of the households remained in the same class in the following year (Table 6).

Tables 6, 7 and 8 reflect large fluctuations between particular groups of households in each period of two consecutive years 2007-2008, 2008-2009 and 2009-2010. Additionally, the 'very high saving rates' class is the most stable one because 60% of households from that group remained in the same group throughout 2007-2010. The retention of households in the 'lowest saving rates' group is ca. 30% of the size of this group. The 'lowest saving rates' group consists not only of households with very low income but also of households that take large mortgage credits that leads to higher savings from current disposable income in the next months (years) when the credits are in repayment.

The number of households in each saving rates' class and shifts from a particular class in the initial year to other classes of households in the following year are presented in the cumulated bar charts (Figures 1, 2 and 3).

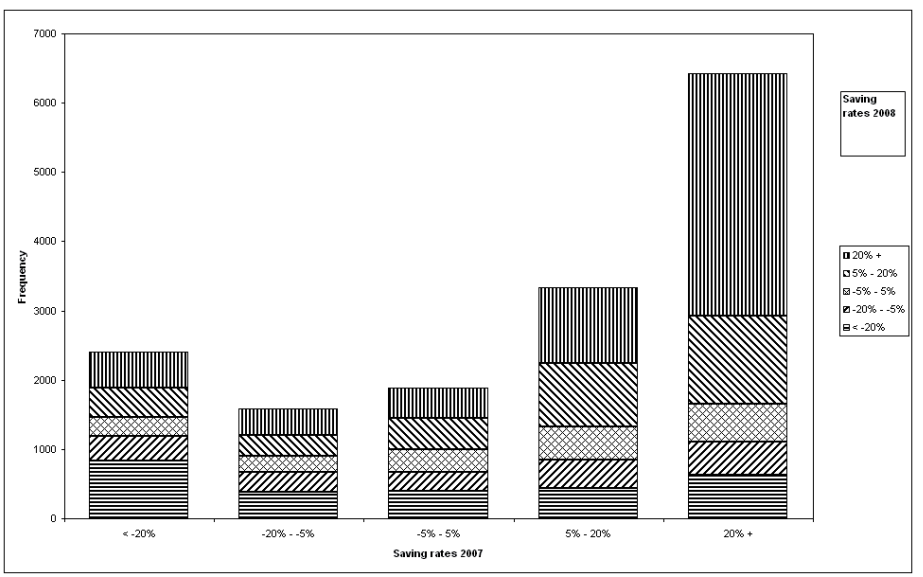


Figure 1. Households' mobility subject to saving rates, 2007-2008.

Source: Own calculations based on panel data for 2007-2008 from Household Budget Surveys 2007-2008, Central Statistical Office (GUS), Warsaw.

Each bar in Figures 1, 2 and 3 represents one row in Tables 6, 7 and 8, respectively. It shows how the households that belonged in the initial year of observation (e.g. 2007) to a particular class of savings (for example minus 20% of

the household income) are distributed between classes of households with different saving rates in the second year of observation (e.g. 2008).

Table 7. Households' mobility matrix subject to saving rates, 2008-2009

		Saving rates 2009					Total
		<-20%	-20% to -5%	-5% to 5%	5% to 20%	20%+	
Saving rates 2008	<-20%	31.7	12.4	13.0	17.1	25.7	100
	-20% to -5%	19.9	15.2	15.3	22.2	27.3	100
	-5% to 5%	14.9	13.9	16.8	25.4	29.0	100
	5% to 20%	11.9	9.8	13.7	27.3	37.2	100
	20%+	8.3	6.3	7.2	18.7	59.5	100
Total		15.0	9.9	11.5	21.4	42.2	100

Source: Own calculations based on panel data for 2008-2009 from Household Budget Surveys 2008-2009, Central Statistical Office (GUS), Poland, Warsaw.

Each bar in the chart (and a row in the table) sums up to 100% of households of this particular saving rate class. The height of the bar reflects the number of households that belong to one class of saving rates in the initial year of observation and the distribution of this class of households between classes of households with different saving rates in the following year.

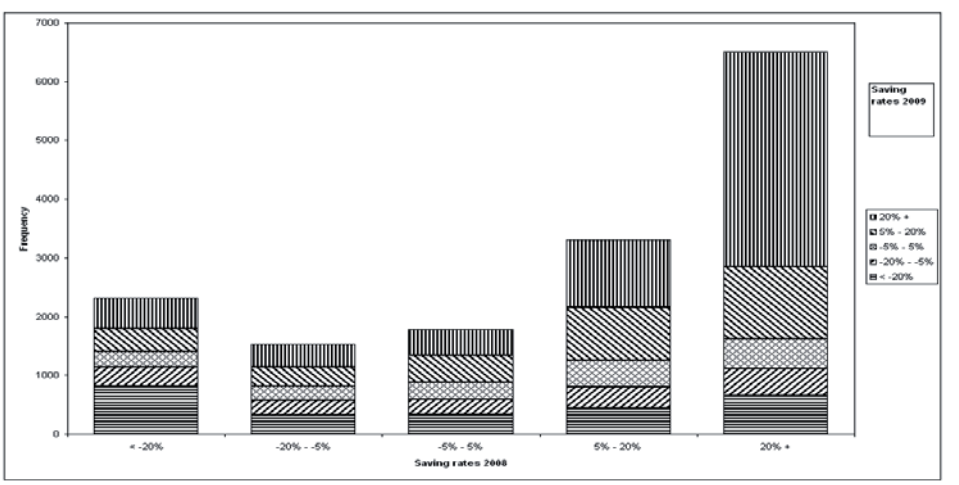


Figure 2. Households' mobility subject to saving rates, 2008-2009.

Source: Own calculations based on panel data for 2008-2009 from Household Budget Surveys 2008-2009, Central Statistical Office (GUS), Poland, Warsaw.

The differences in households' mobility between particular years are small. The groups of households that stayed right where they were a year before are those that save the most and those that save the least.

Table 8. Households' mobility matrix subject to saving rates, 2009-2010

		Saving rates 2010					Total
		<-20%	-20% to -5%	-5% to 5%	5% to 20%	20%+	
Saving rates 2009	<-20%	30.6	13.6	11.6	18.5	25.7	100
	-20% to -5%	21.7	15.3	15.2	22.4	25.5	100
	-5% to 5%	16.8	11.1	16.4	24.0	31.7	100
	5% to 20%	10.6	10.5	12.8	25.9	40.2	100
	20%+	7.0	6.0	7.1	17.3	62.6	100
Total		14.0	9.7	10.9	20.6	44.8	100

Source: Own calculations based on panel data for 2009-2010 from Household Budget Surveys 2009-2010, Central Statistical Office (GUS), Poland, Warsaw.

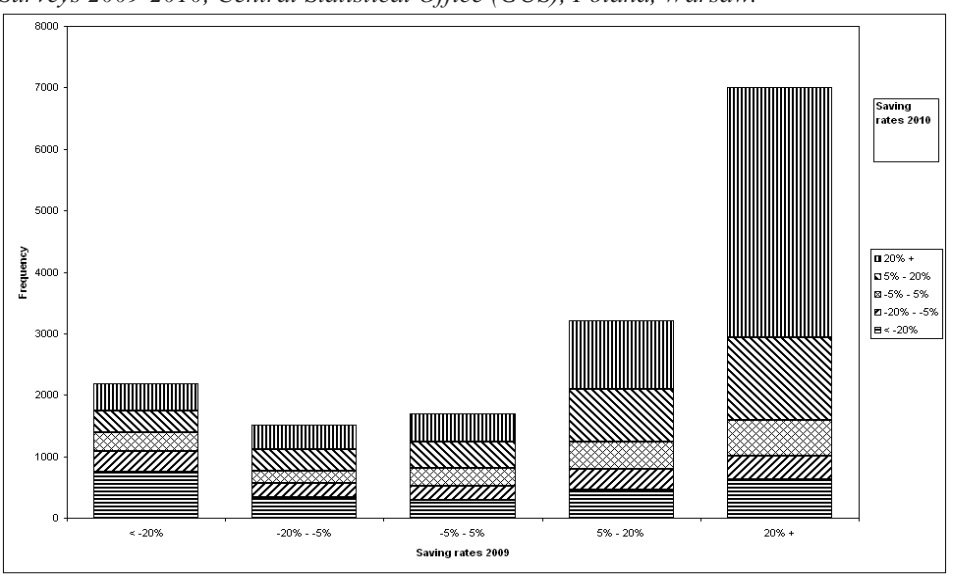


Figure 3. Households' mobility subject to saving rates, 2009-2010.

Source: Own calculations based on panel data for 2009-2010 from Household Budget Surveys 2009-2010, Central Statistical Office (GUS), Poland, Warsaw.

Assuming a constant household mobility we could draw some conclusions about the long-term distribution of households with regard to the saving rates. Such procedure would allow for a construction of an ergodic distribution. This distribution would illustrate a long-term household structure, which reflects the probability of falling into a higher, lower or the same level of saving rate.

In order to verify such assumption, the long-term household distribution subject to the saving rate was calculated, initiating a Markov process from mobility matrix in particular years (2007-2008, 2008-2009 and 2009-2010). Table 9 shows the long-term household distribution subject to saving rates for particular initial periods.

Table 9. Long-term distribution of households subject to saving rates 2007-2010

	Saving rates levels				
	<-20%	-20% to -5%	-5% to +5%	5% to 20%	20%+
	Long-term households' distribution				
Years*					
2007-2008	0.145	0.097	0.117	0.213	0.428
2008-2009	0.143	0.096	0.113	0.214	0.434
2009-2010	0.133	0.093	0.106	0.204	0.464

*Starting dates of the Markov processes

Source: Own calculations based on panel data for 2007-2008, 2008-2009 and 2009-2010 from Household Budget Surveys 2007-2010, Central Statistical Office (GUS), Poland, Warsaw.

Table 9 indicates that in three of the long-term household distributions with regard to saving rates the probability of falling into a group of very high saving rates (above 20% of the household disposable income) was the highest (ca. 0.44). The probability of belonging to the lowest saving rate class was close to 0.14. Groups falling into 'minus 20% to minus 5%' saving rate group and 'minus 5% to plus 5%' saving rate group were the fewest.

The results of the households' mobility to different groups of saving during 2007-2010 are compared to the authors' previous research on households' saving mobility for Poland for years 1997-2000 (Table 10).

Table 10. Long-term distribution of households subject to saving rates 1997-2000

	Saving rates levels				
	<-20%	-20% - -5%	-5% - +5%	5% - 20%	20%+
	Long-term households' distribution				
Years*					
1997-1998	0.211	0.127	0.137	0.196	0.328
1998-1999	0.198	0.125	0.125	0.221	0.332
1999-2000	0.198	0.132	0.130	0.236	0.303

*Starting dates of the Markov processes

Source: Liberda, Peczkowski, 2005, p. 40; own calculations based on the 4-year panel data of 3001 households from Household Budget Surveys 1997-2000 for Poland, Central Statistical Office (GUS), Warsaw.

In both long-term distributions of households with regard to saving rates for years 1997-2000 and 2007-2010 (Tables 9 and 10) the probabilities of belonging to the two top groups and to the bottom group of saving rates are the biggest. However, the shares of the households with the highest and with the lowest saving rates differ in the two studied periods of 1997-2000 and 2007-2010. The group of households with the highest saving rates is relatively bigger a decade later and the group with the lowest saving rates is relatively smaller. The share of households in the middle groups of the savings' distribution is diminishing: from 46-50% of the total number of households in 1997-2000 to 40-43% in 2007-2010.

The change in the distribution of households between 1997-2000 and 2007-2010 demonstrates the polarization of households with regard to saving rates. Polarization is understood as a relative shrinking of groups in the middle of the households' distribution by saving rates. The methodical in-depth analysis of different measures of polarization is done by Kot (2008). Between 1997-2000 and 2007-2010 the polarization of Polish households was asymmetrical toward the highest saving rate groups.

5. Conclusions

In this paper we examined the mobility of households in Poland with regard to saving rates in two periods: 1997-2000 and 2007-2010. The analysis was based on the household budget panel data selected by authors from the Household Budget Surveys for years 1997-2000 and 2007-2010. We applied the Markov mobility matrix and estimated the long-term ergodic distribution of households with regard to saving rates. It illustrates the probability of a household to fall into one of the saving rates classes ranging from negative savings of minus 20% or less to 20% or more of the household disposable income.

Our results concerning the years 2007-2010 show that during four consecutive years (2007-2010) almost one third of households that had negative savings of minus 20% of the household disposable income in the first year of observation remained in their class also the following year. But one fourth of this group of households shifted up in the following year to the group with positive savings of more than 20% of the household disposable income. In the class of households with the highest saving rates three fifths of households kept these high saving rates.

The long-term (ergodic) distribution of households reveals a very high probability (0.44) of a household to fall into a group with the highest saving rates and a relatively high probability (0.14) of belonging to a group with the lowest saving rates. The share of households in the middle of the savings' distribution is lower than is the sum of the shares of households with the highest and with the lowest saving rates. It indicates a tendency toward polarization of households with regard to saving rates.

Comparing the results for 2007-2010 with the authors' previous research on the households' saving mobility for years 1997-2000, we proved that the long-term change in the distribution of households was asymmetrical between 1997-2000 and 2007-2010 toward the highest saving rate groups. The group of households with the highest saving rates is bigger a decade later and a group with the lowest saving rates is smaller. The asymmetrical polarization of households towards the highest saving rate groups is significant in explaining why Polish households could maintain positive and rising savings during the highly uncertain period of the financial crisis in 2007-2010.

REFERENCES

BAXTER, M., JERMANN, U. J., (1999). Household Production and the Excess Sensitivity of Consumption to Current Income, *American Economic Review*, Vol. 89, No. 4, pp. 902–920.

CAMPBELL, J. Y., (1987). Does Saving Anticipate Declining Labour Income? An Alternative Test of the Permanent Income Hypothesis, *Econometrica,* 55, pp. 1249–1273.

CARROLL, C. D., (1994). How Does Future Income Affect Current Consumption?, *Quarterly Journal of Economics,* February, pp. 111–147.

CARROLL, C. D., (2009). Precautionary Saving and the Marginal Propensity to Consume Out of Permanent Income, *Journal of Monetary Economics,* Vol. 56, No. 6, pp. 780–790.

DEATON, A., (1992). Understanding Consumption, Clarendon Press, Oxford.

FLAVIN, M. A., (1981). The Adjustment of Consumption to Changing Expectations about Future Income, *Journal of Political Economy,* Vol. 81, No. 5, pp. 974–1009.

HALL, R. E., (1978). Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence, *Journal of Political Economy*, Vol. 86, No. 6, pp. 971–987.

Household Budget Surveys in 1997-2000 and 2007-2010, Poland*,* Central Statistical Office (GUS), Warsaw.

JENKINS, S. P., VAN KERM, P., (2006). Trends in income inequality, pro-poor income growth and income mobility, *Oxford Economic Papers*, 58(3), pp. 531–548.

KOT, S. M., (2008). Polaryzacja ekonomiczna. Teoria i zastosowanie, PWN, Warszawa.

LIBERDA, B., GÓRECKI, B., PĘCZKOWSKI, M., (2003). Uncertainty of Households' Income in European Countries and Poland, *EMERGO*, Vol. 10, No. 4, pp. 2–15.

LIBERDA, B., GÓRECKI, B., PĘCZKOWSKI, M., (2004). Saving from Permanent and Transitory Income. The Case of Polish Households, *Ekonomia,* No. 14, pp. 7–22.

LIBERDA, B., PĘCZKOWSKI, M., (2005). Mobilność oszczędzania gospodarstw domowych w Polsce, *Ekonomia,* No. 17, pp. 28–42.

Metodologia badania budżetów gospodarstw domowych (Methodology of Household Budget Surveys), (2011). GUS (Central Statistical Office), Poland, Warsaw.

MODY, A., OHNSORGE, F., SANDRI, D., (2012). Precautionary Savings in the Great Recession, *IMF Economic Review,* Vol. 60, No. 1, pp. 114–138.

PODGÓRSKA, M., ŚLIWKA, P., TOPOLEWSKI, M., WRZOSEK, M., (2000). Łańcuchy Markowa w teorii i w zastosowaniach, Oficyna Wydawnicza SGH, Warszawa.

POTERBA, J. M., ed., (1994). International Comparisons of Household Saving, University of Chicago Press, Chicago.

QUAH, D. T., (1996). Empirics for economic growth and convergence, *European Economic Review*, Vol. 40, No. 6, pp. 1353–1375.

TYROWICZ, J., WÓJCIK, P., (2010). Active Labour Market Policies and Unemployment Convergence in Europe, *Review of Economic Analysis*, Vol. 2, No. 1, pp. 46–72.

WÓJCIK, P., (2009). Are Polish Regions Converging? Simple spatial approach, in: *Civilizational competences and regional development in Poland*, eds. B. Liberda, A. Grochowska, Warsaw University Press, Warsaw, pp. 87–99.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 287–318

COHERENCE AND COMPARABILITY AS CRITERIA OF QUALITY ASSESSMENT IN BUSINESS STATISTICS[1][2]

Andrzej Młodak[3]

ABSTRACT

The problems of coherence and comparability exceed the classical notion of analysis of survey errors, because they do not concern single surveys or variables but the question of how results of two or more surveys can be used together and how relevant data can effectively be compared to obtain a better picture of social and economic phenomena over various aspects, e.g. space or time. This paper discusses characteristics of the main concepts of coherence and comparability as well as a description of differences and similarities between these two notions. Types of coherence and various aspects of perception of these notions in business statistics are analysed. Main sources of lack of coherence and comparability, factors affecting them (e.g. methodology, time, region, etc.) and methods of their measurement in context of information obtained from businesses will be also presented.

Key words: coherence, comparability, mirror statistics, benchmarking, data precision, complex indices.

1. Introduction

Coherence and comparability of data are one of the most important aspects of each survey, especially, if it is regularly repeated over time. It is obvious that properly coherent and consistent methodology guarantees efficient long–time

[1] The paper was prepared for presentation at the Conference *Statistics–Knowledge–Development* organized to celebrate the International Year of Statistics in Łódź, Poland, 17th – 18th October 2013.

2 The paper is based on the output of the ESSnet project Methodology for Modern Business Statistics – MeMoBuSt (Specific Grant Agreements No. 61001.2010.006-2010.702 and 61001.2010.006-2012.273). The author expresses his gratitude to Mrs. Judit Vigh (Hungarian Central Statistical Office) as well as to two anonymous reviewers for valuable comments and suggestions.

3 Address: Statistical Office in Poznań, Small Area Statistics Centre, Branch in Kalisz, pl. J. Kilińskiego 13, 62–800 Kalisz, Poland. E-mail: a.mlodak@stat.gov.pl.

analysis combining many variables and taking into account various cross–sections, divisions or clusters of investigated objects (economic entities, spatial areas, branch groups of enterprises, etc.). In modern reality, where many globalization processes based on highly developed strict interregional and international economic connections are observed, a collection and use of multi–aspect coherent and comparable statistical data is one of key priorities. They enable efficient monitoring of occurring changes and reacting when some problems or threats are recognized.

The role of coherence and comparability of data in European statistics was appreciated by the Statistical Office for the European Union (Eurostat), which in its Code of Practice formulated these requirements as the Principle 14, expressed as follows (Eurostat (2005)):

European statistics should be consistent internally, over time and comparable between regions and countries; it should be possible to combine and make joint use of related data from different sources.

Indicators

- *Statistics are internally coherent and consistent (e.g. arithmetic and accounting identities observed).*
- *Statistics are coherent or reconcilable over a reasonable period of time.*
- *Statistics are compiled on the basis of common standards with respect to scope, definitions, units and classifications in the different surveys and sources.*
- *Statistics from the different surveys and sources are compared and reconciled.*
- *Cross–national comparability of the data is ensured through periodical exchanges between the European statistical system and other statistical systems; methodological studies are carried out in close cooperation between* the Member States and Eurostat.

This paper discusses and analyses the basic characteristics of the main concepts of coherence and comparability as well as a description of differences and similarities between these two notions. It is worth noting that "coherence" has much more wider sense than 'comparability'. Particular types of coherence and various aspects of perception of these notions in business statistics will be analysed. Main sources of lack of coherence and methods of its assessment in the context of information obtained from businesses will be analyzed. Similarly, problems concerning comparability will be characterized indicating possible reasons of its lack and factors affecting it, such as time, region, national accounts and economic benchmark as well as complex methods of measurement of comparability level.

The paper consists of two main sections devoted to coherence and comparability. The first of them (Section 2) discusses methodological fundaments of concepts of coherence in the context of accuracy and recognition of incoherence. Types of coherence taking into account the character of statistics and sources of incoherence as well as modelling of assessment of coherence using

various factors having an impact on it are here also presented. Section 3 concerns the comparability and shows its relation to the coherence, methodological and empirical sources of incomparability, basic directions of appraisal of the level of this property. At the end (Section 4), some indices which can measure coherence and comparability in a complex way, characteristics of accuracy and related problems are proposed. In Section 5 some conclusions are formulated.

2. Coherence

To obtain business statistics of satisfactory quality it is necessary to guarantee coherence between relevant data. *Coherence* is a general term referring to the consistency between a set of statistical variables describing finite population parameters in terms of various but – from the methodological point of view – mutually connected social or economic phenomena observed in business reality. More precisely, the level of coherence informs us whether and to what degree some statistics can be analysed jointly and how to indicate their 'optimum' levels. That is, if two variables are strictly methodologically connected and changes of one variable affect values of the other (and vice versa) then they can and should be analysed jointly. For example, such a joint analysis may concern production output and employment, production output and foreign trade turnover, sold production and wages and salaries, etc. These variables may also be combined to obtain one synthetic result.

2.1. Definition of coherence

According to Eurostat (2009), **the coherence of two or more statistical outputs refers to the degree to which the statistical processes by which they were generated relied on the same concepts – classifications, definitions, and target populations – and harmonised methods**. If statistical variables are coherent, it means that they have the potential to be validly combined and used jointly. As it was noted earlier, examples of such joint use reflect the situation where statistical outputs refer to the same level of aggregation forms (population, reference period, territorial level etc.) but concern different sets of data items (e.g. data on wages and salaries and production). Coherence can also occur in the opposite situation, i.e. where the investigated variables comprise the same data items (e.g. employment data) but are collected for different reference periods, regions or other domains.

Various data are usually collected using different processes (e.g. different reporting forms are used for the survey of employment and production output for medium and large economic entities – in addition, the former is conducted monthly and the latter quarterly, hence, in this situation one can look for coherence only for quarters). According to Eurostat (2009), the term coherence is usually used when assessing the extent to which outputs from different statistical processes have the potential to be reliably used in combination. More precisely,

coherence concerns the possibility of combining the two aforementioned variables for the same population and time period. Coherence depends on statistical processes leading to a given output and its level is assessed on the basis of final results in terms of quality of these processes.

Another aspect of coherence is **statistical accuracy**. In order to use several sets of variables or one large set of various variables it is very important to indicate which ingredients of the definition of these variables are constant and which vary and to what extent. For example, it is desirable to know what methodological rules were used to collect statistics about production output, the number of employees or turnover in various surveys and how they can be compared across various geographical areas or types of enterprises. The precision and accuracy of collected data can be a significant factor. For example, if one of several successive surveys contains missing values about turnover for many units, it can be dropped from the study (provided, of course, this action does not affect the resulting analysis).

It is worth noting that **there is a significant difference between coherence and accuracy**. Namely, if results of two different processes or of the same process at different aggregation levels are compared, differences between them are an effect of inconsistency of relevant estimates due to their various precision. This is an example of inaccuracy, which is different from incoherence, since it does not involve analyzing the possibility of combining these processes. In general (cf. Eurostat (2009)), coherence/comparability refers to descriptive (design) metadata (i.e., concepts and methods) about the processes, whereas accuracy is measured and assessed in terms of operational metadata (sampling rates, data capture error rates, etc.) associated with the actual operations that produced the data. Accuracy also concerns the problem of differences between various estimates of target variables and their errors (especially if error profiles are incomplete or unknown). The authors of the document by Eurostat (2009) note that where error profiles of the statistical processes are known and included within the description of accuracy there is no need for further reference to them under coherence (and also comparability). They give an example where it is supposed that sampling error bounds are published for two values of the same data item for adjacent time periods indicating the range within which a movement from one period to the next may be due to chance alone and does not reflect any actual change in the phenomenon being measured. If and only if the measured movement is larger than these bounds, one can discuss whether the movement is real or due to the lack of coherence (and comparability). On the other hand, the assessment of coherence should include elements not analysed in terms of accuracy, e.g. on non-response or non-sampling errors if they were not taken into account when assessing accuracy.

Körner and Puch (2011) discuss Eurostat's attempts to define coherence and note that two sources providing non-deviating results are therefore not necessarily coherent (as various effects in the underlying processes might mutually compensate each other). Therefore, it is easier to recognize incoherence than

coherence. Their operational definition starts from such deviations of results of two different processes and takes the expectations of users regarding consistency of results into account. It assumes that statistical outputs referring to identical concepts have to be numerically consistent. That is, two related surveys are coherent if the value of some basic variable observed in both surveys does not deviate (given that identical concepts are being used) from that expected by the user. For instance, it can be the number of enterprises in monthly employment surveys and quarterly reports on financial results).

The authors of the document by Eurostat (2009) view the relationship between coherence/comparability and accuracy by noting that the numeric consistency of estimates depends on two factors:

- logical consistency (called here coherence/comparability) of the processes that generated those estimates; and
- errors that actually occurred in those processes in generating the estimates,

and conclude that coherence (and comparability) is a prerequisite for numerical consistency. The degree of coherence/comparability determines the potential for numerical consistency (but does not guarantee it, since it also depends on errors).

2.2. Types of coherence

In terms of business statistics, two types of coherence are usually distinguished:

- ***internal coherence*** – coherence within a uniform set of annual and short–term business statistics or between data derived from different sources (surveys, administrative INTRASTAT, etc.);
- ***external coherence*** – consistency between business statistics and main macroeconomic indicators, e.g. with national accounts, statistics on prices and wages, external trade, etc.

In the first case, there are many possible causes of divergent trends between short-term indices and annual business statistics. Davies (2000) and Bergdahl *et al.* (2001) analyze several aspects in terms of which coherence should be perceived.

First, it is important how a variable is defined. A typical definition includes the main features of the variable to be collected such as statistical measure (total, mean, median, etc.), unit of measurement (e.g. different units are used to measure production output, number of hours worked, number of employees, wages and salaries, etc.), unit of observation (enterprise, KAU, LKAU, local units, etc.), domain (definition of subpopulation – e.g. using the NACE classification – or classes, e.g. by number of employees) and reference times (a time point or period which units and variable values relate to).

The problem of coherence with respect to these aspects is equivalent to the comparability of data over such factors. That is, one expects comparability over time, between countries, between non–geographical groups or functional areas

and between statistics from several surveys. These issues will be described broadly in Section 3. Now it is enough to note that in practice reference times are mainly time intervals, such as calendar year, quarter, or month, sometimes they are also points in time, e.g. 1st January of the reference year. The reference time should be the same for all units and variables. Usually reference times agree for all variables and units in a FPP[1]. For monthly statistics, for instance, this means that the delineation of units should refer to the current month.

The second type of coherence in business statistics (i.e. external coherence) can be perceived in terms of logical relationships between target statistics. For example, when both monthly and annual statistics on expenditures are used, the sum of twelve monthly values should sum up to the annual total.

The authors of the document by Eurostat (2009) also indicate several other types of coherence related strictly to business statistics:

- ***coherence between short-term (i.e. sub-annual) and annual statistics***: a good example are monthly and annual production data for the same industries in the same region
- ***coherence with the National Accounts*** – underlines the key importance of coherence for economic surveys using the national accounts; the National Accounts compilation process will detect the possible lack of coherence.
- ***coherence with other statistics***: for example, coherence between employment produced by a labour force survey of household members and the number of employees produced by an economic survey of enterprises.

Bergdahl *et al.* (2001) noted that difficulties encountered by the user often depend strongly on the 'distance' between statistics used jointly (since definitions of variables used even in the same survey can vary in terms of reference times – a period in one case, a point in time in another). Moreover, reference times may correspond to that of the sample frame or the variable and definitions of enterprises can vary in different countries across the EU.

2.3. Sources of incoherence

Körner and Puch (2011) note that deviating results have two main sources:

- ***differences due to concepts*** (like the target population, the reference period, or the definition of the items for analysis) and
- ***differences due to methods used*** (e.g. data collection methods and procedures, the data processing approach, or the sampling design).

They investigate the German Labour Force Survey and National Accounts by identifying and quantifying definitional differences and then recognizing methodological differences. Methodological differences can cover various

1 FPP – Finite Population Parameters, a general notion covering basic descriptive statistics for finite population which distinguish it from characteristics of the infinite population in statistical modelling, e.g. population total and average of a given variable, the ratio of the population averages of two variables, the population variance or the population median of a given variable, etc. (cf. Bergdahl *et al.* (2001)).

aspects, such as all elements of survey design and errors, but also the accounting rules and estimation methodology. If methodological differences are significant, then methods-related differences are usually not easily identified and some additional algorithms are necessary.

Sources of incoherence can also be combined. That is, errors of one type are usually to some extent inherent in the second. For example (Körner and Puch (2011)), the sampling error is a combination of effects due to the sampling design and deviations from it in practical implementation. Similarly, the survey mode, albeit part of the methodology, will directly influence the results which can be obtained.

Differences between short–term and annual statistics result from frames and various units used in the sources. Also, a lot depends on the choice of the domain for unit aggregation (e.g. in surveying the same statistics (variables), different data can be obtained depending on whether units are grouped by the type of activity or by location). It is worth noting that if short–term changes over time – i.e. the realization and trend of time series – are analysed, coherence can only refer to intra–annual results. Körner and Puch (2011) note that coherence of time series will therefore be covered in dedicated sections, focusing on aspects specific to the shorter (e.g. monthly and quarterly) time series.

Because coherence and comparability have usually similar reasons, their various aspects will be described in detail in the part of this paper devoted to comparability (i.e. Section 3). This subsection was only an overview of the general nature of these problems.

2.4. Measurement of coherence

The measurement of coherence (and more precisely, incoherence) is sophisticated. Of course, the simplest way involves computing differences between results of analyzed surveys or data sources. One should remember, however, that not all aspects of coherence are measurable and numerical assessment may be the result of a number of factors, such as properties of concepts and methods used. Cook (2007) notes that *coherence in economic statistics is increasingly dependent on integration of information obtained from individual large enterprises from different sources. Making balancing adjustments at industry or industry group level is of diminishing effectiveness as economic activity is concentrated in fewer enterprises, where coverage rates differ across surveys. In some industries, such as pension funds, the scale of inter-company transfers is so large that it makes it difficult to estimate industry change in an unbiased manner. Coherence is improved by survey design, and most countries have a long way to go, in doing as well as they could. Much improvement is still possible in the way that statistical frames are adopted in individual surveys*.

It is, therefore, very important to be aware of how these aspects affect inconsistencies. A good example of such quality assessment is the study

conducted by Casciano *et al.* (2012), who analyse coherence by studying the difference in the final estimate for turnover and value added. They compare the estimation of turnover obtained solely on the basis of the survey with that based on the survey combined with data from administrative sources. The new estimate is very close to the initial one with a percentage difference of +0.03%, although there is a high variability in results when a breakdown by economic activities and size classes is used. Other factors that have an impact on the assessment are weighting and non-response (the difference due to unit non-response is higher in construction activities and lower in service sector), which has produced a higher estimate of turnover for micro and small firms and a strong underestimate for medium enterprises especially in service sectors.

The measure of overall inaccuracy should concern all data sources, not only the sample. One should also expect significant (i.e. related to methodology or non-response) and non-significant (generated by, e.g. econometrically modelled 'white noise', i.e. random disturbances) inaccuracies. One can expect that relationships and connections between parameters will also be reflected in their estimates. For example, the sum of monthly estimates of the number of employees should be equal to the relevant annual estimate. Of course, in some situations (e.g. in the case of indices), consistency can only be approximate, but such approximations should be as precise as possible. Significant differences in estimates for the same parameter imply incoherence. One should, however, bear in mind that such problems can result not only from definitions but also from systematic errors. Therefore, sources of these problems should always be carefully analyzed and recognized.

P. Davies (2000) provides a comparison between short-term statistics, annual statistics, and national accounts in Sweden. In his example short-term statistics measures own production output, whereas the other two sets of statistics estimate the value added. He observed that the growth rate between different years was unusually large, but he also noticed that if growth rates for output and value added were assumed to be equal/similar, some short-term statistics for output could be regarded as a first estimate of the value added.

3. Comparability

In many professional papers the terms "coherence" and "comparability" are used interchangeably. Such usage can be justified by the fact that comparability is a special case of coherence and involves situations where statistical outputs refer to the same data items and the aim of combining them is to make comparisons over time or across domains. However, the idea of comparability has its special features, and therefore deserves a broader description and analysis. Therefore, its definition will be presented along with highlighting how it differs from the main concepts of coherence and discussing some methods that enable us to assess and improve the comparability of statistical outputs.

3.1. Definition of comparability and its relation to the problem of coherence

The authors of the document by Eurostat (2009) note that unlike coherence, comparability is used when assessing the extent to which outputs from (nominally) the same statistical process but for different time periods and/or for different regions have the potential to be reliably used in combination (recall that coherence concerns the possibility of combining different variables for the same region or time period). Therefore, coherence is a stronger notion. More precisely, if several different time periods and the same population and region are analysed then the validity of the combined use of, e.g. employment data for this structure will be perceived in the context of comparability, whereas combining this variable with, e.g. data on wages and salaries for the same reference points will be considered in terms of coherence.

Hence, the comparability refers to the possibility of making efficient comparisons between units and variables according to various aspects and criteria. One most frequently used basis of such comparisons is time. To make such comparisons possible, the stability of definitions in successive survey administrations should be ensured and the 'state of the art' in all time periods should be well described. To do so, one should remember that when a change is made, special measures are recommended to improve comparability, for example, producing statistics in both ways on one occasion or even re–estimating part of the old series in terms of new definitions (according to Bergdahl *et al.* (2001)).

3.2. Sources of incomparability

The document by Eurostat (2003) provides the following list of sources of the lack of comparability, which can be the basis for designing each quality analysis and quality report:

Concepts
- Statistical characteristics
- Statistical measure (indicator)
- Statistical unit
- Target population
- Frame population
- Reference period and frequency
- Study domains
- Geographical coverage (for comparability over time)
- Standards
- Structure effects
- Conceptual aspects specific for a domain under study

Measurement Quality Dimensions
- Sample design
- Data collection

- Data processing
- Estimation

Measurement aspects specific for a domain under study (these include characteristics pertinent to any specific domain, e.g. thresholds for Foreign Trade Statistics).

As one can see, the possible reasons are numerous, but the authors of the next version of this document (Eurostat (2009)) describe these problems in a much more consolidated form. They argue that possible reasons for the lack of comparability between outputs of statistical processes may be summarised under two broad headings used already to present causes of incoherence, i.e. – differences in concepts and differences in methods. This is the starting point for introducing more detailed classification. First, circumstances connected with concepts which affect comparability will be described.

Target population – units and coverage. Problems with comparability may result from the fact that target populations could differ for two statistical processes, or for the same process over time, in a variety of different ways. These differences can also concern the spatial aspect of definitions of a given notion or the lack of its harmonization. For example, many EU countries use a different definition of economically active population in Labour Force Survey (LFS). This can result from various legislations concerning the retirement age (for example, in Poland it is 60 years for women and 65 years for men, but in France the retirement age was established at 60 for both genders). In the USA persons waiting to start a new job are classified as employed, but in EU LFS – as unemployed. On the other hand, even in one country this definition can differ over time. For example, many countries are planning to extend the age for retirement – e.g. in Poland to 67 and in France to 62 for both genders. This fact will result in changes of the definition of the target population and the comparability of future statistics over time. Eurostat (2009) also provides another example – monthly statistics of industry might only include manufacturing enterprises, whereas another statistical output with the same name might include manufacturing and electricity, gas and water production. Two related but different surveys may use various units to be investigated – one can use an enterprise whereas another – local unit (LAU) or kind of activity unit (LKAU).

Geographic coverage – some spatial areas may be included in the survey of one country and excluded in advance from another. For example, wages and salaries in manufacturing can be investigated only in urban areas (where this kind of economic activity is mostly concentrated) or also in rural areas. Moreover, the internal structure of sampled territorial areas in one country may be different from those in another one. As a result, the coverage of functional areas will also be different.

Reference period – time points of measurement of a given feature within a given survey may differ. For example (cf. Eurostat (2009) in a survey of employees an enterprise might be asked for the number of full-time employees on

the third Monday in the month or on the first day of the month; an annual survey may refer to a fiscal year beginning in March, another to a calendar year.

Data item definitions, classifications – it is a very important cause of the lack of comparability. A variety of non-harmonised methodological definitions are in use. A good example are differences between countries, which seem to be unavoidable (e.g. in terms of accounting systems, definitions of units and variables, etc.). In the common EU methodology (e.g. expressed in URBAN AUDIT project) the concepts of "employees" and "employed persons" are not really distinguished. However, in Polish official statistics each of these terms has a different meaning. Namely, employees are defined as persons performing work and receiving earnings or income, i.e.:

- employees hired on the basis of employment contracts (labour contract, postings, appointment or election),
- employers and own account workers,
- outworkers,
- agents (including contributing family workers and persons employed by agents),
- members of agricultural co–operatives, (agricultural producers' co-operatives, other co-operatives engaged in agricultural production and agricultural farmers' co–operatives),
- clergy fulfilling their obligations.

Data about employees are presented without converting part-time employees into full-time employees and applying the principle that each person is listed once according to their main job. Instead, the number of employed persons is defined as the sum of the number of full-time employees and the number of part–time employees. Full–time employees are persons employed on a full–time basis, as defined by a given company or for a given position, as well as persons who, in accordance with regulations, work a shortened work-time period, e.g. due to hazardous conditions or a longer work-time period, e.g. property caretakers. Part-time employees are persons who, in accordance with labour contracts, regularly work on a part-time basis. In the metodology adopted in URBAN AUDIT, on the other hand, the category of 'employed persons' includes only persons performing work on the basis of a relevant contract – that is, it does not include employers, for example.

Eurostat (2009) also provides a definition of unemployed person in LFS, which includes any economically active person who does not work, is actively looking for a job and is available for employment during the survey or any economically active person who does not work, is actively looking for a job and is or will be available for employment in the period of up to two weeks after the survey's reference week. A well–known reason for incomparability are changes in classification schemes or in particular revisions in accordance with new versions of international standards. For instance, the current version of the NACE classification can have significant modifications in relation to its older own

outputs or variants used in particular countries (e.g. the Polish Classification of Activity is based on NACE but is more detailed at lowest levels and takes into account types of economic activity specific for this country). On the other hand (cf. Eurostat (2009)), even without a change in classification, the procedures for assigning classification codes may be different or change over time, for example with improved training of staff or the introduction of an automated or computer assisted schemes.

The next part of our classification will concern the impact of methods of data collection, processing and analysis on comparability.

Frame population – within the same survey based on a generally established target population, the actual coverage of a survey may depend on the frame used. That is, e.g. employment and wages and salaries in enterprises with fewer than 5 employees are surveyed only once a year and it is only a sample survey, whereas for larger entities such a survey is conducted monthly and in general (maybe with some small exceptions) it is exhaustive. Problems may also occur when frames are based on various administrative sources, e.g. one is constructed using the tax register and another one – using the business register. Apart from the difference resulting from various definitions of units to be registered in these databases, one can also consider possible changes in registration (which may occur in relation to one or both sources) and survey designs (one can be, e.g. cross sectional and the second – longitudinal, which produces significant difference in estimates). One should also account for possible changes of panels or rotation patterns or training or application of new methods (e.g. for creation, amalgamation, clustering, allocation, etc.) occurring over time or between various countries.

Source(s) of data and sample design – this problem concerns a situation when in one statistical structural survey data are obtained from an administrative data source and in another – from a direct survey, or when actual sample designs are different. A good example may be financial data for large and medium enterprises, which can be obtained either from the tax register or from the current reporting.

Data collection, capture and editing – the same survey can differ in terms of the incidence of non–response problems and possibilities of their reduction. These issues are described in details, e.g. by Yancheva and Iskrova (2011). At this point, let us just give one example: in one case the non-response rate may amount to 10% and in another (e.g. in another round of the same survey) – to 40%. This fact may result either from some changes in the frame, sample or questionnaire or from the application of more efficient methods or reduction of burdens.

Imputation and estimation – methods of imputation and estimation may differ depending on the place or time when a given survey is conducted. For example (cf. Eurostat (2009)), in one survey zeroes may be imputed for missing financial items, whereas in another survey non-zero values may be imputed based on a 'nearest neighbouring' record. Likewise, in dealing with missing records in an enterprise survey there are various options, such as assuming that the

corresponding enterprises are non-operational or assuming that they are similar to enterprises that have responded. However, if the number of missing data is relatively large, the hot–deck (such as nearest neighbour, for example) methods of imputation are recommended (cf. de Waal *et al.* (2011)).

To illustrate the problem of incomparability, an example based on empirical experience of Polish business statistics is given. During processing data on economic activity of entities, some important data are generalized on the level of group for some NACE Rev. 2 sections, others on the level of division. Data are generalized also on the ownership sectors for NACE Rev. 2 sections. In Poland the obligation of transmission of reports on economic activity pertains to all economic entities belonging to established NACE Rev 2. sections (they cover mainly manufacturing units) having more than 49 employees and sampled entities employing up to 49 persons. To generalize collected data, the number of employees is used. That is, for entities which have transmitted the report this number if taken from these reports (and is called the variable number of employees). On the other hand, the sampling frame contains data on number of employees collected from business register or earlier reports (called the variable number of employees). Thus, to obtain the estimates of various values for a population the following generalization coefficient is used

$$u_d \stackrel{\text{def}}{=} \frac{\sum_{j\in N_{dr}} \vartheta_j + \sum_{j\in N_d \setminus N_{dr}} \theta_j}{\sum_{j\in N_{dr}} \vartheta_j}$$

for a level of aggregation d, where N_d is the set of all units at the level d and N_{dr} is the set of units which have transmitted their reports belonging to the level d, ϑ_j is the variable number of employees of the j–th unit and θ_j – fixed numer of employees for j–th unit.

On the basis of this coefficient various quantities, such as revenues firm sale of products and services turnover in current process, gross wages and salaries, retail sale, etc., are estimated using the following general formula (with small variants in some particular cases):

$$\widehat{\varphi}_{(d)} \stackrel{\text{def}}{=} u_d \sum_{i\in N_{dr}} \varphi_i$$

for every level of aggregation d, where φ_i is the value of interest for i-th unit. Thus, two important problems concerning coherence and comparability occur:

a) the number of employees in register can be significantly different than the actual number of employees of a given entity,
b) the generalization coefficient and, in consequence, estimate depend on the level of aggregation. Thus, the sums of relevant estimates for, e.g. NUTS 4 units belonging to the NUTS 2 unit obtained on the basis of generalization for NACE groups can be significantly different from relevant estimates obtained when generalization is made at the level of sectors of ownership.

The first of these inconveniences can have an important impact on the quality of estimation due to an error occurring when unknown actual number of employees for those units which have not transmitted the report is replaced with the number of employees based on the business register. However, some average impact of such bias can be estimated as

$$v_d \stackrel{\text{def}}{=} \left| \frac{1}{\sum_{j \in N_{dr}} \vartheta_j} - \frac{1}{\sum_{j \in N_{dr}} \theta_j} \right| \frac{\sum_{j \in N_d \setminus N_{dr}} \theta_j}{u_d}$$

and expressed in percents. Summing (or averaging) these indices over domains d within a given larger domain one can estimate a level of incoherence/ incomparability in this case.

To reduce the second problem an imputation at the level of units is recommended. The aforementioned index can be useful to identify areas where the imputation is most necessary. Now, in Polish statistics some works on this were undertaken.

3.3. Types of comparability

It should be noted that the Handbook issued by Eurostat (2009) mentions the following types of coherence/comparability that are to be included in a reliable quality report:

- ***comparability over time*** – data should show whether collected information for a given region in several different time point was gathered under the same circumstances in terms of definitions, population, etc.
- ***comparability over region*** – for example, data for the same month from the structural business survey conducted in two Member States, indicate which regions are covered in both surveys and which not and why;
- ***comparability over other domains*** – domains over which comparisons are often made include economic activity group, occupational group, and sex. An example would be annual structural data for agriculture with annual structural data for manufacturing collected by a different survey.
- ***internal comparability*** – referring to data produced by a (single) statistical process (but possibly comprising several different segments) for a single time period and region.
- ***comparability between short-term (sub-annual) and annual statistics*** – for example monthly and annual production data for the same industries in the same region.
- ***comparability with the National Accounts*** – for economic surveys that feed into the national accounts, coherence is vital and, in so far as it is lacking, the National Accounts compilation process will detect it.
- ***comparability with other statistics*** – for example, coherence between employment produced by a labour force survey of members of households and numbers of employees produced by an economic survey of enterprises.

3.4. Assessment of comparability

An efficient assessment of comparability requires the knowledge of most useful methods including characteristic features of definitions and design. Therefore, the current type of incomparability according to the typology presented in Section 3.3 has to be recognized. This presentation of methods for assessing and reporting comparability will start from the general approach and next some problems will be addressed by referring to the aforementioned typology in the context of constructing quality reports.

The **General Approach** is intended to explain causes of the lack of coherence/comparability. Namely, in quality reports effects of the main sources of incomparability should be recognized and described. Their impact on estimates has to be determined. An attempt to reconcile these estimates on the basis of such knowledge is also required. The authors of the document by Eurostat (2009) emphasize that any general changes that have occurred, which may have an impact on comparability, should be reported, for example changes in legislation affecting data sources or definitions, reengineering or continual improvement of statistical processes, changes in operations resulting from reductions or increases in processing budget, deviations from relevant ESS legislation and other international standards, etc. Although in many cases comparability is equivalent to coherence, there are some situations when such equivalence does not hold. That is, even if coherent statistical data describes the same phenomenon, this information can be incomparable due to the occurrence of some errors.

A proper comparability analysis should contain a presentation of concepts and methods, explanation of what causes a given problem and each possible source of problems concerning comparability should be separately described. According to Eurostat (2009), the first step is to conduct a systematic assessment of possible reasons for the lack of comparability, based primarily on the examination of key metadata elements, and identification and analysis of differences. This action can enable us to make a picture concerning the magnitude of any lack of comparability. The second step consists in predicting the likely effect of such a difference on statistical outputs. The final step is to aggregate and summarise in some way the total possible effect, in other words, to form an impression of the degree of (lack of) comparability.

A different question is how to detect possible problems in methodology or response when only ‘raw’ data with many reports and possibilities to aggregate them in given categories (e.g. time, space, by unit incomes, by the number of employees, etc.) are at disposal. The **quality control of variables** should provide the first indication of such problems. One can use general rules proposed in the document by Eurostat (2008).

Multi–variate type of control – checking if calculated sums of different variables and their values coincide with the total provided. For example, one should verify whether the sum of the number of entities by the number of

employee groups corresponds to the total number of units. If not, the cause of this fact has to be looked for.

Hierarchy type of control – it consists in comparing sums of the same variable against the variable value provided at an upper spatial level. For example, it is checked whether in any case the sum of the number of units in NUTS 5 areas included in a given NUTS 4 area corresponds to the number provided at the relevant NUTS 5 level. To allow rounded numbers and estimates, sometimes a tolerance or deviation of 3% of the checked value from the control value is set. This type of control refers to the same variable but to different spatial levels.

In practice, many variables have the form of indices. In other words, they are obtained as relevant ratios of given data expressed in absolute values (e.g. average wage and salary per employee is the ratio of the total sum of paid wages and salaries and the total number of employees). The 'sum control' is inappropriate in this case and the extreme values have to be looked for instead. The authors of the document by Eurostat (2008) propose three ways of **detection of outliers** which are usually 'suspected' to be incredible:

Classical interval of variation – its construction is based on the average and the standard deviation of the indicator values over a specified population. The control range is determined by the interval $\bar{x} \pm z \cdot \sigma_x$, where $\bar{x}$ is the arithmetic mean and σ_x – the standard deviation of the index X. The control consists in checking whether the individual indicator value is out of the control range. The value of z is set by the user (more often it is set as $z = 2$). In extreme cases, when the indicator cannot have a negative value, z is set to the ratio of average and standard deviation.

Median interval of variation – a construction of an interval of variation based on the median. That is, the control range is determined by the interval $\mathrm{med}(X) \pm z \cdot \mathrm{mad}(X)$ where $\mathrm{med}(X)$ is the median of X and $\mathrm{mad}(X) = \mathrm{med}_{i=1,2,\dots,n}|x_i - \mathrm{med}(X)|$ (x_i is the i-th value of X, $i = 1,2, \dots, n$, n – number of observations) – its median absolute deviation. The parameter z is established as previously.

Growth rates – this method makes use of growth rates of the dataset over various periods. It is an extension of the method based on the classical interval of variation and takes into account gaps between the analysed years. The threshold limits are usually specified as the maximum allowed growth rates per year. If there are missing time points between the observed ones, this has to be taken into account. For example, if monthly data on employment are analyzed, months where the growth rate exceeds the arbitrarily accepted tolerance interval (e.g. +/- 20%) are indicated. For missing time point the limits should be adjusted respectively.

Some aspects of comparability quality will be recalled now by referring to its typology introduced in subsection 3.3.

Comparability over time is very important if statistical outputs are published at a number of consecutive time points. Hence, changes over time of economic or social phenomena can be observed and analysed and ensuring comparability over

time seems to be crucial for such an analysis to be efficient. The user (when using time series provided by the respondent or the statistical office or constructed by him-/herself on the basis of available 'raw' data) should obtain information about possible limitations and problems in data use concerning comparability with respect of time. This information also has to be included in the quality report. According to Eurostat (2009), in assessing comparability over time the first step is to determine (from the metadata) the extent of changes in the underlying statistical process that have occurred from one period to the next. There are three broad possibilities:

1) There have been no changes, in which case this should be reported;
2) There have been some changes but not enough to warrant the designation of a break in series;
3) There have been sufficient changes to warrant the designation of a break in series.

If changes result in negligible effects on statistical outputs, then it is sufficient to make a relevant note in the metadata. However, if the effect is significant, two cases should be considered. If an effect is too small to warrant a series break, then (cf. Eurostat (2009)) the NSI may wedge in the changes to the outputs over a period of time so that, between any two periods, the adjustment being made to move from old to new values is less than the sampling error and thus cannot by itself be detected and interpreted as a real change. In the second case, i.e. if changes are significant and sufficiently large to cause the break, the user should be precisely informed about this circumstance, their location and consequences. Moreover, the authors of the document by Eurostat (2009) recommend three possible ways to handle these inconveniences:

- The most comprehensive treatment is ***to carry forward both series for a period of time and/or to backcast the series***, i.e., to convert the old series to what it would have been with the new approach by duplicating the measurement in one time period using the original and the revised definitions/methods.
- A less expensive treatment is ***to provide the users with transition adjustment factors giving them the means of dealing with the break***, for example by doing their own backcasting.
- The least expensive treatment is t***o simply describe the changes that have occurred and provide only qualitative assessments of their probable impact upon the estimates***. Obviously, this is the least satisfactory from the user's perspective.

Comparability over time is especially important for short-term business statistics, which can have different priorities and are usually more sensitive to any changes in methodology and the current situation than, e.g. annual data. Given comparable data, users have more possibilities of adjusting relevant data to their individual preferences and priorities. For example, revenue or sold production of enterprises need to be analysed both in the current period and over a longer time.

Comparability in time enables the user to make a better assessment of the situation, the effect of current business strategies and future prospects. Therefore, users can also be interested in much more advanced tools such as trend separation, regular seasonal variations or random and non-random disturbances. Thus, time series modelling should also strongly rely on comparability over time.

Comparability over region is an aspect which could be assessed in two different ways: pairwise comparisons of metadata across regions or comparison of metadata for the region with a standard (such standard can be perceived in terms of norms valid in ESS or best practices of statistical institutes). According to Eurostat (2009), two broad categories of situation can be identified:

- where essentially the same statistical processes are used, e.g. a labour force survey designed in accordance with ESS standard, and differences across regions are expected to be quite small; and
- where a different sort of statistical process is used, for example a direct survey in one case and a register based survey in another. In such cases, differences are likely to be more profound.

A complex measure of differences could be constructed as a sum of partial absolute differences quantified by a scoring system. Of course, the simplest solution in this context is to recognize the key metadata elements for which a difference occurs and use the binary code: no difference or difference. If the aim is to categorize the levels of discrepancies, several categories, e.g. from 5 to 0 by 1, i.e. from the most essential difference (5) to no difference (0), have to be assumed. The intensity of such difference could be quantified using arbitrarily assumed classification based, e.g. on the range of values of a given variable or experts' opinions. Apart from this, one should also consider the possible effect of such differences. This differs from analyzing 'pure' differences because, for example, in some circumstances (e.g. for ratios) even significant 'pure' differences could generate negligible effects in terms of comparability. Therefore, assigning a weight to each metadata element where the difference occurs according to its potential effect on comparability and computing a weighted score across all metadata elements is desirable. Such assessment can be summarised not only within a given country but over all countries within ESS.

One can say that owing to the intensive integration of economies of countries across the world, precise interaction analyses are particularly desirable. They help to recognize the position of a particular company in a given country in the context of other countries, especially those affiliated in such prestigious international economic organizations as OECD, EU, EEA, EFTA, CEFTA, G–8, NAFTA, ASEAN, OPEC, etc. The variety of producers of statistical data, economic practices, legal regulations (e.g. in terms of conducting economic activity or taxes), methodologies used to collect and process statistical data result in many serious difficulties in obtaining internationally comparable data. In recent years many projects designed to improve international data comparability have been initiated. One should mention the so-called 2007 Operation, which has resulted in the harmonization of major economic classifications and nomenclatures used in

the EU with those used in other parts of the world. Some attempts are also undertaken to harmonize the programme of statistical monitoring of urban areas URBAN AUDIT (e.g. in terms of the scope of the functional impact of cities). M. Bergdahl *et al.* (2001) discuss other harmonization actions, e.g. in terms of NACE and PRODCOM applied in Germany and UK.

Comparability over other domains – areas or time periods are, of course, not the only domains, over which comparisons can be performed. Usually, official statistical and various statistical studies present breakdowns by, e.g. size groups of enterprises, economic activity group, occupational group, sex, education of employees, etc. These aggregates can be treated similarly to areas and thus methods for assessing comparability are usually analogous to those mentioned earlier. One should, however, consider the possible differences between various statistical tools.

Internal coherence/comparability – users of statistical data expect, which is obvious and logical, that each set of published outputs should be internally coherent, i.e. all the appropriate arithmetic and logical dependencies should be observed. However, this requirement is sometimes difficult to satisfy. For example, some efficient estimation methods have a drawback or the implemented process comprises more than one segment with data from different sources or for different units in each segment. The authors of the document by Eurostat (2009) suggest giving a brief explanation to users and mentioning these problems in a quality report, with the reasons for publishing non-coherent results explained.

Coherence/comparability between short-term (sub–annual) and annual statistics – another natural expectation on the part of users is that sub-annual and annual statistical outputs are consistent. On the one hand, it means that similar arithmetic dependencies should occur (sum of gross monthly revenue for economic entities should be equal to its annual revenue or the number of newly employed persons in a given month should not be greater than its total in the relevant quarter) and, on the other hand, statistical processes producing these data are often quite different. Thus, all causes of the lack of coherence need to be assessed and accounted for. Eurostat (2009) suggests that a comparison of sub-annual and annual estimates should be the starting point for assessing the magnitude of differences due to the lack of coherence:

- if both annual and sub-annual series measure the same phenomenon in absolute values or indices (other than growth rates), annual aggregates can be constructed from sub-annual estimates and compared to totals from the annual series;
- if one or other of the series produces only growth rates, then comparison can be made of year over year growth rates.

Possible discrepancies can often be explained in terms of sampling error or some other measures of accuracy. Sometimes, however, they have their source in other problems (e.g. methodological) and then their explanation requires an assessment of the possible causes by metadata comparison, as for all forms of coherence assessment.

Coherence/comparability with the National Accounts – the authors of the document by Eurostat (2009) note that the National Accounts compilation process is a method for detecting the lack of coherence in data received from its various source statistical processes, whether they be direct surveys, register based surveys or indices. Thus, the level of comparability can be efficiently assessed and studied (e.g. in terms of causes of possible incomparabilities) using relevant data obtained from this system and properties of the adjustments aimed at obtaining the reliable balance of accounts. For more details see e.g. DESTATIS (2008).

Another interesting tool enabling an efficient assessment of comparability are **mirror statistics**. There are some variables for a given unit (area) that have their counterparts in another region or even country. For example, if employment-related population flows (commuting) from unit A to unit B is observed, then the number of employed persons coming from A to B should be equal to the number of employees coming to B from A. Similarly, in classical migration studies, emigration from Poland to UK and immigration to UK from Poland should coincide. Using such data a flow matrix can be constructed and analysed (see, e.g. T. Józefowski and A. Młodak (2009)). Another example: exports of country A to country B should be equal to imports of country B from country A. Thus, an effort should be made to ensure that relevant two-way statistics coincide. Hence, differences in definitions have to be carefully studied and minimized. As one can see, mirror statistics can be used to verify coherence, geographical comparability and accuracy. Thus, the difference in 'mirror' data can indicate the lack of accuracy in either or both of the outputs and/or may reflect the lack of comparability between regions (or countries) for the same data items. For example (cf. Eurostat (2009)), if the Polish estimate of emigration to UK in a particular year exceeds that of the British estimate of immigration from Poland for the same year by 10%, then this fact may indicate the lack of accuracy due to overestimation in Poland or underestimation in UK. One should then look for causes, e.g. in methodologies, such as the definition of emigrants/immigrants. In general, there are regulations concerning Business Registers (BRs) and statistics, like Structural Business Statistics (SBS) and Short–Term Statistics (STS), regulations about statistical units for the observation and analysis of the production system in the Community, where unit delineation and the BR together form an important part of the basis of statistical output. To overcome the most important problems, one should guarantee co-ordination between statistical surveys, for example in questionnaires, instructions for respondents, data processing, etc. This should, ideally, be done within respective National Statistical Institutes, where it would be much easier than elsewhere. According to M. Bergdahl *et al.* (2001), using the same BR as a frame, constructing the frame at the same time, updating units in the same way at the same time (with regard to business structure, classifications, etc.), addressing questionnaires to the same unit, etc., are further actions influencing coherence and accuracy.

The next problem is **coherence/comparability with other statistics and benchmarking**. It means that there may be other statistical surveys or data sources such that their statistical outputs can be used in combination with the results of a given survey. For example, some data from reporting on financial results of enterprises can be combined with relevant data from the tax register. Any such combination can show discrepancies, which should be (all possible) contained in the quality report. This problem is also connected with the question of benchmarking. According to Iurcovich *et al.* (2006), benchmarking is understood as an improvement process in which a company, organisation or any other (multi-organisational) system carries out three processes:

1) compares its performance against best-in-class systems;
2) determines how these systems have achieved their superior performance; and
3) uses the collected information to improve its own performance. Basically, all processes can be the object of benchmarking.

In the statistical context, benchmarking can be treated as a continuous process in which systems continuously seek to challenge their practices. More precisely, benchmarking is aimed at improving processes based upon the insights on what makes processes effective and efficient. Iurcovich *et al.* (2006) observe that benchmarking stems from the private sector business and has become increasingly popular for political systems over the past years, as nations and regions face increased competition from other competing systems. Statistics also uses benchmarking tools. For example, Barcellan (2005) provides a wide overview of such methodology from the point of view of the national accounts and analyses specific aspects of the European aggregates benchmarking, such as composition and quality of the indicators as well as possible effects of breaks generated by foreseen methodological changes and the impact of the introduction of the new approach to price and volume measures (chain-linking) on the benchmarking-based compilation of the European aggregates. The general benchmarking used in analyzing data comparability takes into account benchmarking of political systems and infrastructures at national level and less so at regional level by adoption and adaptation of modern techniques (like benchmarking) for improvement of data comparability. Statistical benchmarking comprises also the collection, analysis and documentation of good practices and use them to elaborate new solution, as much harmonized across countries and regions as possible.

To enable any comparison, some definitions (units, reference time, variables, sample frame, etc.) should be equivalent. The user should obtain precise information about all differences, problems and their practical consequences. Such information is usually contained in special quality reports. The problems of accuracy should also be mentioned in this context.

4. Measurement and improving degree of coherence and comparability

It is obvious that an efficient and complete quality report should contain basic measures of survey coherence and comparability. These measures should be constructed on the basis of expert opinions on coherence and comparability in terms of various particular aspects described in previous sections of this paper. Such proceeding is motivated by the fact that the degree of coherence and comparability is, in general, not measurable and may depend on subjective assessment of importance of particular factors of them. Our proposal in this respect consists of three main indices, where the third of them is constructed using two previous ones and can be interpreted as total complex index of coherence and comparability.

Let m, p, q and t be natural numbers and $X_1, X_2, \ldots, X_m$ denote variables collected in successive t editions of the analyzed surveys. Assume that assessment is made by p independent experts. Our first index will describe **methodological differences** in definitions of variables used in the aforementioned editions. To define it, it should be assumed that the subjects of assessment are q aspects of definition of analysed variables. It seems to be obvious that various experts can have different hierarchy of importance of the evaluation aspects, i.e. what is crucial for one expert can be of no significance for another. The index can be then computed as a sum of non-negative values expressing evaluations made by these experts:

$$I_{M(r)} \stackrel{def}{=} \frac{1}{pq_0(t-1)} \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ij} \sum_{k=2}^{t} d_{ijkr} \tag{1}$$

where $d_{ijkr} \in [0,1]$ denotes the level of consistency of the definition of variable X_r in k-th edition of the survey in comparison of its $k-1$–th edition in context of j–th aspect according to the opinion expressed by i–th expert (where 0 denotes total inconsistency, 1 – full consistency while an increasing evaluation note denotes a decreasing level of consistency in terms of a given aspect) and $w_{ij} \in [0,1]$ is the weight associated by i–th expert with j–th aspect expressing its potential effect on comparability (0 denotes that relevant aspect is dropped, the larger is the weight, the more important it is for an expert), $i = 1,2,\ldots,p$, $j = 1,2,\ldots,q$, $k = 2,3,\ldots,t$ and $r = 1,2,\ldots m$; $0 < q_0 \leq q$ is the number of aspects which have non–zero weight according to at least one expert. It is also assumed that (according to classical construction of weights)

$$\sum_{j=1}^{q} w_{ij} = 1 \tag{2}$$

for every $i = 1,2,\ldots,p$.

The index I_M takes values from $[0,1]$ – larger value indicates better consistency. It will be next assumed that the variable X_r for which $I_{M(r)} \geq \theta$ is regarded as sufficiently consistent from the point of view of its definition. The parameter $\theta \in (0,1]$ is the arbitrarily assumed threshold of consistency (it can be established, e.g. at the level of 0.6). Let S_M be the set of variables which are sufficiently consistent according to (1), i.e. $S_M \stackrel{def}{=} \{X_r: r \in \{1,2,\dots,m\} \wedge I_{M(r)} \geq \theta\}$.

The construction (1) can be perceived as an extension of the Borda group evaluation method. That is, the expert opinions are here only one of several factors affecting the final evaluation. On the other hand, the weights w_{ij} can be obtained on the basis of 'hidden' (i.e. subjective and not commonly published) individual matrices of preferences of experts concerning the importance of the investigated aspects for the comparability using the fuzzy Borda approach (see, e.g. García Lapresta and Martínez Panero (2002)). An interesting alternative could be here the use of especially modified Litvak's method (cf. Litvak (1983)). In the classical version of this method such option is preferred by an expert, which is ranked first in any preference order that minimizes the sum of distances between vectors of preferences. In this situation one can propose to set the expert weight of an aspect using the distance (computed according to the aforementioned definition) which is minimal for all preference vectors where this alternative is ranked first. The weights will be set as these minimal distances after normalization to satisfy the condition (2).

If the index I_M yields satisfactory results (i.e. if a sufficient number – e.g. more than a half of n – variables are consistent), a second index, based on similar rules, should be constructed. It is **related to comparability over time and over domains of interest** (i.e. spatial units, branch clusters of economic entities, etc.) – restricted to those variables with similar definitions. It should take into account that for some variables comparability over time can be more important than comparability over domains (e.g. total revenues), for others this expectation can be opposite (e.g. for some newly introduced variable having form of an index, which uses results from new phenomena and needs of users). Formally, the index I_D is constructed in the following way:

$$I_{D(r)} \stackrel{def}{=} \frac{1}{ap(t-1)} \sum_{i=1}^{p} \left(w_{ir(T)} \sum_{k=2}^{t} d_{ikr(T)} + w_{ir(U)} \sum_{k=2}^{t} d_{ikr(U)} \right) \qquad (3)$$

where $w_{ir(T)}$ and $w_{ir(U)}$ denote weights associated by i–th expert with time and analyzed domains for the variable X_r, respectively ($w_{ir(T)}, w_{ir(U)} \in [0,1]$), $d_{ikr(T)}$ and $d_{ikr(U)}$ are assessments of comparability over time and domains ($d_{ikr(T)}, d_{ikr(U)} \in [0,1]$, where 0 denotes total lack of comparability, 1 – full comparability in a given context), $i = 1,2,\dots,p$, $k = 2,3,\dots,t$ for every $r \in \{1,2,\dots,m\}$ such that $X_r \in S_M$. The parameter a takes values 1 or 2 depending on whether only one or both of these aspects of comparability have non–zero weight in the opinion of at least one expert. Similarly as in the case of (1) also the index

(3) takes values from [0,1] where 1 indicates the ideal comparability and zero – its total lack.

Finally, for each edition of the survey a complex **index of complex coherence and comparability** based on (1) and (2) should be computed. It will be defined as

$$I_C \stackrel{\text{def}}{=} \frac{\underset{r=1,2,\ldots,m}{\text{med}} I_{M(r)} + b \underset{r\in\{1,2,\ldots,m\}:X_r\in S_M}{\min} I_{D(r)}}{2}.$$

where $b = 1$ if $|S_M| \geq \left[\frac{n}{2}\right] + 1$ and $b = 0$, otherwise[1].

As one can see, the comparability of selected variables has here especial importance. It is motivated by the fact that methodologically consistent survey should have relevant and high level of comparability. Of course, $I_C \in [0,1]$ and higher value inform about better quality of the survey in this context. These indices can significantly help to assess survey quality. It is worth noting that the expert weights have a subjective character. That is, every time when the set of experts is changed a new value of the index I_C (and, of course, previous indices) is obtained. Thus, it is recommended to have stable and competent team of experts, i.e. such that possible change in it will not affect significantly the structure of preferences. On the other hand, the change of the weights may be also a result of importance of actual or foreseen changes in quality of data collection of a given variable of its methodology over time. In this context such elasticity of weights is rather desirable.

A small example illustrating these problems can be as follows. Assume that a survey do journeys to work is conducted. The data following three variables are collected: X_1 – average time of journey to main workplace, X_2 – distance between place of residence of main workplace and X_3 – means of transport from the place of residence to the main workplace. Data on variables X_1 and X_3 are collected from the sample survey of 20% of the population and data on X_2 are determined on the basis of the tax register. The following methodological aspects are considered: 1) frame population, 2) sources of data and sample design, 3) data collection and processing and 4) imputation and estimation. The problem of data comparability on the level of NUTS 2 (voivodships) and NUTS 4 (powiats) Polish regions and over time will be also considered. In case of the index (1) the aspects 1), 3) and 4) will have greater importance for consistency than 2) because most data are collected in sample surveys. Hence, $w_{ij} > w_{i2}$ for any i. The value of d_{ijkr}'s will depend on changes in methodology occurring in comparison with previous edition of the survey. If the changes are very small then d_{ijkr}'s will be close to 1 in any case. Otherwise, e.g. if in the next edition of the survey the data on X_2 are collected also from sample survey then $d_{ij(k+1)2}$ will be smaller,

[1] The symbol $[c]$ denotes the integer part of the real number c, i.e. the greatest integer not greater than c. The symbol $|Z|$ denotes the number of elements of the set Z.

especially for $j = 1,2,4$. As regards the index (3), the comparability over time can be here less important than other regions, and hence one can expect that $w_{ir(T)} < w_{ir(U)}$ for any r. However, taking the character of data collection for particular variables into account, one can observe that the comparability for NUTS 4 regions for X_1 and X_3 will be significantly smaller than for X_2, i.e. $d_{ikr(U)} \ll d_{ik2(U)}$ for $r = 1,3$. For NUTS 2 regions, due to higher quality of estimation, the comparability of X_1 and X_3 can significantly increase.

One more problem, which was not indicated previously due to its practical complexity and importance, is how to treat variables which are expressed using various measurement units and vary from the point of view of the range or significance of information provided. In many cases – despite efforts made to harmonize methodologies – such discrepancies are unavoidable owing to specific needs of users of data produced as a result of particular surveys, diversification of traditions observed between regions or countries, etc. Therefore, some obstacles cannot be overcome methodologically. Hence, numerical methods for the normalization of variables should be applied.

It is worthwhile to note that the normalization should denote also uniformization of characters of particular variables. Each variable has its own status being arbitrarily established, taking into account the relationship between the values and the reality they describe. The following three types of variables can be then distinguished:

- *stimulant* – the higher the value of the feature, the better (e.g. average monthly wage and salary or GDP per capita),
- *destimulant* – smaller values are much more desirable than higher ones (e.g. unemployment rate),
- *nominant* – has an optimum level of value (called also an imbuement point). Thus, below this point the feature is treated as a stimulant and above it – as a destimulant. Or, conversely – increasing values (and simultaneously lower than the optimum) are "worse" whereas decreasing ones (but greater than the optimum) are regarded as "better".

Destimulants and nominants should be converted into stimulants (e.g. by taking respective "bad" values with opposite signs). Next, they are then normalized. The relevant formula can lead either to unification of basic statistical measures (such as arithmetic mean, median, standard deviation or median absolute deviation) or ranges of features. A rich collection of those methods can be found in articles by Zeliaś (2002) or by Młodak (2006 a). They are based on the following formula:

$$z_{ij} = \frac{z_{ij} - a_j}{b_j},$$

where x_{ij} is the value of j–th variable for i–th unit, z_{ij} is its normalized value, a_j and b_j are parameters for j–th variable ($i = 1,2,\dots,n$, $j = 1,2,\dots,m$, where n denotes the number of analysed units) such as mean, standard, median, median

absolute deviation, j–th coordinate of the Weber median or another singular points of the multidimensional space (cf. (A. Młodak (2006 b)), etc. In general, if the arithmetic mean of each of normalized variable amounts to 0 and its variance is equal to 1, then normalization is called standardisation; if the range is constant and equal (e.g. to 1), then such transformation is called *unitarisation*, and if $a_j = 0$ and b_j is assumed to be arithmetic mean, minimum, maximum, median, sum, sum of squares, square root of the sum of squares of the value of the j–th variable ($j = 1,2, \dots, m$), then it is called *quotient transformation*.

Another option in this context is the use of benchmarking. That is, distances of a given object (represented by a vector of data on $\mathbb{R}^m$) and the benchmark are computed, i.e. a best–in–class object (real or artificial, created i.e. by optimization of values of particular variables), which all investigated objects are compared with. To this end, a special measure of distance between structures can be applied (e.g. Gower's distance, Minkowski's metrics and its particular cases, Canberra metrics, Kaziniec's, Gatew's or Jeffrey's and Matusita's formulas or Pearson's correlation coefficient – cf. Kaziniec (1968), Gatew (1977), Młodak (2006), Bruzzone *et al.* (1995)). Hence, the benchmarking normalization can be performed, e.g. as

$$z_{ij} = \frac{\left|x_{ij} - \varphi_j\right|}{d_i}$$

where $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_m)$ is the benchmark, d_i denotes the distance of i–th object from the benchmark, $i = 1,2, \dots, n$, $j = 1,2, \dots, m$.

Data from various questionnaires can also be used to construct complex indices of development in a given domain. It is very comfortable for the user, who obtains one synthetic information instead of many various data that tend to be directly incomparable. Some concepts concerning this issue are contained in the handbook by OECD (2008), but they are based mainly on simple normalizations and do not exploit all interesting properties of the modelled data. Hence, one can recommend here also Polish output combining the aforementioned normalization and benchmarking and taking into account efficient choice of diagnostic variables (cf. A. Młodak (2006 a)). Also some suggestions concerning theoretical methods of establishing the share of each object in the overall development of their whole collection have been formulated (cf. A. Młodak (2002)).

If this analysis is restricted only to the *uniformization* of variables, one could produce desirable *characteristics of accuracy*. Although it may often be of no primary importance, sometimes (e.g. if results of one of two surveys to be compared exhibit bias errors) its characteristics would be desirable. For example, when the same statistics are obtained in two different surveys, measures of their accuracy may be crucial in order to assess the consistency of resulting estimators. The quality report should also include a description of differences between various data producers with the same country, e.g. organizations, agencies,

statistical offices, etc. which can affect the final quality and comparability of collected data as well as differences between quality measures used in other countries. Accuracy can be quantified by measurement errors for units sampled with probability one (that do not contribute to sampling errors) or differences between basic characteristics of respondents in two surveys. If a unit has different respondents in different surveys and their internal structures also differ, the measurement error can be serious, if it was undetected or not reduced. Some secondary indicators (e.g. wage per employee or profitability rate) can be burdened by missing data, if the number of non-response units is relatively large. Imputation precision can also be perceived as a measure of the survey's accuracy. In many cases, however, accuracy should be assessed using several indicators describing various aspects of accuracy (i.e. variability caused by non-response), precision of the sampling frame, random errors due to sampling, etc. Using all components of accuracy statistics (including estimation error), one can determine a symmetric uncertainty interval around a given point estimate, which informs us about the expected precision of this estimate. The smaller this interval is, the more efficient the resulting statistical inference and data comparisons made on the basis of both surveys. To avoid any output distortion, it is very important to take all error sources into account. On the other hand, some negative effects can be mutually reduced. For example, if indices defined as a ratio of two quantity statistics obtained from the same survey or different surveys with similar structures and intensity of errors are considered, such a ratio can reflect them sufficiently, since they are included (with similar values) both in the numerator or denominator of such an index). A similar problem can be observed if profit (the difference between revenues and costs) is investigated. Hence, quality measurement should also be done using the simplest data, expressed in non-negative, absolute values (e.g. revenue in EUR, number of employees, fixed assets in EUR, etc.). In the case of monetary values observed over time, one should rely on the inflation factor using the fixed-base index (fixed prices, if applicable). They can also be normalized.

M. Bergdahl *et al.* (2001) also consider the method of co–ordinating statistical output called also benchmarking (but it is dynamic, whereas the previously described approach can be perceived as static), where one set of estimates is forced to agree with another. Typically, short-term statistics could be benchmarked against annual statistics, if the former (after aggregation to the calendar year) are an indicator of the latter. This procedure can have two benefits for the user: it unifies the two time series (ensuring that the monthly series has the same annual sum as the annual series); and it improves the accuracy of short–term statistics. For this to be meaningful, the two sets of statistics should have the same target parameters for the calendar year. The necessary comparison should be conducted both at the macro and micro–levels.

5. Conclusions

Coherence and comparability of data are one of the key aspects deciding on the quality of statistical output. Their role is underlined in many official documents and recommendations. There are various dimensions of these problems as well as their sources and typologies. In business statistics coherence and comparability have specific formal backgrounds and depend on quality of integration of information obtained from enterprises using different data sources. The methodological solutions differ between countries and sometimes even regions. Moreover, another inconveniences making difficulties in estimation of sufficient quality can occur, e.g. due to the scale of inter–company transfers in some economic entities as well as a dispersion of local units of the same enterprise over various countries. In any case, the user of statistical data should obtain precise information about all such differences, problems and their practical consequences. Hence, some international standards of quality reports were elaborated.

Thus, the construction of some universal methodological solutions guaranteeing at least roughly international coherence and comparability is very difficult. The only way to the improvement of these aspects of data quality leads through harmonisation of sampling frames, survey designs and used tools. However, to do this efficiently a reliable assessment of the level of problems should be performed. The presented original concepts of aggregated indices of coherence and comparability can be a significant support in realisation of this task. Of course, they are based to a large extent on subjective appraisal of the size and importance of key aspects. However, the stable and competent teams of experts and professional use of presented tools (such as, e.g. benchmarking) can ensure reliable results of such recognition which could be a hint how to permanently improve the data quality in the context of the investigated aspects.

REFERENCES

BARCELLAN, R., (2005). *The use of benchmarking techniques in the compilation of the European quarterly national accounts situation and perspectives*, Working Papers and Studies, European Commission, Euro Indicators, Office for Official Publications of the European Communities, Luxembourg, available at http://www.uni-mannheim.de/edz/pdf/eurostat/05/KS-DT-05-026-EN.pdf.

BERGDAHL, M., BLACK, O., BOWATER, R., CHAMBERS, R., DAVIES, P., DRAPER, D., ELVERS, E., FULL, S., HOLMES, D., LUNDQVIST, P., LUNDSTRÖM, S, NORDBERG, L., PERRY, J., PONT, M., PRESTWOOD, M., RICHARDSON, I., SKINNER, CH., SMITH, P., UNDERWOOD, C., WILLIAMS, M., (2001). *Model Quality Report in Business Statistics*, General Editors: P. Davies, P. Smith, http://users.soe.ucsc.edu/~draper/bergdahl-etal-1999-v1.pdf.

BRUZZONE, L., ROLI, F., SERPICO, S. B., (1995). *An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection*, IEEE Transactions on Geoscience and Remote Sensing, vol. 33, pp. 1318–1321.

CASCIANO, M. C., DE GIORGI, V., OROPALLO, F., SIESTO, G., (2012). *Estimation of Structural Business Statistics for Small Firms by Using Administrative Data*, Rivista Di Statistica Ufficiale No. 2–3, Instituto Nazionale Di Statistica, pp. 55–74.

COOK, L., (2007). *International experience in setting up an economic statistics compilation programme*, Regional Workshop for African countries on Compilation of Basic Economic Statistics jointly organized by United Nations Statistics Division (UNSD) and African Centre for Statistics at Economic Commission for Africa (ACS), United Nations, Department of Economic and Social Affairs, Statistics Division, 16th – 19th October 2007, Addis-Ababa, Ethiopia, Doc. No. ESA/STAT/AC.136.3, document available in the Internet at the website http://unstats.un.org/unsd/economic_stat/intl%20coop%20and%20workshops%20(bes)_files/AddisOct2007/UNSD%20documents/WS-BES-ECA-136-3-Intl-experience-Len%20Cook.pdf.

DAVIES, P., (2000). *Assessing the Quality of Business Statistics*, Office for National Statistics, Proceedings from the Second International Conference on Establishment Surveys, June 17th – 21st, 2000, Buffalo, New York, http://www.amstat.org/meetings/ices/2000/proceedings/S38.pdf.

DESTATIS, (2008). *National Accounts Quarterly Calculations of Gross Domestic Product in accordance with ESA 1995 – Methods and Data Sources.* New version following revision 2005, Fachserie 18 Series S. 23, Statistisches Bundesamt (Federal Statistical Office), Wiesbaden, Germany, https://www.destatis.de/EN/Publications/Specialized/Nationalaccounts/QuarterlyCalculationsGrossDomesticProductAccordance.pdf?__blob=publicationFile.

EUROSTAT, (2009). *ESS Handbook for Quality Reports*, Series: Methodologies and Working papers, Office for Official Publications of the European Communities, Luxembourg.

EUROSTAT, (2008). *Quality Control of Urban Audit Variables*, Unit D2, Office for Official Publications of the European Communities, Luxembourg, April 2008.

EUROSTAT, (2003). *Handbook "How To Make A Quality Report"*, Series: Methodological Documents, Working Group "Assessment of quality in statistics", Sixth meeting, Luxembourg, 2nd – 3rd October 2003, Document No. Doc. Eurostat/A4/Quality/03/Handbook, available at the website http://190.25.231.249/aplicativos/sen/aym_document/aym_biblioteca/Documento%20de%20soporte/Methodological%20documents%20handbook%20%20how%20to%20make%20a%20quality%20report%20-%20NACIONES%20UNIDAS.pdf .

EUROSTAT, (2005). *European Statistics Code of Practice for the National and Community Statistical Authorities*, Statistical Office of European Union, Eurostat, Luxembourg, available at http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2636.

GARCÍA LAPRESTA, J., MARTÍNEZ PANERO, M., (2002). *Borda Count Versus Approval Voting: A Fuzzy Approach*, Public Choice 112(1–2), pp. 167–184.

GATEW, K., (1977). *Статистическо характеризание на структурнии измениения,* Трудовэ на Висшия, Икономический Институт - К. Маркс, София, vol. 3, pp. 10–42 (in Russian).

IURCOVICH, L., KOMNINOS, N., REID, A., HEYDEBRECK, P., PIERRAKIS, Y., (2006). *Mutual Learning Platform. Regional Benchmarking Report. Blueprint for Regional Innovation Benchmarking*, European Commission, Committee of the Regions, IRE Innovation Network, available at http://www.rttn.ru/_files/fileslibrary/90.pdf.

JÓZEFOWSKI, T., MŁODAK, A., (2009). *Observation of flows of population in Polish statistics – problems and challenges*, [in:] E. Elsner, H. Michel (eds.) "Assistance for the Younger Generation. Statistics and Planning in Big Agglomerations", Institut für Angewandte Demographie IFAD, Berlin, pp. 61–76.

KAZINIEC, L. S., (1968). *О методах сводной оценки стуктурных сдвигов*, Вестник Статнстики, No. 11 (in Russian).

KÖRNER, T., PUCH, K., (2011). *Statistics and Science. Coherence of German Labour Market Statistics*, Volume 19, Statistisches Bundesamt (Federal Statistical Office), Wiesbaden. Available at https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band19_CohrenceLabourMarket1030819119004.pdf?__blob=publicationFile.

LITVAK, B., (1983). *Distances and consensus rankings*, 1 Cybernetics and systems analysis, 19 (1), 71{81. (Translated from Kibernetika, No. 1, pp. 57–63, January-February, 1983).

MALINA, A., ZELIAŚ, A., (1998). *On Building Taxonometric Measures on Living Conditions*, Statistics in Transition, vol. 3, pp. 523–544.

MŁODAK, A., (2006a). *Taxonomic analysis in regional statistics*, ed. by DIFIN – Advisory and Information Centre, Warszawa, Poland (in Polish).

MŁODAK, A., (2006b). *Multilateral normalisations of diagnostic features*, Statistics in Transition, vol. 7, pp. 1125–1139.

MŁODAK, A., (2002). *An Approach to the Problem of Spatial Differentiation of Multi-feature Objects Using Methods of Game Theory*, Statistics in Transition, Vol. 5, pp. 857–872.

OECD, (2008). *Handbook on Constructing Composite Leading Indicators: Methodology and User Guide*, Global Inventory of Statistical Standards, Organization for Economic Cooperation and Development, link: http://unstats.un.org/unsd/iiss/Handbook-on-Constructing-Composite-Leading-Indicators-Methodology-and-User-Guide.ashx.

STATCAN, (2006). *Metadata to Support the Survey Life Cycle*, Invited Paper, Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Topic (iii): Metadata and the Statistical Cycle, Submitted by Statistics Canada for the Conference of European Statisticians, United Nations Statistical Commission and Economic Commission for Europe, European Commission Statistical Office of the European Communities (Eurostat) Organisation For Economic Cooperation and Development (OECD), Statistics Directorate, Geneva, 3rd – 5th April 2006, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2006/zip.5.e.pdf.

DE WAAL, T., PANNEKOEK, J., SCHOLTUS, S., (2011). *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, Inc., Hoboken, New Jersey.

YANCHEVA, D., ISKROVA, K., (2011). *Reducing the administrative burden for the business in Bulgaria: Single Entry Point for Reporting Fiscal and Statistical Information*, [in:] Proceedings from BLUE-ETS Conference on Burden and Motivation in Official Business Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011, pp. 189–198.

ZELIAŚ, A., (2002). *Some Notes on the Selection of Normalization of Diagnostic Variables*, Statistics in Transition, vol. 5, pp. 787–802.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 319–328

STATISTICS AS A PROFESSION – STATISTICIAN AS AN OCCUPATION: observations and comments from a panel of experts

Wlodzimierz Okrasa[1], **Beata Witek**[2]

1. Introduction and background

Statisticians rarely devote as much attention to their profession as it does deserve, including such fundamental questions as what actually constitutes *statistics today* - as a discipline in relation to others, primarily to mathematics and observation-based sciences (theoretically or applied oriented), on one side, and in researching and teaching, on the other. Especially, given its inherent dynamics and externally caused transition to new stages of its permanent development, and what is their – statisticians' – own view of their occupational status, including who actually should unambiguously be considered *statistician*. And how s/he ought to be prepared through education and training system to play this important role in various domains – in academia and policy making, as well as in private life and business management. Therefore, any occasion to exchange views on such dual aspects (disciplinary and occupational status) among experts during scientific meetings seems to be worth of reporting. One of such meeting took recently place at the conference on Methods of Assessment of Quality of Teaching held in the University of Lodz last June (see note on it in this issue below) during which a discussion panel was organized to address some of the above issues. As the panel's organizers, we feel deeply indebted to all its participants for sharing their thoughts and opinions: Prof. Prof. Czeslaw Domanski, Jozef Dziechciarz, Miroslaw Krzysko, Marek Rocki, and Janusz Wywial. As a part of our appreciation of their generosity and of the quality of the panel's output, their voices are summarized here, extended a bit by introductory and concluding remarks, while taking into account the voices of the highly competent audience[3], composed of academic teachers and researchers.

[1] Central Statistical Office of Poland and the University of Cardinal Stefan Wyszynski in Warsaw.
[2] Central Statistical Office of Poland and the University of Cardinal Stefan Wyszynski in Warsaw.
[3] Discussants: Prof. Prof. K. Jajuga, S. M. Kot, A. Sokolowski, L. Tomaszewicz, Tadeusz Gerstenkorn. In addition, A. Kupis-Fijalkowska, PhD, was an invited discussant for presenting the Eurostat's European Master in Official Statistics (EMOS) initiative.

While statistics as a profession – meant as a domain of scientific activity, including education – is primarily an object of methodological reflection, statistician as an occupation is basically an example of a labour market category characterized also in sociological terms (status, prestige, ethos, etc.). Although the former was the main focus of the panel discussion, some occupation-related issues are worthwhile to be mentioned here too. In the vein of Neyman's saying - "[S]tatistics is the servant to all sciences" (cf. Chiang [1]), what implies its presence across all subject matter disciplines, through *inter alia* necessary involvement of representatives of those disciplines in applying statistical methods – statisticians constitute one of the most heterogeneous categories in occupational statistics. For instance, US Bureau of Labor Statistics counts about 200 specific occupations under this title, defined as follows: *"Statisticians.* Develop or apply mathematical or statistical theory and methods to collect, organize, interpret, and summarize numerical data to provide usable information. ... Includes mathematical and survey statisticians. Excludes "Survey Researchers" (US BLS 2010, p. 23)[2]. This is supplemented by definition: "*Survey Researchers***.** Plan, develop, or conduct surveys. May analyze and interpret the meaning of survey data, determine survey objectives, or suggest or test question wording. Includes social scientists who primarily design questionnaires or supervise survey teams. Excludes "Market Research Analysts and Marketing Specialists" (ibidem, 35). Such a broad interpretation of the occupation accords with perhaps the most widely accepted among statisticians answer to the question 'Who is the statistician?' given by Platek and Särndal about decade ago (2001[3]). Starting with considerations of what can realistically be expected from statisticians in terms of quality products generated by national statistical agencies, the authors arrived with the following definition: "[T]he statistician ... is anyone who contributes to the ultimate delivery of statistics and data to users.", and specify the main categories of this occupation: "theoretical statistician – survey methodologist – subject matter specialist – information technologist – and survey manager" (ibidem, p. 3). Keeping in mind such a broad interpretation of both 'statistics' and 'statistician' we began the panel discussion with a concern about the quality of the process of creating new generations of performers in the scene of this profession.

[1] "Statisticians in History", http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=11.

[2] US BLS 2010 SOC Definitions U.S. Bureau of Labor Statistics On behalf of the Standard Occupational Classification Policy Committee (SOCPC), http://www.bls.gov/soc/soc_2010_definitions.pdf.

[3] R Platek and C-E Särndal, 2001. Can a Statistician Deliver? Journal of Official Statistics, Vol. 17, No. 1. pp. 1–20.

2. Scoping panel's perspective

All the panelists and discussants agreed that there is a tremendous demand for statisticians and a big need to prepare new cohorts of specialists in the art of using data for sectors of education, government, and industry in a way readying them to meet the challenges from the technologically advanced society. It makes this occupation both an attractive path of carrier for new alumni – along with Tukey's view: "[T]he best thing about being a statistician is that you get to play in everyone else's backyard."[1] – and a highly respected as a job in view of the general public. For instance, according to a US survey of occupational status, statistician is ranked fifth out of about two hundred (together with mathematician and engineer – Kennett, 2011[2]).

Much of the panelists' attention revolved around the issue of *what?* to do and *how?,* in order to equip the new generations of statisticians in tools and abilities assuring the highest standards of professional quality, given the growing expectation concerning the statisticians' deliverables (in Platek and Särndal's meaning) on the one hand, and the existing drawbacks on the other. Especially, the lack of mathematical background among the majority of students as a consequence of the earlier reform of the high school curriculum, and subsequent lowering requirements from candidates for studying statistics, being often tough in the standard-liberal environment. This concerns the whole process of education, including textbooks and other means and conditions of teaching, which are summarized here as they emerged in the panelists' presentations:

(i) the means and conditions of teaching statistics;
(ii) the quality of teaching;
(iii) the problem of *curriculum*;
(iv) professional and occupational aspects of statistics.

3. The means and conditions of teaching statistics

The problem of quality of textbooks for teaching statistics

Because of the great importance and influence of the quality of textbooks on teaching and students' knowledge, this issue was one of the key ones discussed during the Panel. The topic was initiated by Prof. M. Krzysko, who referred to the first Polish textbook on statistics entitled "Outline of statistical methods as applied to anthropology" by Jan Czekanowski (1913), as an exemplary model. The speaker also drew attention to the limitations and weaknesses of modern

[1] American Statistical Association, http://www.amstat.org/careers/whatisstatistics.cfm.
[2] R S Kenneth, Statistics As a Profession.

textbooks on statistics: outdated content, excessive focus on descriptive statistics, presence of elementary errors, unfair reviews allowing authorizing the publication of low-quality books, collections of tasks that relate to outdated data and obsolete problems. Concern can be raised by textbooks on statistics in general secondary schools – it turns out that the textbooks approved by the Ministry of Science and Higher Education are not necessarily a good basis to start education in this area. Agreeing with the above, Prof. K. Jajuga stressed that textbooks should be tailored to different areas of expertise, including consultations with the appropriate persons about the subject matter related to statistics, for example, with economists in the econometric issues.

Prof. Cz. Domanski, as the President of the Polish Statistical Association (PTS), considered creating by PTS (in cooperation with other statisticians) a suitable consultative team for evaluating textbooks, or finding such a good author as, for example, Marek Fisz. He recalled that "a statistician seeks the truth" and one should not accept inappropriate behavior at meetings of a council (those related to, for example, lowering the number of hours or unacceptable combination of mathematics with statistics).

Supporting these suggestions, Prof. Okrasa stressed that it would be desirable in terms of assurance of quality of teaching statistics to set up a PTS's council or a team for new textbooks matters. Another valuable idea would be the one of creating (along other countries, for instance, the United States) up-to-date textbooks for practitioners, under the name of *Best Statistical Practices,* that would keep track of new methods and techniques in the field of applied statistics and lay particular emphasis on the needs of official statistics. This type of instructions for daily work of a statistician would help to raise the prestige of both statisticians and institutions employing them (both public and private ones, such as think-tanks).

Initiation into the profession - motivation and preparation

Learning statistics should start at earlier stage of education than higher education, according to the panelists. Elements of statistics should be taught already in general secondary schools (especially in economic schools), as argued by Prof. J. Dziechciarz. Referring to the need to elicit emotional motivation to follow this difficult field of study (mentioned by Prof. Domanski) Prof. M. Rocki stressed the need to create such attitudes (emotions) even at the level of primary school (which was also addressed by Prof. J.L. Wywial), pointing at the same time to economic universities, such as Warsaw School of Economics which pioneered with some initiatives to this aim, launching programs such as Children's Economic University and the Academy of Young Economist for students of

primary and general secondary schools. And at the subject contests organized to enable the dissemination of issues among young people and to test the knowledge of statistics (e.g. the statistics contest or complement of mathematics or entrepreneurship contests). A good place to stimulate motivation to study statistics could be, according to Prof. Domanski, the Polish Statistical Association, which takes initiatives to "stimulate emotions" and organizes statistical competitions in several Polish cities.

Learning-teaching conditions

The use of larger amount of current, real-world data stored on CDs and conducting exercises in computer rooms which require not only passive use of statistical software, but making conscious choices based on theoretical knowledge was postulated by Prof. M. Krzysko (seconded by Prof. S.M. Kot). Recalling the limitations associated with the need to join groups because of financial constraints, Prof. L. Tomaszewicz stressed the importance of the learning conditions – which consists also of insufficient length of studies – the worst candidate can be made a gem would s/he be met with adequate teaching environment. Noting that the 6-semester bachelor studies do not provide a complete study program that could educate analysts with good knowledge of statistics, Prof. Rocki stressed that the sequence of the subjects itself – analysis, algebra, probability theory, mathematical statistics and econometrics – requires five semesters, which leaves no room for other subjects enriching the knowledge of the graduate. Possible solutions are (i) the struggle for a uniform courses leading to a master's degree, and (ii) the formulation of the university qualification framework so as to ensure proper education. In addition, efforts should be taken for participation in determining the title of the professional – the proper nomenclature should reflect the knowledge and skills of graduates (e.g., along University of Minnesota offering a *Master of Statistics* degree). Bearing in mind the regulation defining the duration of studies of "at least" six semesters, Prof. Domanski shared this view and pointed to the possibility of extending the studies and encouraged taking the initiative and striving for uniform courses leading to a master's degree.

The bottom line: Poor quality of modern textbooks of statistics in Poland results from the use of out-of-date information, referring to obsolete problems, too much focus on descriptive statistics and the lack of a fair selection of textbooks, the consequence of which is the presence of elementary errors in them. In view of this situation it would be reasonable to appoint a consultative team of experts charged with responsibility assess and recommend for distribution only the highest quality textbooks. Providing background and interest in knowledge

oriented towards statistics should be preferably taken as early as in primary school. One should also take steps to extend the duration of the studies, ideally bringing them to a uniform course (that leads to a master's degree).

4. The quality of teaching

The quality of university teachers

As a precondition for producing statisticians as good professionals the availability of quality teachers must be seen, and the Polish educational system has not worked out the mechanism for preparing people to teach statistics properly at each level. It was the main concern of Prof. J. Dziechciarz, who addressed a gap between real-world problems and formal approaches due to the fact that persons teaching statistics are typically graduated in mathematics and educated in the area of advanced statistics. This situation causes two kinds of obstacles:

(i) removing from the curriculum the basics of statistics, mainly tools of descriptive statistics, (ii) the lack of teaching suited to the subject matter disciplines, to the specific needs of economics, medicine, social sciences, etc. In teaching statistics there is a necessity to become familiar with the specific objectives of a given field (which was emphasized also by Prof. Kot).

A step in overcoming problems related to the lack of subject matter specialists in the profession of a statistician has been made by universities which begun introducing new field of studies such as "Commercial Analyst" or "Business Analyst," taking into account all the elements essential to practicing statistician. It is worth noting that members of the Polish Accreditation Committee should conduct the assessment of the quality of education in a way comprising verification of the competence of all teachers given courses in statistics (not only of the staff members, being included into so-called the academic minimum).

The quality of students – the recruitment policy

One of the causes of deterioration of quality of students was seen by the panelists in the student recruitment policy due to the lack of an entry exam that would make it possible to select those who are predestined to study specific subjects from those who simply want to study. According to Prof. Rocki, who addressed this concern, confining the recruitment criteria to the obligatory 'matura' examination results in admitting students who are good at graduation subjects but not necessarily prepared to study the specific subject.

Apart from the selection of candidates, flexibility and multidisciplinarity is advisable in teaching, whereas a statutory need is to enroll the student on a

particular course of study, which may result in wrong choices and waste of public money in the course of studying. Prof. Rocki indicated to results of a survey carried out by the Warsaw School of Economics which showed that in significant proportion the students entering university do not know what choice to make, while among those who have made a choice, one third switches to other fields than the one previously chosen. Therefore, flexibility in studying is necessary, as well as the waver from the requirement to declare the field of study at an early stage – this would give the opportunity for a more informed and consciously made choice of statistics as a main subject of study.

In conclusion: The practical problems are the lack of mechanisms to prepare for training statisticians, the teaching process not suited to the different subject matter areas, as well as the lack of mechanisms for selecting candidates for studies and insufficient flexibility to selecting and changing a field of study. Therefore, the necessity to take into account the specificity of particular field of study is stressed, along with including in the curriculum all the key elements and special tools enabling students to practice statistics. The quality of a student should be improved by introducing university entrance examination and removing the need for the selection of the field of study at an early stage of education.

5. The issue of curriculum

Apart from the question of who is to teach statistics, it is important to ask *what?* should be included in the curriculum. Prof. Wywial talked about dubious legitimacy of the profession that requires specialized education in higher education, in case of the absence of a subject (both I and II degree) that teaches basics of statistical inference in a reliable way. Computer science and econometrics were to be such a subject, which was, however, targeted towards experts in the application of quantitative methods in economics and towards computer science (which was also addressed by Prof. Rocki).

Another problem arises from sometimes observed attempts to remove from the curriculum quantitative methods, and statistics in particular, due to commercial focus on rapid training of graduates (which was also pointed by Prof. L. Tomaszewicz supported by other panelists). What is important, according to Prof. K. Jajuga, is also teaching statistics in subjects other than computer science and econometrics, and especially in these with the reduced number of hours. Prof. Tomaszewicz stressed that the statistical and econometric core of the fields of study such as computer science and econometrics should be maintained by inter-subject actions, even in defining the occupation of a statistician by means of effects for statistical training, introduced pursuant to the National Qualifications

Framework. This would refine the graduate profile, defined differently depending on a variety of specialties, which cover the most difficult challenges of modern data analysis. The requirement set out in the description of a graduate profile, such as "knows basic statistical methods..." is not enough to create a statistics-based program on this basis, as pointed out by Prof. Rocki.

6. Professional and occupational aspects of statistics

If we assume that there is no single profession of a statistician, we recognize that there is no single model of education, and that there is a need to adjust teaching to the type of working areas. An example may be the category of *official statistician* invoked by the panel organizer. The occupation of a statistician working in public statistics institutions has been recognized by Eurostat as important and specific, and an initiative for the program called EMOS "The European Masters in *Official Statistics"* was undertaken. It was initiated by Eurostat currently conducting a series of meetings with the National Statistical Institutes. As explained by Ms. A. Kupis-Fijalkowska (assistant of Prof. Domanski, an expert for Poland and neighboring countries), EMOS project is to bring to universities, starting from 2014, an additional educational module for the final year of Master's Degree, which would allow for educating a statistician prepared to work in the statistical office, or at Eurostat.

A model statistician as seen by mathematical statistics may not be the same as the *official statistician*. They both are needed but, according to Prof. Okrasa, one cannot expect the same qualifications from both of them. It is worth to quote the conclusions of one of the surveys on what statisticians think about their profession, presenting desirable skills of a statistician: (1) mathematical basics, (2) the ability of critical thinking, (3) the ability of active learning, interacting with representatives of other areas, (4) the ability of active listening (communication and contact with a user). Moreover, statisticians asked in surveys why they want to be statisticians generally appreciate independence (higher than salary), autonomy of their work and high esteem among various types of employers (generally higher than of other staff), as well as prestige in the society.

When considering questions about the nature of statistics, Prof. Wywial came to the conclusion that statistics should be considered primarily in terms of profession (referring to the Polish language dictionaries which define profession requiring acquired qualifications in higher education institutions as a concept somewhat different from the term "occupation"). Its subject is the empirical verification of theories produced in other sciences and support in recognition of the characteristics of population, which are the target of studies in other

disciplines. A real solution could be the introduction of elite ordered studies and improvement or maintaining the level of studies on existing subjects, such as computer science and econometrics as well as economic analysis.

Given the high prestige of the profession of a statistician and the growing concern about bringing the state-of-the-art statistical knowledge and skills to official statistics it was suggested (W. Okrasa) to consider launching in Poland a kind of competition-based scholarship for researchers with proven achievements (modeled on the National Science Foundation's program of *Senior Researc*h *Fellowship* at the US Bureau of Labour Statistics and at the US Bureau of Census), who would be working on problems being currently of the focus of official statistics. Such problems are, for example, new modes of conducting census or administrative registers vs. statistical data collection system, or whether and how *electronic future* can provide a threat to statisticians as professionals and to the institutions of official statistics, given that *big data*, generated by other systems of information, are channeled outside of the area remaining under the control of institutions responsible for public statistics.

In spite of being discussed as essentially country-specific, the above issues have been actually internationally recognized for decades, just to mention the presentation by Hartley as the American Statistical Association (ASA) President (1979 – entitled *Statistics As a Science and As a Profession*), who tried to solve a traditional trade-off between professional equipment of mathematical and applied statisticians. While rejecting claims for 'more mathematics' to assure the quality solution of a real-world problem – or to blame its insufficiency for criticism ("standard of our papers is low", he quotes) – he pointed to cooperation between statistician and subject matter specialist as a way to balance between deductive (formal) and inductive (empirical) components of the statistics as a profession (science) and as an occupation (in terms used here). He was seconded by one of his distant successor, J. Stuart Hunter (the ASA President in 1993) who emphasized in his presidential address that while "a professional in statistics is a person whose everyday work consisted of *making sense of data*", there are also others – "the builders of statistical theory and makers of statistical tools" – who are "vital to the health of the statistical profession" [italic added][1].

Avoiding temptation to go beyond the scope of the referred panel's discussion, one may indicate the numerous and thematically reach sessions devoted to that issues being vigorously debated at such prominent meetings as the World Statistics Congress held last August in Hong Kong. It does prove that new approaches are continuously sought in most of the nations as some titles inform

[1] J. S. Hunter, 1994. Statistics As a Profession. Journal of the American Statistical Association Vol. 89, Issue 425, pages 1–6.

about that – let us mention only a few out of several dozen presented over there[1], for instance: Research on the Modes of Statistical Education and Training in China (Xia Rongpo) – Changing Educational Framework in the Transition to New Educational Standards at Russian Universities of Life Science and their Impact on the Teaching of Statistics (Galina Kamyshova and Lyman McDonald, Russia) – Engaging Students in Statistics Education: situated learning in statistics projects (Pieternel S. Verhoeven, the Netherlands) – Good Practice in Using Statistics in Statistics Education Research (Neville Davies and Gemma Parkinson UK) – New Perspectives: A Statistician and a Statistics Educator Discuss the Lessons Learned from Cross Disciplinary Sojourns (Jennifer J. Kaplan *et al.*, USA) – Radical Statistics: Teachers and Students on the Highwire (Bruno de Sousa *et al.*, Portugal and Spain).

A systematic overview of the problems and approaches discussed on the global scale – i.e. at meetings like the 59th WSC – would provide the needed contextualization for those identified as of the key importance during the panel.

[1] 59th ISI World Statistics Congress in Hong Kong, http://www.isi2013.hk/en/index.php.

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 329–336

REPORT

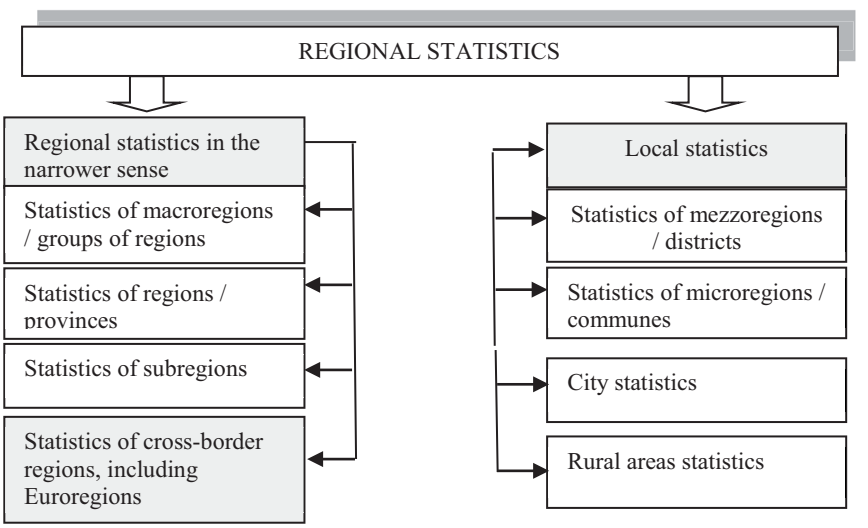
The Regional Statistics – Current Situation and Fundamental Challenges[1]

Regional statistics constitutes an integral and very important component of public statistics regarding its organizational and substantive aspects. **Basic data aggregation method** represents the main criterion for its distinguishing, which facilitates the positioning of phenomena in regional space and, to a lesser extent, the analysis content.

The method for providing regional statistics definition depends on the accepted interpretation of statistics itself – whether statistics is interpreted as *tabular* (data sets), as *functional* ("reporting" interpretation), or as the *branch of science*.

Following tabular interpretation regional statistics represents data sets referring to regions and facilitating social, economic and environmental phenomena understanding in different systems of territorial units. In this sense regional statistics is understood as data sets referring to phenomena occurring at different spatial levels, which is illustrated by picture 1.

Picture 1. The scope of "regional statistics" concept



Source: Author's compilation.

[1] The author's reflections based on his participation in three thematic sessions at the Congress of Polish Statistics – Poznań 2012.

Regional statistics also refers to data sets covering all or some spatial cross-sections of the described phenomena. Additionally, there are no substantial contraindications to include into local statistics (as the type of regional statistics) also these statistics which cover smaller than a district or commune territorial units, e.g. locations.

In relation to *functional interpretation* **regional statistics** refers to collecting, accumulating and updating statistical data, their processing, provision and dissemination in different spatial systems. Finally, with reference to the third interpretation, **regional statistics** represents the scientific branch of statistics which deals with detecting and analyzing certain regularities occurring in mass processes and also characteristic for phenomena studied in different territorial systems.

Currently, Regional Statistics is focused on the following three fundamental problems:

- current situation assessment and the identification of basic challenges put before Polish and European regional statistics,
- data bases as well as regional and local development monitoring,
- city and rural areas statistics.

Basic challenges

The main task of regional statistics is to provide the quantitative description of administrative and functional identities at the level of regions. It refers to the countrywide statistics as the hierarchical system of geographical and infrastructural economic, demographic and social relations.

The second important problem is the assessment of information coverage necessary for the state and local self-government functioning in market economy after 1990. The system of regional statistics should consider possibilities for information systems construction which extend outside administrative divisions based on NUTS. It mainly refers to information provision for special attention areas, such as: economic zones, ethnic groups, environmental advantages, areas at risk of flooding, etc., or goal oriented areas, e.g. generating cross-border statistics (including Euro-regional one), metropolitan statistics, etc.

The development of statistical research methodology in cross-regional sections at the beginning of the 21st century, covering problems which occurred in the process of preparation for Agricultural Census 2010 and National Census 2011 represent the third challenge of Regional Statistics. It mainly refers to conclusions resulting from the assessment of statistical information quality obtained based on traditional censuses and representative research without using alternative data sources collected by state administration.

Integration of sources and special sample surveys applying synergy effect, as well as problems referring to regional data quality and the criteria of assessing the statistical survey results are also considered very important problems.[1]

Challenges for Regional Statistics, at the background of the debate focused on cohesion policy are regarded as the next problem of this type of Statistics. The adoption of the Lisbon Treaty has changed the scope of cohesion policy where, apart from social and economic problems, a significant spatial dimension has also appeared. In order to specify due objectives, in order to perform both monitoring and evaluation of this policy, it is necessary to put more emphasis on the role of public statistics as the source of indispensable data. The organization of statistics in Poland keeps adjusting to the changing demand of data recipients. Integrated operations of numerous public statistics provision units are fundamental for effective information service of database users. These changes result in the establishment of Regional Surveys Centres which focus not only on initiating regional surveys but mainly on immediate response to the reported information demand, especially presented by territorial self-government units. The ongoing improvement of publicly available databases, such as, e.g. Local Data Bank (LDB), must also be remembered.

Significant methodological efforts in regional statistics have been undertaken. In the context of strategic assumptions of implementation monitoring, in line with the carried out development policy, the preparation of new rural communes typology constitutes an important initiative the realisation of which undergoes its final phase.

Publishing activities represent an important component of information and statistical data dissemination. Among numerous regional publications *Rural areas in Poland* is worth mentioning. Additionally, *Statistical Vade Mecum of Local Government* will be issued for the second time since it received very good reviews from the target group.

Challenges put before Regional Statistics are determined not only by socio-economic or political transformations, but they also result from new technical capacities. Suggestions of solution in many thematic areas may become significantly enriched by the cooperation between academic circles and statistics practitioners – it mainly refers to the problems of methodological nature, better specification of information needs typical for particular groups of recipients and

[1] The presentation by Jan Paradysz (Poznań University of Economics – Regional Statistics Centre) - *Regional statistics: state, problems and towards of development,* materials of the Congress of Polish Statistics, Poznań 2012.

also broadly understood analytical work, which probably constitutes the most difficult issue[1].

Database and indicator systems of regional and local development monitoring

The condition for establishing adequate sets of sustainable development indicators referring to a given country, provinces, districts and communes is necessary to substantiate this development concept providing legal and strategic substance by defining qualities, objectives and principles of sustainable development. Indicators represent basic tools for sustainable development monitoring which, in measurable way, unveils the core concept of such development. Therefore, the establishment of indicator-based monitoring system for sustainable development is expected to provide possibly the most precise and unambiguous answer to the following two crucial questions:

- what sustainable development means in an indicator-oriented sense;
- what substantial results illustrate this process and its progress in both spatial and temporal system.

The concept of monitoring turns out crucial at this point since it combines sustainable development measurement with life quality measurement at the local level within the framework of specific and constantly developing "statistical" initiative of the local government sector under the name of Local Government Analysis System (SAS). At the third stage of this system development (i.e. since 2007), apart from the objective already being carried out (public services quality research), a new goal, significantly extending the scope of SAS, was defined – the strengthening of indicator monitoring systems for local development policies/strategies by means of two additional indicator modules construction: sustainable development and inhabitants' life quality. The module of sustainable development indicators for the purposes of the system can be distinguished, at the background of other indicator systems, by the possibility of calculating sustainable development synthetic measures for domains and orders. Within the module four analysis levels of sustainable development indicators are possible to carry out and adjust to the monitoring needs of local strategies implementation. The system also offers life quality survey to local governments, including the survey methodology and the report structure. This system module lists, altogether,

[1] The presentation by Dominika Rogalińska (Central Statistical Office) - *Challenges for regional statistics at the background of debate focused on cohesion policy*; materials of the Congress of Polish Statistics, Poznań 2012.

over 250 sustainable development indicators and a relatively balanced division of indicator set into three orders: social, economic and environmental-spatial order. It is supported by the Regional Data Bank (RDB) of Central Statistical Office (CSO)[1].

Changes in RDB represent one of the major manifestations featuring improvements in the publically available database. These changes mainly refer to:

- the availability of selected qualities also for the level of statistical locations,
- the extension of short-term data scope,
- significant improvements in interface functioning,
- recognition of RDB resources as an important source of information about sustainable development implementation at local and regional level.

A universal module was constructed within the framework of RDB which facilitates sustainable development monitoring at the lower level than the national one. Sustainable development indicators developed by Eurostat became the point of reference for this module construction. The module prepared within RDB does not present the set for the purposes of a substantive development strategy monitoring at either local or regional level, but it offers a basic set of indicators which may become the background for sustainable development analyses at the lower than national level. The established module is supposed to constitute a type of a "core" facilitating the assessment and comparison of particular territorial self-government units.

Such module has to be composed of indicators which are covered by statistical data and these data are comparable for all territorial self-government units. Apart from that, due to the specific nature of particular units, the module may be supplemented by indicators crucial from the perspective of a given territorial unit. The module should be extended by the so-called leading groups in a given area – self-government sector, entrepreneurs and social organizations representatives. The prepared module consists of topics which are essential for sustainable development monitoring at regional and local level. Each topic is divided into sub-topics under which substantive indicators were listed[2].

[1] The presentation by Tadeusz Borys (Wrocław University of Economics) and Tomasz Podkański (The Association of Polish Cities) - *Monitoring regional and local development*, materials of the Congress of Polish Statistics, Poznań 2012.

[2] The presentation by Bartosz Bartniczak (Wrocław University of Economics) – Module of sustainable development indicators in Regional Data Bank, materials of the Congress of Polish Statistics, Poznań 2012.

The statistics of cities and rural areas

Practical application of **Geographic Information System (GIS)** is of crucial significance in city and rural areas statistics as an information system useful for entering, collecting, processing and the visualisation of geographical data since one of its many functions is to provide support for the decision-making processes. GIS is the result of several decades of ongoing methodological changes in geography and, obviously, also the rapid development of computer technology influencing database (information sets) management methods. The establishment of GIS is the effect of combining efforts carried out in different fields: geography, cartography, geodesy, computer technology, electronics and statistics.

GIS is applied, among others, in constructing spatial information sets (records, registers) and also in their processing and analysis. Another application group is related to information processing about the distribution of all types of phenomena, especially these featuring extensive variability in time. This group of applications also covers GIS usage for statistical data analysis and presentation, such as, e.g. crime threat, diseases occurrence, land use structure. GIS may also present a very useful tool for processing data about technical infrastructure of a given area, i.e. water, gas, energy supply system network, communication lines. Such data require frequent modifications. Additionally, their great accuracy and timeliness is indispensable.

GIS was used for the purposes of public statistics in the course of two censuses: Agricultural Census (AC 2010 – from 01 September until 31 October 2010) and National Census of Population and Housing (NCPH 2011 – from 1 April until 30 June 2011). In the process of these censuses modern and cheaper than before solutions of data collecting from 16 data holders, coming from 25 information systems (including administration ones) and electronic communication tools were also extensively used, which completely eliminated paper forms. This allowed for smaller burden on respondents and for cutting census materials printing costs. The following data collecting channels were applied in the above censuses: administration sources, the Internet (CAII – Computer Assisted Internet Interview, online self-enumeration), telephone interview (CATI – Computer Assisted Telephone Interview), by an enumerator (CAPI – Computer Assisted Personal Interview) equipped with a portable hand-held terminal.

To support the census a Computer-based Census System (CCS) was introduced which integrated different technologies (from applications installed on

mobile terminals through applications managing and supporting phone interviews, to specialized database, data warehouses and analytical-reporting tools). Additionally, CCS provided solutions ensuring high level of data processing security, due organizational procedures were implemented which obliged census participants to maintain statistical secrecy and protect personal data.

The innovation introduced in the course of both censuses was the application of GIS technology at every stage. Digital maps constituted essential tools for census enumerators (regarding spatial orientation, sampling frame verification, etc.), for communal leaders, provincial and central dispatchers who could perform remote verification of census advancement, as well as the route or the current location of the census enumerator. GIS applications were using materials obtained by PZGiK *[National Geodetic and Cartographic Resources]* (ortophotmaps, borders of provinces, districts and communes, names of locations), borders of statistical regions and census enumeration areas, and also statistical address points prepared by the official statistics services, as well as registered parcels layer (ARiMR *[Agriculture Restructuring and Modernization Agency]*), roads and streets[1].

GIS represents an effective and attractive tool for statistical data presentation and analysis. All statistical data refer, in a sense, to geographic space and therefore spatial aspect should also be considered in their analysis. Typical information published in tabular form, without presenting it in the form of maps, does not allow for noticing numerous interdependencies. Therefore, statistics uses, more and more often, GIS software which facilitates statistical data visualization by means of maps and also advanced spatial analyses of these data[2].

The application of data from administration sources in city statistics supported by GIS tools represents a very important direction in the development of Regional Statistics. The growing demand for detailed and also complex information about cities and problems with meeting them result in the fact that public statistics keeps looking for new sources of their acquisition. The study of information-oriented demand also indicates the need for analyzing cities perceived in certain separation from administrative borders (functional area, wider urban zone, internal urban zones). The need for a different perspective in relation to a "city" also results

1 The presentation by Janusz Dygaszewicz, Magdalena Jaczur-Knapek and Amelia Wardzińska-Sharif (Central Statistical Office), Two censuses: AC 2010 and NCPH 2011 and GIS in public statistics, materials of the Congress of Polish Statistics, Poznań 2012.

2 The presentation by Paweł Chlebicki (ESRI Poland) - ArcGIS as strongly tool of statistical data visualization and spatial analyse, materials of the Congress of Polish Statistics, Poznań 2012.

from such strategic documents as National Regional Development Strategy, National Spatial Development Concept and the European Commission projects focused on urban development support. These reasons became the leading inspiration to extend the application of data included in administrative sources as supplementary for statistical surveys.

Prepared by:
Tadeusz Borys
Wrocław University of Economics

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 337–340

REPORT

The 22nd Didactic Conference on Teaching Quality Evaluation Methods, 10-11 June 2013, Łódź, Poland,

The 22nd edition of Polish nationwide **Didactic Conference** was titled ***Teaching Quality Evaluation Methods*** (***Metody oceny jakości nauczania***) and was held in **Łódź** from 10th to 11th June 2013. The Conference was organized by the **Institute of Statistics and Demography** of the University of Łódź, with the cooperation of the **Institute of Econometrics** of the University of Łódź, under the honorary auspices of the **President of the City of Łódź, Ms. Hanna Zdanowska**. **The City of Łódź Office** and **StatSoft Polska Sp. z o.o.** were official partners of the 2013 Didactic Conference. **Prof. Czesław Domański** held the function of the Chairman of the Scientific and Organizing Committees. **Aleksandra Kupis-Fijałkowska**, M.Sc. was the first secretary of both Committees and **Anna Majdzińska**, M.Sc. was the second secretary of the Organizing Committee.

The Didactic Conference Organizers aimed to provide an opportunity for academic teachers and tutors, especially in fields such as mathematics, statistics, econometrics, finance, economy and national accounts, demography and demometry, to present their teaching problems and reflections, as well as analyze and discuss recent scientific observations in this area. The 2013 edition of the Conference was focused on the Teaching Quality and methods of its evaluation.

The Conference was opened by the Chairman of the Scientific Committee **Prof. Domański**. On the Opening session the next speakers were respectively: **Prof. Zofia Wysokińska**, Pro-Rector in Charge of International Affairs, who was the representative of the Rector of the University of Łódź, **Prof. Włodzimierz Nykiel** and **Prof. Janusz Świerkocki,** Pro-Dean for International Cooperation and Course of Studies of International Economic Relations, Economics, who was representing the Dean of The Faculty of Economics and Sociology of the University of Łódź, **Prof. Paweł Starosta** and the Pro-Dean for Research, **Prof. Agnieszka Rossa**.

Briefly, the key statistics of the **2013 Didactic Conference** look as follows: 63 participants altogether, including the Senator of the Republic of Poland and the President of Polish Accreditation Committee, **Prof. Marek Rocki**, the representative of the Central Statistical Office of Poland, **Prof. Włodzimierz Okrasa**, academic teachers, scientists and Ph.D. students representing 12 Polish

universities and practitioners; 5 plenary sessions with 13 presented papers, 1 Jubilee session and 1 Discussion Panel.

The first session was organized to celebrate The International Year of Statistics and was led by **Prof. Okrasa** (Cardinal Stefan Wyszyński University in Warsaw, Central Statistical Office of Poland). It was started by **Prof. Rocki** (Senator of the Republic of Poland, Polish Accreditation Committee, Warsaw School of Economics), who spoke about *The Condition of Statisticians Education in Reports of the Polish Accreditation Committee*. The second speech *Statistics – Science About Information Ensuring the Quality of Teaching* was given by **Prof. Domański** (University of Łódź).

These two mentioned above presentations formed a solid introduction to the next part of the Conference, which was a *Discussion Panel Statistics as a Profession – Statistician as an Occupation,* prepared and moderated by **Prof. Okrasa**. The following experts were invited by the Organizers to take part in it (by alphabetical order): **Prof. Domański** (University of Łódź)**, Prof. Józef Dziechciarz** (Wroclaw University of Economics), **Prof. Mirosław Krzyśko** (Adam Mickiewicz University in Poznań), **Prof. M. Rocki** (Senator of the Republic of Poland, Polish Accreditation Committee, Warsaw School of Economics), **Prof. Janusz Wywiał** (University of Economics in Katowice). This event was open for everyone, therefore also other prominent statisticians, as well as less experienced ones and many representatives of other teaching fields, for example demography or econometrics, participated in the discussion. [The detailed information about this panel, as well as important conclusions and further recommendations can be found in this *Statistics in Transition - new series* issue by **the panel organizer**].

The second plenary session was led by **Prof. Rocki** (Senator of the Republic of Poland, Polish Accreditation Committee, Warsaw School of Economics) and the following papers were presented: *Quality measurement in the EU concept of the entrepreneurial university* by **Prof**. **Dziechciarz** (Wroclaw University of Economics), *The Interplay Between Entrepreneurship Education and Students' Entrepreneurial Intentions* investigated by **Jacek Białek** and **Agnieszka Kurczewska** (University of Łódź) and *Methods for Assessing the Quality of Teaching – Quantitative Methods for Economics and Business* by **Grzegorz Kończak** and **Jolanta Bernais** (University of Economics in Katowice).

The third session was chaired by **Prof. Krzyśko** (Adam Mickiewicz University in Poznań) and 3 papers were discussed here: *Analysis of the Results of the Compensatory Courses in Mathematics Realised in the Affiliate Branch of the Warsaw University of Technology in Plock* by **Izabela Józefczyk**, **Romuald Małecki** and **Roman Rumianowski** (Warsaw University of Technology, Affiliate Branch in Plock), *Evaluation of Quality of Teaching at the Example of the Faculty of Economics of the West Pomeranian University of Technology in Szczecin* by **Iwona Bąk** and **Katarzyna Wawrzyniak** (West Pomeranian University of Technology in Szczecin) and *Modernization of mathematics courses*

in departments of economics, by supplementary remarks about elements of complex analysis by **Janusz Kupczun** (University of Łódź).

The next session of the Conference was a special Jubilee one and was led by **Prof. Rossa**. It was organized by the **Institute of Statistics and Demography** of **University of Łódź** with cooperation of the **Polish Statistical Association.** This session took place to celebrate the 45th anniversary of **Prof. Domański**'s scientific and academic work. **Prof. Domański** is the author or co-author of almost 200 research papers and 22 books in the field of mathematical statistics, financial statistics and risk theory. He also acted as supervisor of numerous PhD dissertations and is recognized as a great teacher of many generations of students. **Prof. Domański** was awarded several prestigious prizes. **Prof. Jerzy T. Kowaleski** and **Prof. Alina Jędrzejczak** (University of Łódź) prepared *Scientific Curriculum Vitae of Professor Czesław Domański*. The Authors thoroughly described the biographical profile, including scientific publications and academic achievements of the Jubilarian. Also, **Prof. Jędrzejczak** presented the Session's Participants with a video with the interview of **Prof. Domański** and **Prof. Walenty Ostasiewicz** about Statistics given to the Reporter of the Polish Television, **Monika Włodarczyk**.

On the second day of the 22nd Didactic Conference, 2 sessions had place. Respectively, the fourth plenary session and the first one on the 11th June was chaired by **Prof. Andrzej Sokołowski** (Cracow University of Economics). The following papers were presented: *Didactics and Changes in the Student Community – A Review* by **Lilianna Jarmakowska-Kostrzanowska** (University of Łódź), *Selected Statistical Analysis Methods in Assessment of the Learning Outcomes in the Example of the Descriptive Statistics Exam Results* by **Małgorzata Misztal** (University of Łódź) and *Implementation of Continuous Quality Improvement Method in Higher Education* prepared by **Elżbieta Zalewska** (University of Łódź).

The sixth plenary session, which was led by **Prof. Kowaleski** (University of Łódź) was devoted to the teaching of demography and its evaluation at the higher education level. During this session the following papers were discussed: *Methods of Evaluation in Teaching Demography and Related Subjects* – a presentation delivered by **Anna Majdzińska** on behalf of its Author, **Milena Lange** (University of Łódź), and *Essence of Teaching Demography in the Context of Trends and Country's Development Strategy* prepared and presented by **Anna Majdzińska** and **Anna Wierzbicka** (University of Łódź).

The 22nd edition of the **Didactic Conference *Teaching Quality Evaluation Methods*** was closed by the Vice-Chairman of the **Institute of Statistics and Demography** of the University of Łódź and the Chairman of the Łódź Branch of the Polish Statistical Association, **Prof. Jędrzejczak**. On behalf of the Scientific and Organizing Committees, **Prof. Jędrzejczak** summarized the Conference as very effective. The importance of the high quality education and the problem of its reliable assessment were emphasized. Also, in the context of the Discussion

Panel led by **Prof. Okrasa**, the role of Statistical Literacy in the modern societies and the need of popularizing profession of statistician were mentioned.

The publication entitled ***Acta Universitatis Lodziensis, Folia Oeconomica. Metody oceny jakości nauczania (Methods of Assessment of the Quality of Teaching)*** is the publishing output of the 2013 Didactic Conference. It was edited by **Alina Jędrzejczak** and **Aleksandra Kupis-Fijałkowska** (printed by the Łódź University Press).

The next round of the meeting series known as the Didactic Conference will be devoted to ***Academic teacher and new educational challenges.*** It is planned on **June 9th-10th, 2014** (also in **Łódź) - details will be given in a separate announcement**.

Prepared by
Aleksandra Kupis-Fijałkowska
University of Łódź

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 341–342

REPORT

Summer School of Baltic-Nordic-Ukrainian Network on Survey Statistics 2013

The Summer School of Baltic-Nordic-Ukrainian Network on Survey Statistics 2013 was organised in Minsk (Belarus) on June 13–19. It was the seventeenth annual event organised by the Baltic-Nordic-Ukrainian Network on Survey Statistics (https://wiki.helsinki.fi/display/BNU). The objectives of the summer school were to present recent achievements and results, learn from teachers and colleagues, discuss the future, share opinions and experience, and strengthen contacts between survey statisticians.

There were 32 participants from seven countries – Belarus, Estonia, Finland, Latvia, Lithuania, Sweden, and Ukraine. The audience of the workshop was diverse – consisting of undergraduate students, research students, university teachers and practising statisticians.

There were four main lecturers:

1. Seppo Laaksonen (University of Helsinki, Finland), two lectures entitled "Non-response adjustment in surveys";
2. Risto Lehtonen (University of Helsinki, Finland), two lectures entitled "Analysis of complex survey data: an overview";
3. Aleksandras Plikusas (Vilnius University, Lithuania), two lectures entitled "Calibration";
4. Imbi Traat (University of Tartu, Estonia), two lectures entitled "Main principles of survey sampling".

There were five special invited lecturers:

1. Natallia Bokun (Belarusian State Economic University, Belarus), the lecture "Multivariate sample: main principles, design, estimation";
2. Oksana Honchar (National Academy Statistics, Accounting and Audit, Ukraine), the lecture "Sample surveys in Official Statistics: understanding quality";
3. Danute Krapavickaite (Vilnius Gediminas Technical University, Lithuania), the lecture "Software for survey sampling and analysis";
4. Gunnar Kulldorff (University of Umeå, Sweden), the lecture "The Statistical Science Is International – And Survey Statistics Is Cool and Hot";

5. Mārtiņš Liberts (Central Statistical Bureau of Latvia), the lecture "The choice of a sampling design regarding survey cost efficiency".

The first day of the summer school was devoted to lectures in Russian. The lectures in Russian were given by Danute Krapavickaite, Aleksandras Plikusas, Elena Zapatrina, and Victor Levkevich. There were additionally thirteen participants from Belarus taking part during the first day of the summer school. Nineteen presentations of contributed papers devoted to survey statistics were presented by the summer school participants and discussed by invited discussants. The most of the materials presented during the summer school is available at the summer school website (http://nir.bseu.by/BalticNordicConference/Main%20Page.html).

The sponsors of the workshop were the Nordplus Programme of the Nordic Council of Ministers and the International Association of Survey Statisticians. The next annual event organised by the network will be the workshop in Tallinn (Estonia) in August 2014.

Prepared by
Mārtiņš Liberts
University of Latvia

STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 343–348

THE INTERNATIONAL YEAR OF STATISTICS (IYS) STATISTICS 2013[1]

GŁÓWNY URZĄD STATYSTYCZNY

POLSKIE TOWARZYSTWO STATYSTYCZNE

STATISTICS2013.ORG
INTERNATIONAL YEAR OF STATISTICS

URZĄD STATYSTYCZNY W ŁODZI

INSTYTUT STATYSTYKI I DEMOGRAFII UNIWERSYTETU ŁÓDZKIEGO

The year 2013 is celebrated as the International Year of Statistics (American Statistical Association, Institute of Mathematical Statistics, International Biometric Society, the International Statistical Institute, the Bernoulli Society, and the Royal Statistical Society, are the founding Organizations). National and international organizations from 124 countries from all over the world are involved in the celebration of the IYS. Among them are professional statistical associations, schools, universities, business institutions, research institutes and statistical offices.

Organizations participating in the celebration of the International Year of Statistics are contributing to the promotion of the importance of statistics by organizing conferences, seminars and other educational activities aimed at both the representatives of the scientific, business and media community, as well as at public representatives who are users of statistics (government data users), and policy makers, employers and students.

The International Year of Statistics is celebrated by a number of Polish institutions, such as:

Professional Societies

- The Scientific Club of Quantitative Methods, Faculty of Management, University of Gdansk;
- GAUSS Scientific association of mathematical statistics, Institute of Mathematics and Computer Science, Wroclaw University of Technology;

[1] Text authorized by Prof. Janusz Witkowski, President of the Central Statistical Office of Poland and Chairman of the Organizing Committee of the IYS Conference.

- Polish Mathematical Society, Wroclaw Branch;
- Polish Statistical Association;
- Statistical Forum - Polish First Statistical Forum.

Colleges and Universities

- Collegium Invisible, Warsaw;
- Cross-National Studies: Interdisciplinary Research and Training Program (CONSIRT);
- Maria Curie-Skłodowska University, Faculty of Mathematics, Physics and Computer Science, Lublin;
- Nicolaus Copernicus University, Faculty of Mathematics and Computer Science, Torun;
- University of Economics, Poznan, Department of Statistics, "Estimator" Student Scientific Association;
- University of Lodz, Department of Statistical Methods SKN SFI;
- University of Lodz, Institute of Statistics and Demography, Department of Statistical Methods;
- University of Management and Administration, Zamosc;
- Warsaw Management Academy.

Business

- ABE-IPS;
- Data Intelligence Poland.

Government

- Central Statistical Office of Poland, Warsaw.

Research Institutes and Journals

- Institute of Catholic Church Statistics, Warsaw;
- Scientific Foundation SmarterPoland.pl.

This initiative, which is supported by over 2,100 organizations around the world and about 20 institutions in Poland, involves numerous events aimed, in particular, at:

- Increasing public awareness of the impact of statistics on all aspects of social life (including increasing recognition of the contribution of statistics to improve the quality of life and advancement of the global community);
- Developing the profession of a statistician (especially among young pople);
- Promoting creativity and development in statistics and science of probability.

One of such events is the scientific conference "Statistics – Knowledge – Development" organized in Lodz on 17-18 October by the Central Statistical Office, Statistical Office in Lodz, the Polish Statistical Association and the Institute of Statistics and Demography of University of Lodz. It is the culmination of year-long activity of the Polish Programme Committee of the International Year of Statistics under the leadership of Janusz Witkowski - the President of the Central Statistical Office. The Programme Committee, which brought together distinguished representatives of the Polish science, has set itself the following tasks, among other things: (i) to initiate and popularize research projects undertaken on the occasion of the IYS , (ii) to initiate and review publications associated with the celebration of the IYS , (iii) to monitor the celebration of the IYS and to popularize its purposes in the media, (iv) to actively participate in the preparation of the scientific conference "Statistics – Knowledge – Development" - which is the highlight of the celebration of the International Year of Statistics in Poland.

As mentioned above, the Programme Committee of the International Year of Statistics includes prominent social scientists from all over Poland: the Chairperson Janusz Witkowski, the President of the Central Statistical Office; Deputy Chairpersons and Vice-Presidents of the Central Statistical Office - Halina Dmochowska, PhD, and Grażyna Marciniak, PhD, Advisors to the President of the CSO: Prof. Tadeusz Walczak, Prof. Włodzimierz Okrasa, and Bohdan Wyżnikiewicz, PhD; as well as directors of Statistical Offices and Departments of the Central Statistical Office: Prof. Jerzy Auksztol (Gdansk), Marek Cierpiał-Wolan, PhD (Rzeszow), Piotr R. Cmela, PhD (Łodz), Krzysztof Markowski, PhD (Lublin) and Renata Bielak (Analyses and Comprehensive Studies Department), Mirosław Błażej (Macroeconomic Studies and Finance Statistics Department), Katarzyna Cichońska (Deputy Director, Office of the President) and the other employees of Statistical Offices: Prof. Prof. Andrzej Młodak (Statistical Office in Poznan), Jacek Wesołowski (Methodology, Standards and Registers Department of CSO), Kazimierz Kruszka, PhD (Statistical Office in Poznan) and the leading academics: Prof. Prof. Czesław Domański - President of the Polish Statistical Association (Univeristy of Lodz), Elżbieta Gołata (University of Economics, Poznan), Adam Jakubowski (Nicolaus Copernicus University, Torun), Jan Kordos (Warsaw Management Academy), Mieczysław J. Król (University of Rzeszow), Franciszek Kubiczek (ALMAMER University), Tomasz Panek (Warsaw School of Economics), Zofia Rusnak (Wroclaw University of Economics), Andrzej Sokołowski (Cracow University of Economics), Mirosław Szreder (University of Gdansk), Waldemar Tarczyński (Szczecin University), Grażyna Trzpiot (University of Economics in Katowice), Agata Zagórowska (Opole University), Krzysztof Zagórski, (Kozminski University), Roman Zmyślony (University of Zielona Góra).

The two-day meeting of prominent representatives of social science from across Poland includes a number of interesting sessions listed below:

The Scientific Conference Statistics – Knowledge – Development

CONFERENCE PROGRAMME

Agenda:

16.10.2013 - day preceding the conference
Arrival of the participants at the hotel in the afternoon

17.10.2013 - Conference day 1
Conference opening
Plenary session I **Statistics in the face of global challenges**
Organizer and Chairperson Mirosław Szreder (University of Gdansk)
Plenary session II **Statistics friendly to all**
Organizer and Chairperson Czesław Domański (University of Lodz)
Plenary session III **Quality in statistics**
Organizer and Chairperson Jan Kordos (Warsaw Management Academy)

18.10.2013 - Conference day 2
Plenary session IV **Statistics in socio-economic practice**
Organizer Andrzej Sokołowski (Cracow University of Economics)

- ✓ *Social statistics;* Chairperson Janusz Witkowski (Central Statistical Office)
- ✓ *Economic statistics;* Chairperson Andrzej Sokołowski (Cracow University of Economics)

Plenary session V **Methodology of statistical surveys in theory and practice**

Organizer and Chairperson Elżbieta Gołata (Poznan University of Economics)

- ✓ *Poster session;* Chairperson Elżbieta Gołata (Poznan University of Economics)
- ✓ *Mathematical aspects of statistical research methodology;* Chairperson Elżbieta Gołata (Poznan University of Economics)

- ✓ *Representative surveys versus alternative sources of information;* Chairperson Eugeniusz Gatnar (NBP)

Closing of the conference

During the conference an exhibition ***Statistics of the Lodz Region*** will be presented (Statistical Office in Lodz, The Jozef Pilsudski Regional and Municipal Public Library in Lodz). A reviewed monograph containing the conference papers is also planned to be released.

The conference is held under the patronage of Ms Hanna Zdanowska - Mayor of the City of Lodz, the Ministry of Science and Higher Education, Polish Academy of Sciences and the National Bank of Poland. The partners of the project are the City of Lodz Office (the conference received funding under the project "Cooperation with higher education institutions" to promote the city as a centre of science) and The Jozef Pilsudski Regional and Municipal Public Library in Lodz. TVP Lodz and Radio Lodz are media partners of the conference.