STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Other articles

Comparative Surveys

Congress of Polish Statistics: The 100th Anniversary of the Polish Statistical Association

Congress Information

Volume 13, Number 1, March 2012

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 1—4

FROM THE EDITOR

The Spring 2012 issues of the *Statistics in Transition new series* is being released somewhat earlier than usually in order to contribute in this way to the upcoming Congress of Polish Statistics, which is under preparation to celebrate the hundredth anniversary of establishing of the Polish Statistical Association. Accordingly, in addition to Journal's regular sections – on *estimation and sampling issues* and *other articles*, and also on *comparative surveys* – a special congressional section is included in this volume, containing voices (‘occasional statements’) of several members of the Journal's Editorial Board, and of important Congress' information materials.

The first part starts with paper by **Przemysław Ciepiela, Małgorzata Gniado, Jacek Wesołowski** and **Małgorzata Wojtyś** on ***Dynamic K-composite estimator for an arbitrary rotation scheme*.** Authors begin with an overview of the properties of classical K-composite estimator proposed by Hansen, Hurwitz, Nisselson and Steinberg (1955) and intensively studied in Rao and Graham (1964). It gives an alternative solution to quasi-optimal estimation under rotation sampling when it is allowed that units leave the sample for several occasions and then come back. Since the K-composite estimator suffers from certain disadvantages – as being designed for a stable situation in the sense that its basic parameter is kept constant on all occasions and restricted only to a certain family of rotation designs – authors propose a dynamic version of the K-composite estimator (DK-composite estimator), without any restrictions on the rotation pattern. Although the proposed algorithm is simpler than the one for the classical K-composite estimator with optimal weights, it is precise, in the sense that it does not use any approximate or asymptotic approach.

Diwakar Shukla, Sharad Pathak and **Narendra Singh Thakur** in paper entitled ***Estimation of Population Mean Using Two Auxiliary Sources in Sample Surveys*** propose families for estimation of population mean of the main variable using the information on two different auxiliary variables, under simple random sampling without replacement (SRSWOR) scheme. Three different classes of estimators are constructed and examined with a completive study with other existing estimators. The expression for bias and mean squared error of the proposed families are obtained up to first order of approximation. Usual ratio estimator, product estimator, dual to ratio estimator, ratio-cum-product type estimator and many more estimators are identified as particular cases of the suggested family; theoretical results are supported by numerical examples.

In the next paper, ***Modified Estimators of Population Variance in Presence of Auxiliary Information*** by **Rajesh Tailor** and **Balkishan Sharma** proposed is an estimator of population variance using information on known parameters of auxiliary variable. It has been shown that using modified sampling fraction the proposed estimators are more efficient than the usual unbiased estimator of population variance and usual ratio estimator for population variance under certain given conditions. Empirical study is also carried out to demonstrate the merits of the proposed estimators of population variance over other estimators considered in this paper.

G. C. Tikkiwal and **Alka Khandelwal** in paper ***Crop Acreage and Crop Production Estimates for Small Domains – Revisited*** discuss the problem of advance and final estimates of yield of principal crops, at national and regional (State) levels, which are of great importance for country's macro level planning. For decentralized planning and for other purposes like crop insurance, loan to farmers, etc., the reliable estimates of crop production for small domains are also in great demand. This paper, therefore, discusses and review critically the methodology used to provide crop acreage and crop production estimates for small domains, based on indirect methods of estimation, including the SICURE model approach. The indirect methods of estimation so developed use data obtained either through traditional surveys, like General Crop Estimation Surveys (GCES) data, or a combination of the surveys and satellite data.

In paper ***Estimation of Population Mean in Post-Stratified Sampling Using Known Value of Some Population Parameter(s)*** by **A.C**. **Onyeka** a general family of combined estimators of the population mean in post-stratified sampling (PSS) scheme is presented, following Khoshnevisan et.al. (2007) and Koyuncu and Kadilar (2009), and using known values of some population parameters of an auxiliary variable. Properties of the proposed family of estimators, including conditions for optimal efficiency, are obtained up to first order approximations, and the results are illustrated empirically.

The second group of articles ('other articles') is opened by paper of **Edgar Mauricio Bueno Castellanos** on ***Nonresponse Bias in The Survey of Youth Understanding of Science and Technology in Bogotá.*** The Colombian Observatory of Science and Technology – OCyT – developed in 2009 a survey about understanding of Science and Technology in students of high school in

Bogotá, Colombia. The sampling design was stratified according to the nature of school (public or private). Two sources of unit nonresponse were detected. The first one corresponds to schools that did not allowed to collect information. The second source corresponds to students who did not assist during the days when survey was applied. Estimates were obtained through two different approaches. Results obtained in both cases do not show visible differences when estimating ratios; even though, some great differences were observed when estimating totals. Results obtained using the second approach are believed to be more reliable because of the methodology used to handle item nonresponse.

R. Sankle, **J.R. Singh** and **I.K. Mangal** in paper ***Cumulative Sum Control Charts For Truncated Normal Distribution under Measurement Error*** constructed Cumulative Sum (CUSUM) Control Charts for mean under truncated normal distribution and measurement error. For different truncation points and different sizes of measurement error tables have been prepared for the average run length, lead distance and the angle of mask. They analyze the sensitivity of the parameters of the V-Mask and the Average Run Length (ARL) through numerical evaluation for different values of r.

Elżbieta Gołata's paper ***Data Integration and Small Domain Estimation in Poland – Experiences and Problems*** has twofold objective, encompassing, on the one hand, a presentation of Polish experiences with the methodological issues considered currently as one of the most important – i. e., data integration (DI) and statistical estimation for small domains (SDE); and, on the other hand, it attempts to determine relationship between these two types of methods. Given convergence of the goals of both methods, SDE and DI (i.e., to increase efficiency of the use of existing sources of information), simulation study was conducted in order to verify the hypothesis of synergies referring to combined application of both groups of methods: SDE and DI.

The third section, *comparative surveys*, is represented in this volume by one item, by **Piotr Tarka's** paper on ***Customers Research and Equivalence Measurement in Factor Analysis.*** Author discusses the problem of between population validity of the measurement, when extracted factors may hard to be equally compared on the reflective basic level (unless all conditions of invariance measurement are met). Hence, implementation of customers research and any inter-cultural studies require a multi-cultural model describing statistical differences in both cultures with invariance as underlying assumption. In the article employed was a model for analysis of customers' personal values pertaining to hedonic consumption aspects in two culturally opposite populations.

Data were generated through survey conducted in two countries, in the following cities: Poland (Poznan) and The Netherlands (Rotterdam and Tilburg), using probability samples of youth. This model made it possible to test invariance measurement under cross-group constraints and thus examining structural equivalence of latent variables' values.

The section devoted to the Congress of Polish Statistics, concludes this volume.

Włodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 5

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, March 2012
Vol. 13, No. 1, pp. 7-20

DYNAMIC K-COMPOSITE ESTIMATOR FOR AN ARBITRARY ROTATION SCHEME

Przemysław Ciepiela[1], Małgorzata Gniado[2], Jacek Wesołowski[3] and Małgorzata Wojtyś[4]

ABSTRACT

Classical K-composite estimator was proposed in Hansen et al. (1955). Its optimality properties were developed in Rao and Graham (1964). This estimator gives an alternative solution to quasi-optimal estimation under rotation sampling when it is allowed that units leave the sample for several occasions and then come back. Such situations happen frequently in real surveys and are not covered by the recursive optimal estimator introduced by Patterson (1955). However the K-composite estimator suffers from certain disadvantages. It is designed for a stable situation in the sense that its basic parameter is kept constant on all occasions. Additionally it is restricted only to a certain family of rotation designs. Here we propose a dynamic version of the K-composite estimator (DK-composite estimator) without any restrictions on the rotation pattern. Mathematically, the algorithm, we develop, is much simpler than the one for the classical K-composite estimator with optimal weights. Moreover, it is precise, in the sense that it does not use any approximate or asymptotic approach (opposed to the method used in Rao and Graham (1964) for computing optimal weights).

1. Introduction

It is well known that, while looking for optimal estimators in surveys which repeat in time with the same time spacing, taking under account observations not only from the present edition of the survey (occasion) but also from previous occasions may significantly improve the quality of estimation.

[1] Bank PEKAO, Warszawa, POLAND

[2] Towarzystwo Ubezpieczeń na Życie "Warta", Warszawa, POLAND

[3] Główny Urząd Statystyczny and Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warszawa, POLAND, e-mail: wesolo@mini.pw.edu.pl

[4] Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warszawa, POLAND, e-mail: wojtys@mini.pw.edu.pl

For the best linear unbiased estimators (BLUEs) of the mean on a given occasion, to reduce time and memory requirements, it is desirable to have a recursive form for such an estimator which refers only to certain (possibly, small) number of optimal estimators from recent occasions and, additionally, observations from those occasions. Such a problem was completely solved in a seminal paper by Patterson (1950) for a family of rotation patterns which do not allow for a come-back of a unit to the sample, after leaving it for some occasions. The solution gives a formula for the BLUE of μ_h, the mean on the hth occasion, as a linear combination of the BLUE of μ_{h-1} and observations from $(h-1)$th and hth occasions.

However, in many practical surveys, the rotation pattern allows holes, i.e. some units stay in a sample for a number of occasions, leave it for a number of occasions, then return to the survey for a number of occasions. Important examples include the Current Population Survey (CPS) in the US, where the units follow the pattern 1111000000001111 (a unit is in the sample for subsequent 4 occasions, leaves it for subsequent 8 occasions, is again in the sample for subsequent 4 occasions and then never returns to the sample), or polish Labour Force Survey with the pattern 110011 (see Szarkowski and Witkowski (1994) or Popiński (2006)). Unfortunately the recurrent form of the BLUE in such situations of rotation patterns with holes is not known in general, see for instance Yansaneh and Fuller (1998). (Actually, the recurrent form of the optimal estimators for any rotation pattern with holes of size 1 has been derived only recently in Kowalski (2009) and for the Szarkowski scheme 110011 even more recently in Wesołowski (2010).) A widely accepted solution in the general situation is the K-composite estimator introduced in Hansen et al. (1955). Its optimality properties were studied for several models in Rao and Graham (1964) (shortened to RG in the rest of this paper).

By definition K-composite estimator makes use only of the most recent past composite estimator and observations from the present and the most recent past occasions. More precisely K-composite estimator on hth occasion, $\hat{\mu}_h$, has the following form

$$\hat{\mu}_h = Q\left(\hat{\mu}_{h-1} + \bar{X}^{(h)}_{h-1,h} - \bar{X}^{(h-1)}_{h-1,h}\right) + (1-Q)\bar{X}_h \ , \tag{1}$$

where $\hat{\mu}_{h-1}$ is the K-composite estimator on $(h-1)$th occasion, $\bar{X}^{(h)}_{h-1,h}$ is the sample mean for the units common to both $(h-1)$th and hth occasions calculated for the hth occasion, $\bar{X}^{(h-1)}_{h-1,h}$ is the sample mean of for the units common to both $(h-1)$th and hth occasions calculated for the $(h-1)$th occasion, $\bar{X}_h$ is the sample mean for all the units on hth occasion and $Q \in [0,1)$ is a numerical parameter which does not depend on h(!). Additionally in RG only a restricted though natural family of rotation patterns is investigated:

a group of units remains in the sample for r occasions, then leaves it for m occasions, comes back to the sample for r occasions, leaves it for m occasions, and so on. In such a setting strengthened by assuming exponential (Model 1) or arithmetic (Model 2) correlation pattern the optimal choice of Q is considered in that paper (in passing, let us note that Model 3 for correlation pattern is impossible since the resulting covariance matrix may not be positive definite). To attain this goal in RG it is taken $h \to \infty$, since otherwise, apparently, the optimal Q has to depend on h. Numerical solutions are then obtained since the resulting formula ((14) in RG) for the variance of the estimator is analytically non-treatable.

As it already has bee mentioned K-composite estimator has been used for years with some adjustments in the CPS - see for instance Bailar (1975), Breau and Ernst (1983) or Lent et al. (1994). A complete description can be found for instance in Current Population Survey (2002). The adjustments known as AK-composite estimator introduced in Gurney and Daly (1965) has been further developed, e.g. in Cantwell (1988) and Cantwell and Caldwell (1998). A more recent approach through regression composite estimator has been considered in Bell (2001), Fuller and Rao (2001), Singh et al. (2001) (with implications for Canadian Labour Force Survey). It is based on modified regression method proposed in Singh (1996). The difficulty in recursive estimation in repeated surveys for patterns with holes was raised in Yansaneh and Fuller (1998), who analyzed variances of composite estimators in several rotation schemes. For a relatively current description of the state of art in the area one can consult Steel and McLaren (2008), in particular Sec. IV on different rotation patterns and Sec. V on composite estimators. A very recent paper on optimal estimation under rotation is by Towhidi and Namazi-Rad (2010).

In the present paper we develop the idea of K-composite estimator in two new directions. First, $Q = Q_h$ is allowed to depend on the number of occasion. Then it appears that the optimal solution for Q_h is very simple: it is attained through minimizing certain quadratic function F_h (which has to be determined on each occasion). Second, any rotation pattern is allowed. The price for such a development is surprisingly cheap: we only have to keep track of subsequent Q_h's (to be able to determine F_h's).

2. Dynamic K-composite estimator

We consider a double array of random variables $(X_{i,j})$ which may be column-wise or row-wise infinite or finite, where the rows are for values of the variable of interest for different units on the same occasion, while columns are for values of the variable for the same unit on different occasions. Thus $X_{i,j}$ represents the value of the variable on ith occasion for the jth unit of the population. We assume that on a given occasion all the variables have the same mean, which is the parameter we want to estimate, i.e.

$$\mathbb{E}\, X_{i,j} = \mu_i\ , \quad i,j = 1,2,\ldots$$

Also it is assumed that there is no correlation between different units, i.e.

$$\mathbb{C}\mathrm{ov}(X_{i,j}, X_{l,k}) = 0 \quad \text{for } j \neq k, \ \ i,j,k,l = 1,2,\ldots$$

These two assumptions are crucial for further development of our result. The remaining two are not important for the derivation we propose but, first, make formulas somewhat simpler, second, since they include some parameters which are assumed to be known, it is desirable to have as few such parameters as possible. Thus, additionally we assume the exponential correlation pattern between the values of the variable for the same unit on different occasions (which, while not so important here, is a crucial condition for the Patterson scheme), i.e.

$$\mathbb{C}\mathrm{orr}(X_{i,j}, X_{i+k,j}) = \rho^k, \text{ for any } k = 0,1,\ldots,\ i,j = 1,2,\ldots,$$

for some $\rho \in [-1,1]$. Finally it is assumed that the variances of all variables are constant, i.e.

$$\mathbb{V}\mathrm{ar}\, X_{i,j} = \sigma^2 > 0\ , \quad i,j = 1,2,\ldots$$

The dynamic version of K-composite estimator, which called here DK-composite estimator, has the form

$$\hat{\mu}_h = Q_h \left(\hat{\mu}_{h-1} + \bar{X}^{(h)}_{h-1,h} - \bar{X}^{(h-1)}_{h-1,h}\right) + (1-Q_h)\bar{X}_h\ , \quad h = 2,3,\ldots \tag{2}$$

while $\hat{\mu}_1 = \bar{X}_1$, where all the symbols where introduced in (1) except of Q_h which plays the role of the former Q. Let us point out the we do not impose any a priori restrictions on the range of (Q_h) (restriction imposed in RG on the range of Q, $Q \in (0,1)$, made it possible to pass to the limit with $h \to \infty$ in the expression for the variance of $\hat{\mu}_h$). Our goal is to choose Q_h in a dynamic way, i.e. on each occasion $h \geq 2$, the value Q_h has to minimize the variance of $\hat{\mu}_h$.

The rotation scheme is described by the rotation matrix $R = (r_{i,j})$, where $r_{i,j} = 1$ if the jth unit is in the sample on the ith occasion, otherwise $r_{i,j} = 0$. There is absolutely no restriction on the rotation pattern. By n_i we denote the sample size on the ith occasion, and m_i denotes the size of overlap between samples on occasions $(i-1)$th and ith.

Denote also for $k = 2,3,\ldots$

$$D_{i,k} = \begin{cases} Q_i Q_{i+1} \cdot \ldots \cdot Q_k, & \text{for } i = 1, \ldots, k, \\ 1, & i = k+1 \,. \end{cases} \tag{3}$$

Then we define weights which will be responsible for the form of quadratic functions (F_k) to be minimized:

$$w_{i,j}^{(1)} = \frac{1}{n_1}$$

and for any $i = 1, \ldots, k > 1$ and any $j = 1, 2, \ldots$

$$w_{i,j}^{(k)} = r_{i,j} \left[D_{i,k} \left(\frac{r_{i-1,j}}{m_i} - \frac{1}{n_i} \right) + D_{i+1,k} \left(\frac{1}{n_i} - \frac{r_{i+1,j}}{m_{i+1}} \right) \right] , \tag{4}$$

where in the last expression we adopt the rule that $r_{k+1,j} = r_{0,j} = 0$. Note that with such a little abuse of notation the formula for $w_{i,j}^{(1)}$ agrees with (4). Let us emphasize that to find the weights $\left(w_{i,j}^{(k)}\right)$ for a given occasion k nothing more is needed but $k-1$ numbers $Q_2, \ldots, Q_k$ (note that $Q_1 = 0$, by the definition of $\hat{\mu}_1$).

Now we are ready to present our main result which explains how to choose, occasion by occasion, the values (Q_h) which make the estimator $\hat{\mu}_h$ optimal in the model we consider here. Though the formulas, in particular (6), do not look very friendly, it has to be emphasized that actually to find Q_h one needs just $Q_2, \ldots, Q_{h-1}$ to calculate $w_{i,j}^{(h-1)}$ and consequently, A_h, B_h and C_h.

Theorem 1. *In the model described above the optimal value of Q_h which minimizes the variance of DK-composite estimator $\hat{\mu}_h$, $h \geq 1$, is*

$$Q_h = \frac{C_h - B_h}{A_h - 2B_h + C_h} \tag{5}$$

with $C_1 = B_1$ and $A_h - 2B_h + C_h > 0$, where for $h \geq 2$

$$A_h = \sigma^2 \sum_j \left[\sum_{i=1}^{h-1} \left(w_{i,j}^{(h-1)} \right)^2 r_{i,j} + 2 \sum_{1 \leq i_1 < i_2 \leq h-1} w_{i_1,j}^{(h-1)} w_{i_2,j}^{(h-1)} r_{i_1,j} r_{i_2,j} \rho^{i_2 - i_1} \right]$$

$$+2 \frac{(1-\rho)\sigma^2}{m_h} \left[1 - \sum_j \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{h-1,j} r_{h,j} r_{i,j} \rho^{h-1-i} \right] , \tag{6}$$

$$B_h = \frac{\sigma^2}{n_h} \left[1 - \rho + \sum_j \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{i,j} r_{h,j} \rho^{h-i} \right] , \tag{7}$$

$$C_h = \frac{\sigma^2}{n_h}. \tag{8}$$

Moreover,

$$\hat{\mu}_h = \sum_{i=1}^{h} \sum_{j} w_{i,j}^{(h)} r_{i,j} X_{i,j}$$

with the weights $(w_{i,j}^{(h)})$ *defined in (4).*

Actually, as it will be observed during the proof, which is given in Section 3,

$$A_h - 2B_h + C_h = \mathbb{V}\mathrm{ar}\left(\hat{\mu}_{h-1} + \bar{X}_{h-1,h}^{(h)} - \bar{X}_{h-1,h}^{(h-1)} + \bar{X}_h\right)$$

which is always positive since $\sigma^2 > 0$.

3. Proof

The proof is by induction with respect to h. For $h = 1$ the result holds true since then $C_1 = B_1$ yields $Q_1 = 0$. Moreover, (4) for $k = 1$ agrees with the formula for $w_{i,j}^{(1)}$. We assume that it holds for $h - 1$ and we will prove it for h.

Compute the variance of $\hat{\mu}_h$:

$$\mathbb{V}\mathrm{ar}\,\hat{\mu}_h = Q_h^2 \Big[\mathbb{V}\mathrm{ar}\,\hat{\mu}_{h-1} + \mathbb{V}\mathrm{ar}\,\bar{X}_{h-1,h}^{(h)} + \mathbb{V}\mathrm{ar}\,\bar{X}_{h-1,h}^{(h-1)} + 2\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_{h-1,h}^{(h)}\right)$$

$$-2\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_{h-1,h}^{(h-1)}\right) - 2\mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h)}, \bar{X}_{h-1,h}^{(h-1)}\right)\Big]$$

$$+2Q_h(1-Q_h)\Big[\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_h\right) + \mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h)}, \bar{X}_h\right) - \mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h-1)}, \bar{X}_h\right)\Big] +$$

$$(1-Q_h)^2\,\mathbb{V}\mathrm{ar}\,\bar{X}_h = Q_h^2 A_h + 2Q_h(1-Q_h)B_h + (1-Q_h)^2 C_h,$$

where the last equality defines the quantities A_h, B_h and C_h.

By the induction assumption

$$\hat{\mu}_{h-1} = \sum_{i=1}^{h-1} \sum_{j} w_{i,j}^{(h-1)} r_{i,j} X_{i,j}\ .$$

Then a direct computation gives

$$\mathbb{V}\mathrm{ar}\,\hat{\mu}_{h-1} = \sigma^2 \sum_j \left[\sum_{i=1}^{h-1} \left(w_{i,j}^{(h-1)} \right)^2 r_{i,j} + 2 \sum_{1 \le i_1 < i_2 \le h-1} w_{i_1,j}^{(h-1)} w_{i_2,j}^{(h-1)} r_{i_1,j} r_{i_2,j} \rho^{i_2 - i_1} \right] ,$$

$$\mathbb{V}\mathrm{ar}\,\bar{X}_{h-1,h}^{(h)} = \mathbb{V}\mathrm{ar}\,\bar{X}_{h-1,h}^{(h-1)} = \frac{\sigma^2}{m_h} ,$$

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_{h-1,h}^{(h)} \right) = \frac{\sigma^2}{m_h} \sum_j \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{i,j} r_{h-1,j} r_{h,j} \rho^{h-i} ,$$

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_{h-1,h}^{(h-1)} \right) = \frac{\sigma^2}{m_h} \sum_j \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{i,j} r_{h-1,j} r_{h,j} \rho^{h-1-i} ,$$

$$\mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h)}, \bar{X}_{h-1,h}^{(h-1)} \right) = \frac{\sigma^2 \rho}{m_h} .$$

Combining the last five formulas we observe that the definition of A_h agrees with the expression (6).

Similarly to check if (7) holds we have to compute

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_{h-1}, \bar{X}_h \right) = \frac{\sigma^2}{n_h} \sum_j \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{i,j} r_{h,j} \rho^{h-i} ,$$

$$\mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h)}, \bar{X}_h \right) = \frac{\sigma^2}{n_h} ,$$

$$\mathbb{C}\mathrm{ov}\left(\bar{X}_{h-1,h}^{(h-1)}, \bar{X}_h \right) = \frac{\sigma^2 \rho}{n_h} .$$

Finally, (8) follows since

$$\mathbb{V}\mathrm{ar}\,\bar{X}_h = \frac{\sigma^2}{n_h} .$$

Minimizing

$$F_h(x) = (A_h - 2B_h + C_h)x^2 + 2(B_h - C_h)x + C_h$$

we get the solution (5).

The DK-composite estimator is a linear estimator so in general it has a form

$$\hat{\mu}_h = \sum_{i=1}^{h} \sum_{j} v_{i,j} r_{i,j} X_{i,j}$$

with some weights $(v_{i,j})$. To finish the proof we have to show that $v_{i,j} r_{i,j} = w_{i,j}^{(h)} r_{i,j}$ as defined in (4) for any $i = 1, \ldots, h$ and any $j = 1, 2, \ldots$.

Note that by the definition of $\hat{\mu}_h$ given in (2) we have

$$\hat{\mu}_h = Q_h \left(\sum_{j} \sum_{i=1}^{h-1} w_{i,j}^{(h-1)} r_{i,j} X_{i,j} + \frac{1}{m_h} \sum_{j} r_{h-1,j} r_{h,j} X_{h,j} \right.$$

$$\left. - \frac{1}{m_h} \sum_{j} r_{h-1,j} r_{h,j} X_{h-1,j} \right) + (1 - Q_h) \frac{1}{n_h} \sum_{j} r_{h,j} X_{h,j}.$$

Comparing the coefficients of $X_{i,j}$ in the last two expressions we get for $i = h$

$$\begin{aligned} v_{h,j} r_{h,j} &= \frac{Q_h}{m_h} r_{h-1,j} r_{h,j} + \frac{1 - Q_h}{n_h} r_{h,j} \\ &= r_{h,j} \left[D_{h,h} \left(\frac{r_{h-1,j}}{m_h} - \frac{1}{n_h} \right) + \frac{1}{n_h} \right] = w_{h,j}^{(h)} r_{h,j} \ , \end{aligned}$$

for $i = h - 1$

$$v_{h-1,j} r_{h-1,j} = Q_h \left(w_{h-1,j}^{(h-1)} r_{h-1,j} - \frac{1}{m_h} r_{h-1,j} r_{h_j} \right)$$

$$= Q_h r_{h-1,j} \left[D_{h-1,h-1} \left(\frac{r_{h-2,j}}{m_{h-1}} - \frac{1}{n_{h-1}} \right) + \frac{1}{n_{h-1}} \right] - Q_h \frac{1}{m_h} r_{h-1,j} r_{h,j}$$

$$= r_{h-1,j} \left[D_{h-1,h} \left(\frac{r_{h-2,j}}{m_{h-1}} - \frac{1}{n_{h-1}} \right) + D_{h,h} \left(\frac{1}{n_{h-1}} - \frac{r_{h,j}}{m_h} \right) \right] = w_{h-1,j}^{(h)} r_{h-1,j} \ ,$$

and for any $i < h - 1$

$$v_{i,j} r_{i,j}$$

$$= Q_h w_{i,j}^{(h-1)} r_{i,j} \left[D_{i,h-1} \left(\frac{r_{i-1,j}}{m_i} - \frac{1}{n_i} \right) + D_{i+1,h-1} \left(\frac{1}{n_i} - \frac{r_{i+1,j}}{m_{i+1}} \right) \right] = w_{i,j}^{(h)} r_{i,j}$$

since $Q_h D_{k,h-1} = D_{k,h}$ for $k = i, \ i + 1$.

Thus the proof is completed. $\square$

4. Numerical examples

Below, similarly to numerical comparisons in RG, we consider percentage gain in efficiency for the DK-composite estimator compared to the mean of the observations from the last hth occasion. It is defined as

$$g_h = \frac{\mathbb{V}\mathrm{ar}\,\bar{X}_h - \mathbb{V}\mathrm{ar}\,\hat{\mu}_h}{\mathbb{V}\mathrm{ar}\,\hat{\mu}_h} \times 100. \tag{9}$$

We took $h = 20$, since the parameter Q_h behaves quite stable with respect to occasion number h. In Tables 1 and 2 we give the optimal weight Q_h, the variance of μ_h and g_h for different values of correlation ρ in two schemes: Szarkowski's 110011 (Table 1) and CPS 1111000000001111 (Table 2). We can easily see that the largest gain is achieved for strong correlations and the smallest when there is no correlation between occasions for the same unit.

Table 1. Szarkowski scheme

ρ	Q_{20}	$\mathbb{V}\mathrm{ar}\,\hat{\mu}_{20}$	g_{20}
-0.9	-0.16	0.235	6.275
-0.8	-0.14	0.238	5.173
-0.7	-0.13	0.240	4.138
-0.6	-0.12	0.242	3.181
-0.5	-0.10	0.244	2.317
-0.4	-0.08	0.246	1.560
-0.3	-0.07	0.248	0.926
-0.2	-0.05	0.249	0.437
-0.1	-0.02	0.250	0.116
0	0.00	0.250	0.000
0.1	0.03	0.250	0.135
0.2	0.06	0.249	0.592
0.3	0.09	0.246	1.473
0.4	0.13	0.243	2.940
0.5	0.17	0.238	5.256
0.6	0.22	0.230	8.877
0.7	0.29	0.218	14.695
0.8	0.38	0.200	24.797
0.9	0.51	0.171	46.499

Table 2. CPS scheme

ρ	Q_{20}	$\mathbb{V}\mathrm{ar}\,\hat{\mu}_{20}$	g_{20}
-0.9	-0.3	0.116	7.469
-0.8	-0.27	0.118	5.772
-0.7	-0.24	0.120	4.341
-0.6	-0.20	0.121	3.149
-0.5	-0.17	0.122	2.172
-0.4	-0.14	0.123	1.390
-0.3	-0.11	0.124	0.787
-0.2	-0.07	0.125	0.355
-0.1	-0.04	0.125	0.091
0	0.00	0.00	0.000
0.1	0.04	0.250	0.098
0.2	0.08	0.249	0.414
0.3	0.12	0.124	1.000
0.4	0.17	0.123	1.947
0.5	0.23	0.121	3.417
0.6	0.29	0.118	5.729
0.7	0.36	0.114	9.586
0.8	0.45	0.107	16.908
0.9	0.58	0.092	35.564

Source: own calculations

Consider now a cascade rotation scheme which is defined through a rotation pattern $(1, \epsilon_2, \ldots, \epsilon_{k-1}, 1)$, $\epsilon_l \in \{0, 1\}$, $l = 2, \ldots, k-1$, which moves one unit down the rotation matrix with subsequent occasions, that is

$$(r_{i,i}, \ldots, r_{i,i+k}) = (1, \epsilon_2, \ldots, \epsilon_{k-1}, 1)$$

for any $i = 1, 2, \ldots$, otherwise $r_{i,j} = 0$. The number k is called the rotation pattern length.

Taking the advantage of the fact that the DK-composite estimator allows for any rotation scheme, we calculated percentage gain (as defined in (9)) in efficiency for all possible cascade schemes with rotation patterns of length up to 10. Table 3 contains 10 schemes with the smallest and 10 with the largest gain among such $2^8 = 256$ schemes. Here, again, the results for $h = 20$ are presented. The largest gain is achieved for "sparse" schemes with small number of elements in the rotation pattern (and strong correlations) while the lowest gain is observed for schemes with complete or almost complete rotation patterns (and for weak correlations). Similar comparisons of variances for particular rotation cascade patterns in the time series framework can be found in McLaren and Steel (2000) (see also Steel and McLaren (2002)).

Table 3. The worst and the best rotation patterns

worst patterns	ρ	g_{20}	best patterns	ρ	g_{20}
1111111111	-0.1	0.045	1000010001	0.9	84.087
1111111111	0.1	0.046	1000100001	0.9	84.087
111111111	-0.1	0.049	101	0.9	93.570
111111111	0.1	0.051	1001	0.9	108.519
11111111	-0.1	0.054	10001	0.9	117.448
11111111	0.1	0.056	100001	0.9	122.970
1111111	-0.1	0.060	1000001	0.9	126.075
1111111	0.1	0.063	10000001	0.9	127.970
1101010101	-0.1	0.064	1000000001	0.9	129.264
1011010101	-0.1	0.064	100000001	0.9	129.305

Source: own calculations

Table 4. Comparison between K-composite and DK-composite estimators

ρ	0.5			0.6			0.7		
	Q_{20}	g_{20}	diff	Q_{20}	g_{20}	diff	Q_{20}	g_{20}	diff
m	$r = 2$								
2	0.23	3.42	1.74	0.29	5.73	2.77	0.36	9.59	4.64
4	0.17	5.26	0.01	0.22	8.88	-0.19	0.29	14.70	0.58
8	0.17	5.29	-0.03	0.23	9.02	-0.33	0.29	15.23	0.09
∞	0.17	5.29	-0.03	0.23	9.02	-0.33	0.29	15.23	0.09
m	$r = 3$								
3	0.24	2.34	1.83	0.30	3.97	3.28	0.38	6.56	5.41
6	0.21	4.27	-0.04	0.27	7.20	0.13	0.34	12.03	0.40
9	0.21	4.27	-0.04	0.27	7.23	0.10	0.34	12.16	0.28
∞	0.21	4.27	-0.04	0.27	7.23	0.10	0.34	12.16	0.28
m	$r = 4$								
4	0.25	1.84	1.47	0.32	3.02	2.76	0.39	4.96	4.82
8	0.23	3.42	-0.11	0.29	5.73	0.06	0.36	9.56	0.32
12	0.23	3.42	-0.11	0.29	5.73	0.06	0.36	9.59	0.29
∞	0.23	3.42	-0.11	0.29	5.73	0.06	0.36	9.59	0.29
m	$r = 6$								
6	0.25	1.25	1.08	0.32	2.04	1.92	0.39	3.32	3.38
12	0.24	2.40	-0.07	0.30	3.97	-0.01	0.38	6.56	0.14
18	0.24	2.40	-0.07	0.30	3.97	-0.01	0.38	6.56	0.14
∞	0.24	2.40	-0.07	0.30	3.97	-0.01	0.38	6.56	0.14
m	$r = 8$								
8	0.26	0.94	0.86	0.32	1.54	1.45	0.40	2.50	2.53
16	0.25	1.84	-0.03	0.31	3.02	-0.03	0.39	4.96	0.07
∞	0.25	1.84	-0.03	0.31	3.02	-0.03	0.39	4.96	0.07

Source: own calculations

Table 4. Comparison between K-composite and DK-composite estimators, continuation

ρ	0.8			0.9		
	Q_{20}	g_{20}	diff	Q_{20}	g_{20}	diff
m	$r = 2$					
2	0.45	16.91	5.94	0.58	35.72	3.06
4	0.38	24.80	1.88	0.51	46.50	4.15
8	0.38	26.75	0.37	0.52	54.65	1.01
∞	0.38	26.77	0.35	0.52	55.07	1.18
m	$r = 3$					
3	0.47	11.46	8.74	0.60	24.23	12.42
6	0.43	20.77	1.53	0.56	40.48	4.80
9	0.43	21.44	1.00	0.57	44.28	2.43
∞	0.43	21.47	0.98	0.57	44.89	2.09
m	$r = 4$					
4	0.48	8.57	8.55	0.61	17.93	14.47
8	0.45	16.69	1.10	0.58	33.67	4.92
12	0.45	16.91	0.89	0.58	35.56	4.23
∞	0.45	16.91	0.89	0.58	35.72	4.30
m	$r = 6$					
6	0.49	5.67	6.31	0.61	11.61	14.41
12	0.47	11.44	0.78	0.60	23.79	4.56
18	0.47	11.46	0.76	0.60	24.22	4.28
∞	0.47	11.46	0.76	0.60	24.23	4.28
m	$r = 8$					
8	0.49	4.23	5.00	0.62	8.55	12.08
16	0.48	8.57	0.70	0.61	17.83	3.50
∞	0.48	8.57	0.70	0.61	17.93	3.42

Source: own calculations

Table 4 shows Q_h, g_h and the difference ($diff = g - g_h$) between the gains obtained in two ways: g for the K-composite estimator as computed in Table 1 of RG and g_h for the DK-composite estimator as proposed in the present paper. We took $h = 100$ though in many particular cases the values for Q_h and g_h stabilized much earlier. The numbers r and m are responsible for the rotation pattern, i. e. a unit stays in the sample for r occasions, leaves the sample for m occasions, comes back into the sample for r occasions, and so on. Table 4 is named "comparison" nevertheless we cannot actually in strictly mathematical sense compare these two values of g and g_h because the two methods involve different models: RG considered a finite population case in which a given unit returns to the survey infinitely often whereas in the present paper an infinite population model is investigated and a unit returns to the survey after a gap of m occasions for another sequence of r occasions and then leaves the survey. In the course of simulations we noted that Q_h for the DK-composite estimator are quite stable, even for relatively small values of h. Moreover, their values observed in simulations were quite similar to those of Q, obtained in Table 1 of RG. In our Table 4 it is visible that differences in gains of efficiency are remarkable for strong correlations and small gaps m, while for small correlations and large gaps m they are insignificant. Small negative values which appear in some cases are due to the fact that the two methods are not precisely equivalent, otherwise negative values would not be possible since the method we present here is optimal within considered class of estimators.

REFERENCES

BAILAR, B. (1975). The effects of rotation group bias on estimates from panel surveys. J. Amer. Statist. Assoc. 70, 23-30.

BELL, P. (2001). Comparison of alternative Labour Force Survey estimators. Survey Meth. 27(1), 53-63.

BREAU, P., ERNST, L. (1983). Alternative estimators to the current composite estimator. Proc. Sec. Survey Res. Meth., Amer. Statist. Assoc., 397-402.

CANTWELL, P.J. (1988). Variance formulae for the generalized composite estimator under balanced one-level rotation plan. SRD Research Report Census/SRD/88/26, Bureau of the Census, Statistical Research Division, 1-16.

CANTWELL, P.J., CALDWELL, C.V. (1998). Examining the revisions in monthly retail and wholesale trade surveys under a rotation panel design. J. Offic. Statist. 14, 47-54.

Current Population Survey (2002). Design and Methodology, Technical Paper 63RV, Bureau of Labour Statistics, U.S. Census Bureau.

FULLER, W., RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. Survey Meth. 27(1), 45-51.

GURNEY, M., DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. Proc. Amer. Statist. Assoc., Sect. Soc. Statist., 242-257.

HANSEN, M.H., HURWITZ, W.N., NISSELSON, H., STEINBERG, J. (1955). The redesign of the census current population survey. J. Amer. Math. Assoc. 50, 701-719.

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

LENT, J., MILLER, S., CANTWELL, P. (1994). Composite weights for the Current Population Survey. Proc. Sec. Survey Res. Meth., Amer. Statist. Assoc., 867-872.

MCLAREN, C.H., STEEL, D.G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. Survey Meth. 26(2), 163-172.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. J. Royal Statist. Soc., Ser. B 12, 241-255.

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

RAO, J.N.K., GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. Ann. Math. Statist. 35, 492-509.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc., 120-129.

SINGH, A.C., KENNEDY, B., WU, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. Survey Meth. 27, 33-44.

STEEL, D., MCLAREN, C. (2002), In search of a good rotation pattern. In: Advances in Statistics, Combinatorics and Related Areas. Singapore, World Scientific, 309-319.

STEEL, D., MCLAREN, C. (2008). Design and analysis of repeated surveys. Centre for Statist. Survey Meth., Univ. Wollonong, Working Paper 11-08, 1-13, http://ro.uow.edu.au/cssmwp/10

SZARKOWSKI, A., WITKOWSKI, J. (1994), The Polish labour force survey. Statist. Transit. 1(4), 467-483.

TOWHIDI, M., NAMAZI-RAD, M.-R. (2010). An optimal method of estimation in rotation sampling. Adv. Appl. Statist. 15(2) , 115-136.

WESOŁOWSKI, J. (2010). Recursive optimal estimation in Szarkowski rotation scheme. Statist. Transit. 11(2), 267-285.

YANSANEH, I.S., FULLER, W. (1998). Optimal recursive estimation for repeated surveys. Survey Meth. 24, 31-40.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 21—36

ESTIMATION OF POPULATION MEAN USING TWO AUXILIARY SOURCES IN SAMPLE SURVEYS

Diwakar Shukla[1], Sharad Pathak[1] and Narendra Singh Thakur[2]

ABSTRACT

This paper proposes families for estimation of population mean of the main variable under study using the information on two different auxiliary variables under simple random sampling without replacement (SRSWOR) scheme. Three different classes of estimators are constructed, examined with a complete study with other existing estimators. The expression for bias and mean squared error of the proposed families are obtained up to first order of approximation. Usual ratio estimator, product estimator, dual to ratio estimator, ratio-cum-product type estimator and many more estimators are identified as particular members of the suggested family. Expressions of optimization are derived and theoretical results are supported by numerical examples.

Key words: Family of estimators, SRSWOR, Bias and Mean squared error. **AMS Subject Classification:** 94A20, 62D05

1. Introduction

To improve the exactitude in sample surveys theory the use of two auxiliary variables for estimation of population mean of a variable under study has played an influential role. A number of estimators are accessible in the literature of sample surveys where supporting information is the contributor to improve the methodology. Out of all ratio and product estimators are good examples as evidence to state this. The ratio estimation method is practical when the correlation coefficient between the study and auxiliary variable is positive [Cochran (1940, 42)]. If the correlation coefficient between the study and auxiliary variable is negative then the use of product estimation will make the study valuable [Robson (1957) and Murthy (1964)].

[1] Department of Mathematics and Statistics. Dr. Hari Singh Gour Central University, Sagar, M.P., India - 470003. E-mail: sharadpathakstats@yahoo.com, diwakarshukla@rediffmail.com.

2 Center for Mathematical Sciences, Banasthali University, Rajasthan, India.

There are so many situations in survey sampling where the record of more auxiliary variable is available for the investigators (at least for two variables). There are so many researchers who used the information of more than two auxiliary variables to contribute in the field. Dalabehara and Sahoo (1994) presented a class of estimators in stratified sampling with two auxiliary variables for estimation of mean. In another contribution Dalabehara and Sahoo (2000) proposed an unbiased estimator in two-phase sampling using two auxiliary variables.

Abu-Dayyeh et al. (2003) used auxiliary variables to show estimators of finite population mean. Sahoo and Sahoo (1993) suggested a class of estimators in two-phase sampling using two auxiliary variables. In another work Sahoo and Sahoo (2001) discussed about predictive estimation of finite population mean in two-phase sampling using two auxiliary variables. Singh and Shukla (1987) have a discussion on one parameter family of factor type ratio estimator. In a study Shukla et al. (1991) transformed factor type estimator to make the estimation more effective. Shukla (2002) Studied F-T estimator and sampling procedure undertaken was two-phase sampling. In this sequence Singh and Singh (1991) provided Chain type estimator with two auxiliary variables under double sampling scheme. In another study Singh et al. (1994) suggested a class of chain-ratio estimator with two auxiliary variables and the study completed under double sampling scheme. Kadilar and Cingi (2004) took two auxiliary variables in simple random sampling to find population mean. Moreover, Kadilar and Cingi (2005) derived a new estimator using two auxiliary variables. Perri (2007) analysed the work of Singh (1965, 1967b) and suggested a new improved work on ratio-cum-product type estimators with the application of Srivenkataramana, T. (1980) estimator on pervious proposed work of Singh (1967b).

Many authors including Srivastava (1971), Srivastava and Jhajj (1983), Ray and Sahai (1980), Khare and Srivastava (1981), Hansen et al.(1953) and Desraj (1965) used more than one supporting information to make the study more impressive. Some other useful contributions over applications of auxiliary information are due to Mukhopadhyay (2000), Cochran (2005), Murthy (1976), Sukhatme et al. (1984), Naik and Gupta (1991), Singh and Shukla (1993) and Shukla et al (2009) etc.

2. Notations and Assumptions

Notations for the study are:

$\bar{Y}, \bar{X}_1, and\ \bar{X}_2$: Population Parameters

$\bar{y}, \bar{x}_1\ and\ \bar{x}_2$: Mean per unit estimates for a simple random sample of size n.

n : Sample size

f : Sampling friction ($f = n/N$)

N : Population size

ρ_{01} : Correlation between variable *Y* and *X₁*

ρ_{02} : Correlation between variable *Y* and *X₂*

ρ_{12} : Correlation between variable *X₁* and *X₂*

$C_Y = {S_Y}/{\overline{Y}}$: Coefficient of variation for variable Y (C_0)

$C_{X_1} = {S_{X_1}}/{\overline{X}_1}$: Coefficient of variation for variable $X_1 (C_1)$

$C_{X_2} = {S_{X_2}}/{\overline{X}_2}$: Coefficient of variation for variable $X_2 (C_2)$

3. Some Estimators

In the literature of survey sampling so many estimators and estimation procedures exist. This literature is the basic motivation to work in this direction and contribution in this area. Let Y is the main variable and X_1, X_2 are two auxiliary variable then some well known estimators are as follows.

3.1. Ratio estimator

$$\overline{y}_R = \overline{y}\left(\frac{\overline{X}}{\overline{x}}\right) \tag{3.1a}$$

$$Bias(\overline{y}_R) = E(\overline{y}_R - \overline{Y}) = \overline{Y}M_1\left[C_X^2 - \rho C_Y C_X\right] \tag{3.1b}$$

$$MSE(\overline{y}_R) = \overline{Y}^2 M_1\left[C_Y^2 + C_X^2 - 2\rho C_Y C_X\right] \; ; \; M_1 = \left(\frac{1}{n} - \frac{1}{N}\right) \tag{3.1c}$$

3.2. Product estimator

$$\overline{y}_P = \overline{y}\left(\frac{\overline{x}}{\overline{X}}\right) \tag{3.2a}$$

$$Bias(\overline{y}_P) = E(\overline{y}_P - \overline{Y}) = \overline{Y}M_1 \rho C_Y C_X \tag{3.2b}$$

$$MSE(\overline{y}_R) = \overline{Y}^2 M_1\left[C_Y^2 + C_X^2 + 2\rho C_Y C_X\right] \; ; \; M_1 = \left(\frac{1}{n} - \frac{1}{N}\right) \tag{3.2c}$$

3.3. Dual to ratio estimator [By Srivenkataramana, T. (1980)]

$$\overline{y}_{VR} = \overline{y}\,\frac{N\overline{X} - n\overline{x}}{(N-n)\overline{X}} \tag{3.3a}$$

$$Bias(\bar{y}_{VR}) = E(\bar{y}_{VR} - \bar{Y}) = -\frac{\bar{Y}}{N}\rho C_Y C_X \qquad (3.3b)$$

$$MSE(\bar{y}_{VR}) = \bar{Y}^2 M_1\left[C_Y^2 + \alpha^2 C_X^2 - 2\alpha\rho\, C_Y C_X\right] \;; M_1 = \left(\frac{1}{n} - \frac{1}{N}\right), \alpha = n/(N-n)$$

… (3.3c)

3.4. Ratio-cum-product type estimator

Singh (1965, 1967b) proposed some ratio-cum-product type estimators as

$$\bar{y}_{R1} = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{x}_2}{\bar{X}_2} \qquad (3.4a)$$

$$Bias(\bar{y}_{R1}) = E(\bar{y}_R - \bar{Y}) = \bar{Y}M_1\left[C_1^2 + \rho_{02}\, C_0 C_2 - \rho_{01}\, C_0 C_1 - \rho_{12}\, C_1 C_2\right] \qquad (3.4b)$$

$$MSE(\bar{y}_{R1}) = \bar{Y}^2 M_1\left[C_0^2 + C_1^2 + C_2^2 - 2\rho_{01}\, C_0 C_1 + 2\rho_{02}\, C_0 C_2 - 2\rho_{12}\, C_1 C_2\right]$$

(3.4c)

$$\bar{y}_{R2} = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{X}_2}{\bar{x}_2} \qquad (3.5a)$$

$$Bias(\bar{y}_{R2}) = E(\bar{y}_{R2} - \bar{Y}) = \bar{Y}M_1\left[C_1^2 + C_2^2 - \rho_{02}\, C_0 C_2 - \rho_{01}\, C_0 C_1 + \rho_{12}\, C_1 C_2\right]$$

(3.5b)

$$MSE(\bar{y}_{R2}) = \bar{Y}^2 M_1\left[C_0^2 + C_1^2 + C_2^2 - 2\rho_{01}\, C_0 C_1 - 2\rho_{02}\, C_0 C_2 + 2\rho_{12}\, C_1 C_2\right]$$

(3.5c)

$$\bar{y}_{P1} = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{x}_2}{\bar{X}_2} \qquad (3.6a)$$

$$Bias(\bar{y}_{P1}) = E(\bar{y}_{P1} - \bar{Y}) = \bar{Y}M_1\left[\rho_{01}\, C_0 C_1 + \rho_{02}\, C_0 C_2 + \rho_{12}\, C_1 C_2\right] \qquad (3.6b)$$

$$MSE(\bar{y}_{P1}) = \bar{Y}^2 M_1\left[C_0^2 + C_1^2 + C_2^2 + 2\rho_{01}\, C_0 C_1 + 2\rho_{02}\, C_0 C_2 + 2\rho_{12}\, C_1 C_2\right]$$

(3.6c)

$$\bar{y}_{P2} = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{X}_2}{\bar{x}_2} \qquad (3.7a)$$

$$Bias(\bar{y}_{P2}) = E(\bar{y}_{P2} - \bar{Y}) = \bar{Y}M_1\left[C_2^2 + \rho_{01}\, C_0 C_1 - \rho_{02}\, C_0 C_2 - \rho_{12}\, C_1 C_2\right]$$

(3.7b)

$$MSE(\bar{y}_{P2}) = \bar{Y}^2 M_1\left[C_0^2 + C_1^2 + C_2^2 + 2\rho_{01}\, C_0 C_1 - 2\rho_{02}\, C_0 C_2 - 2\rho_{12}\, C_1 C_2\right]$$

(3.7c)

where $M_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$

4. Proposed Estimator(s)

Singh and Shukla (1987) discussed a family of factor-type *(F-T)* ratio estimator for estimating population mean. In another contribution Singh and Shukla (1993) derived efficient factor-type estimator for estimating the same population parameter. Deriving motivation from both some proposed estimators are given below.

$$\left.\begin{aligned} (\bar{y}_{F-T})_1 &= \bar{y}T_1T_2 \\ (\bar{y}_{F-T})_2 &= \bar{y}\frac{T_1}{T_2} \\ (\bar{y}_{F-T})_3 &= \bar{y}\frac{T_2}{T_1} \end{aligned}\right\} \tag{4.1}$$

Where $T_i = \dfrac{(A_i + C_i)\bar{X}_i + fB\,\bar{x}_i}{(A_i + fB_i)\bar{X}_i + C_i\bar{x}_i}$ (4.2)

$$A_i = (K_i - 1)(K_i - 2); B_i = (K_i - 1)(K_i - 4); C_i = (K_i - 2)(K_i - 3)(K_i - 4)$$ … (4.3)

Remark 4.1 Here we have a combination of K_i where $i = (1,2)$. Some of the factors are shown in the following table where $(K_1 = K_2)$. As above concerned K_i where $i = (1,2)$ is constant to choose suitably so that the resulting mean squared error of proposed estimators may become least. For example let $K_i = 1$ then the values of T_1 and T_2 will be $\dfrac{\bar{X}_1}{\bar{x}_1}$ and $\dfrac{\bar{X}_2}{\bar{x}_2}$ respectively and so on.

Remark 4.2 By proposed estimator we can obtain so many different estimators. For each combination of (K_1, K_2) an estimator exists.

Table 4.1. Some Members of the proposed estimation.

$t_1 = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{X}_2}{\bar{x}_2}$ (At $K_1 = K_2 = 1$)	$t_2 = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{x}_2}{\bar{X}_2}$ (At $K_1 = 1, K_2 = 2$)	$t_3 = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$ (At $K_1 = 1, K_2 = 3$)	$t_4 = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}$ (At $K_1 = 1, K_2 = 4$)
$t_5 = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{X}_2}{\bar{x}_2}$ (At $K_1 = 2, K_2 = 1$	$t_6 = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{x}_2}{\bar{X}_2}$ (At $K_1 = K_2 = 2$)	$t_7 = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$ (At $K_1 = 2, K_2 = 3$)	$t_8 = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}$ (At $K_1 = 2, K_2 = 4$)

Table 4.1. Some Members of the proposed estimation (cont.).

t_9 $= \bar{y}\dfrac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1}\dfrac{\bar{X}_2}{\bar{x}_2}$ (At $K_1 = 3, K_2 = 1$)	t_{10} $= \bar{y}\dfrac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1}\dfrac{\bar{x}_2}{\bar{X}_2}$ At $K_1 = 3, K_2 = 2$	t_{11} $= \bar{y}\dfrac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1}\dfrac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$ (At $K_1 = K_2 = 3$)	t_{12} $= \bar{y}\dfrac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1}$ (At $K_1 = 3, K_2 = 4$)
$t_{13} = \bar{y}\dfrac{\bar{X}_2}{\bar{x}_2}$ (At $K_1 = 4, K_2 = 1$)	$t_{14} = \bar{y}\dfrac{\bar{x}_2}{\bar{X}_2}$ (At $K_1 = 4, K_2 = 2$)	$t_{15} = \bar{y}\dfrac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$ (At $K_1 = 4, K_2 = 3$)	$\bar{y}$ (At $K_1 = K_2 = 4$)

5. Properties of Proposed Estimator

For large sample approximation we assume that

$$\bar{y} = \bar{Y}(1 + e_0);\ \bar{x}_1 = \bar{X}_1(1 + e_1)\ ;\ \bar{x}_2 = \bar{X}_2(1 + e_2)\,;\alpha_i = \frac{fB_i}{A_i + fB_i + C_i}\ ;$$

$$\beta_i = \frac{C_i}{A_i + fB_i + C_i}$$

$$E(e_0) = E(e_1) = E(e_2) = 0\,;\ E(e_0^2) = M_1C_0^2\ ;\ E(e_1^2) = M_1C_1^2\,;$$

$$E(e_2^2) = M_1C_2^2\,;\delta_{1i} = \alpha_i - \beta_i$$

$$E(e_0e_1) = M_1\rho_{01}C_0C_1\,; E(e_0e_2) = M_1\rho_{02}C_0C_2\,; E(e_1e_2) = M_1\rho_{12}C_1C_2\,;$$

$$M_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$$

THEOREM 5.1:

[1]: The estimator $(\bar{y}_{F-T})_1$ in terms of e_0, e_1 and e_2 up to first order of approximation could be expressed as:

$$(\bar{y}_{F-T})_1 = \bar{Y}\left[1 + e_0 + \delta_1(e_1 + e_0e_1 - \beta_1e_1{}^2) + \delta_2(e_2 + e_0e_2 - \beta_2e_2{}^2) + \delta_1\delta_2e_1e_2\right] \tag{5.1}$$

[2]: Bias of $(\bar{y}_{F-T})_1$ up to first order approximation is:

$$B(\bar{y}_{F-T})_1 = \bar{Y}M_1\left[\delta_1(\rho_{01}C_0C_1 - \beta_1C_1{}^2) + \delta_2(\rho_{02}C_0C_2 - \beta_2C_2{}^2) + \delta_1\delta_2\rho_{12}C_1C_2\right] \tag{5.2}$$

[3]: Mean squared error of $(\bar{y}_{F-T})_1$ up to first order approximation is:

$$M(\bar{y}_{F-T})_1 = \bar{Y}^2M_1\left[C_0{}^2 + \delta_1{}^2C_1{}^2 + \delta_2{}^2C_2{}^2 + 2\delta_1\rho_{01}C_0C_1 + 2\delta_2\rho_{02}C_0C_2 + 2\delta_1\delta_2\rho_{12}C_1C_2\right]$$

… (5.3)

Proof 5.1:

[1]:

$$(\bar{y}_{F-T})_1 = \bar{y}\,\frac{(A_1+C_1)\bar{X}_1 + fB_1\,\bar{x}_1}{(A_1+fB_1)\bar{X}_1 + C_1\bar{x}_1}\,\frac{(A_2+C_2)\bar{X}_2 + fB_2\,\bar{x}_2}{(A_2+fB_2)\bar{X}_2 + C_2\bar{x}_2}$$

$$(\bar{y}_{F-T})_1 = \bar{Y}\,(1+e_0)(1+\alpha_1 e_1)(1+\alpha_2 e_2)(1+\beta_1 e_1)^{-1}(1+\beta_2 e_2)^{-1}$$

$$(\bar{y}_{F-T})_1 = \bar{Y}\left[1 + e_0 + \delta_1(e_1 + e_0 e_1 - \beta_1 {e_1}^2) + \delta_2(e_2 + e_0 e_2 - \beta_2 {e_2}^2) + \delta_1\delta_2 e_1 e_2\right]$$

[2]:

$$E\left[(\bar{y}_{F-T})_1 - \bar{Y}\right] = E\left[\bar{Y}\,\{e_0 + \delta_1(e_1 + e_0 e_1 - \beta_1 {e_1}^2) + \delta_2(e_2 + e_0 e_2 - \beta_2 {e_2}^2) + \delta_1\delta_2 e_1 e_2\}\right]$$

$$B(\bar{y}_{F-T})_1 = \bar{Y}M_1\left[\delta_1\,(\rho_{01}C_0C_1 - \beta_1 {C_1}^2) + \delta_2\,(\rho_{02}C_0C_2 - \beta_2 {C_2}^2) + \delta_1\delta_2\rho_{12}C_1C_2\right]$$

[3]:

$$\left[(\bar{y}_{F-T})_1 - \bar{Y}\right]^2 = \bar{Y}\,[e_0 + \delta_1(e_1 + e_0 e_1 - \beta_1 {e_1}^2) + \delta_2(e_2 + e_0 e_2 - \beta_2 {e_2}^2) + \delta_1\delta_2 e_1 e_2]^2$$

$$M(\bar{y}_{F-T})_1 = \bar{Y}^2 M_1\,[{C_0}^2 + {\delta_1}^2{C_1}^2 + {\delta_2}^2{C_2}^2 + 2\delta_1\rho_{01}C_0C_1 + 2\delta_2\rho_{02}C_0C_2 + 2\delta_1\delta_2\rho_{12}C_1C_2]$$

THEOREM 5.2:

[4]: The estimator $(\bar{y}_{F-T})_2$ in terms of e_0, e_1 and e_2 up to first order of approximation could be expressed as:

$$(\bar{y}_{F-T})_2 = \bar{Y}\left[1 + e_0 + \delta_1(e_1 + e_0 e_1 - \beta_1 {e_1}^2) - \delta_2(e_2 + e_0 e_2 - \alpha_2 {e_2}^2) - \delta_1\delta_2 e_1 e_2\right] \tag{5.4}$$

[5]: Bias of $(\bar{y}_{F-T})_2$ up to first order approximation is:

$$B(\bar{y}_{F-T})_2 = \bar{Y}\,M_1\left[\delta_1 C_1(\rho_{01}C_0 - \beta_1 C_1) + \delta_2 C_2(\alpha_2 C_2 - \rho_{02}C_0) - \delta_1\delta_2\rho_{12}C_1C_2\right] \tag{5.5}$$

[6]: Mean squared error of $(\bar{y}_{F-T})_2$ up to first order approximation is:

$$M(\bar{y}_{F-T})_2 = \bar{Y}^2 M_1\left[C_0^2 + \delta_1^2 C_1^2 + \delta_2^2 C_2^2 + 2\delta_1\,\rho_{01}C_0C_1 - 2\delta_2\,\rho_{02}C_0C_2 - 2\delta_1\,\delta_2\,\rho_{12}C_1\,C_2\right]$$

… (5.6)

Proof 5.2:

[4]:

$$(\bar{y}_{F-T})_2 = \bar{y}\,\frac{(A_1+C_1)\bar{X}_1 + fB_1\,\bar{x}_1}{(A_1+fB_1)\bar{X}_1 + C_1\bar{x}_1}\,\frac{(A_2+fB_2)\bar{X}_2 + C_2\bar{x}_2}{(A_2+C_2)\bar{X}_2 + fB_2\,\bar{x}_2}$$

$$(\bar{y}_{F-T})_2 = \bar{Y}\,(1+e_0)\,(1+\alpha_1 e_1)(1+\beta_2 e_2)(1+\beta_1 e_1)^{-1}(1+\alpha_2 e_2)^{-1}$$

$$(\bar{y}_{F-T})_2 = \bar{Y}\left[1 + e_0 + \delta_1(e_1 + e_0 e_1 - \beta_1 {e_1}^2) - \delta_2(e_2 + e_0 e_2 - \alpha_2 {e_2}^2) - \delta_1\delta_2 e_1 e_2\right]$$

[5]:

$$E\left[(\bar{y}_{F-T})_2-\bar{Y}\right]=\bar{Y}\,E\left[e_0+\delta_1(e_1+e_0e_1-\beta_1{e_1}^2)-\delta_2(e_2+e_0e_2-\alpha_2{e_2}^2)-\delta_1\delta_2e_1e_2\right]$$

$$B(\bar{y}_{F-T})_2=\bar{Y}M_1\left[\delta_1C_1(\rho_{01}C_0-\beta_1C_1\ )+\delta_2C_2(\alpha_2C_2-\rho_{02}C_0)-\delta_1\delta_2\rho_{12}C_1C_2\right]$$

[6]:

$$M\,(\bar{y}_{F-T})_2=E\left[(\bar{y}_{F-T})_2-\bar{Y}\right]^2$$

$$M(\bar{y}_{F-T})_2=\bar{Y}^2M_1\left[C_0^2+\delta_1^2C_1^2+\delta_2^2C_2^2+2\delta_1\,\rho_{01}C_0C_1-2\delta_2\,\rho_{02}C_0C_2-2\delta_1\,\delta_2\,\rho_{12}C_1\,C_2\right]$$

THEOREM 5.3:

[7]: The estimator $(\bar{y}_{F-T})_3$ in terms of e_0, e_1 and e_2 up to first order of approximation could be expressed as:

$$(\bar{y}_{F-T})_3=\bar{Y}\left[1+e_0+\delta_1(\alpha_1e_1^2-e_1-e_0e_1)+\delta_2(e_2-\beta_2e_2^2+e_0e_2)-\delta_1\delta_2e_1e_2\right]\tag{5.7}$$

[8]: Bias of $(\bar{y}_{F-T})_3$ up to first order approximation is:

$$B(\bar{y}_{F-T})_3=\bar{Y}M_1\left[\delta_1(\alpha_1C_1^2-\rho_{01}C_0C_1)+\delta_2(\rho_{02}C_0C_2-\beta_2C_2^2)-\delta_1\delta_2\rho_{12}C_1C_2\right]\tag{5.8}$$

[9]: Mean squared error of $(\bar{y}_{F-T})_3$ up to first order approximation is:

$$M(\bar{y}_{F-T})_3=\bar{Y}^2M_1\left[C_0^2+\delta_1^2C_1^2+\delta_2^2C_2^2-2\rho_{01}C_0C_1\delta_1+2\rho_{02}C_0C_2\delta_2-2\delta_1\delta_2\rho_{12}C_1C_2\right]\tag{5.9}$$

Proof 5.3:

[7]:

$$(\bar{y}_{F-T})_3=\bar{y}\;\frac{(A_2+C_2)\bar{X}_2+fB_2\,\bar{x}_2}{(A_2+fB_2)\bar{X}_2+C_2\bar{x}_2}\,\frac{(A_1+fB_1)\bar{X}_1+C_1\bar{x}_1}{(A_1+C_1)\bar{X}_1+fB_1\,\bar{x}_1}$$

$$(\bar{y}_{F-T})_3=\bar{Y}(1+e_0)(1+\alpha_2e_2)(1+\beta_1e_1)(1+\beta_2e_2)^{-1}(1+\alpha_1e_1)^{-1}$$

$$(\bar{y}_{F-T})_3=\bar{Y}\left[1+e_0+\delta_1(\alpha_1e_1^2-e_1-e_0e_1)+\delta_2(e_2-\beta_2e_2^2+e_0e_2)-\delta_1\delta_2e_1e_2\right]$$

[8]:

$$(\bar{y}_{F-T})_3=\bar{Y}\left[1+e_0+\delta_1(\alpha_1e_1^2-e_1-e_0e_1)+\delta_2(e_2-\beta_2e_2^2+e_0e_2)-\delta_1\delta_2e_1e_2\right]$$

$$B(\bar{y}_{F-T})_3=\bar{Y}M_1\left[\delta_1(\alpha_1C_1^2-\rho_{01}C_0C_1)+\delta_2(\rho_{02}C_0C_2-\beta_2C_2^2)-\delta_1\delta_2\rho_{12}C_1C_2\right]$$

[9]:

$$E\left[(\bar{y}_{F-T})_3-\bar{Y}\right]^2=E\left[\bar{Y}\left\{e_0+\delta_1(\alpha_1e_1^2-e_1-e_0e_1)+\delta_2(e_2-\beta_2e_2^2+e_0e_2)-\delta_1\delta_2e_1e_2\right\}\right]^2$$

$$M(\bar{y}_{F-T})_3=\bar{Y}^2M_1\left[C_0^2+\delta_1^2C_1^2+\delta_2^2C_2^2-2\rho_{01}C_0C_1\delta_1+2\rho_{02}C_0C_2\delta_2-2\delta_1\delta_2\rho_{12}C_1C_2\right]$$

6. Minimum Mean Squared Error & Optimal Choices for Proposed Estimator(s)

In this proposed estimator we have multiple choices of the combination $K_i; i=(1,2)$ and optimal conditions obtained by mean squared error of all proposed designs.

For minimum mean squared error by $(\bar{y}_{F-T})_1$ differentiating (5.3) with respect to δ_1 and δ_2 respectively and equating to zero.

$$\left.\begin{aligned}C_1^2\delta_1+C_1C_2\rho_{12}\delta_2+\rho_{01}C_0C_1=0\\ \rho_{12}C_1C_2\delta_1+C_2^2\delta_2+\rho_{02}C_0C_2=0\end{aligned}\right\}\qquad(6.1)$$

By solving these simultaneous equations, we have

$$\delta_1=\frac{C_0}{C_1}\frac{\rho_{02}\rho_{12}-\rho_{01}}{(1-\rho_{12}^2)}=\hat{\delta}_{11}\text{ and }\delta_2=\frac{C_0}{C_2}\frac{\rho_{01}\rho_{12}-\rho_{02}}{(1-\rho_{12}^2)}=\hat{\delta}_{12}\quad\ldots\quad(6.2)$$

At these values of $\hat{\delta}_{11}$ and $\hat{\delta}_{12}$ the minimum mean square error of the proposed estimator is

$$MSE(\bar{y}_{F-T})_1\big|_{Min}=\bar{Y}^2C_0^2M_1\left[1+V(V+2\rho_{01})+U(U+2\rho_{02})+2UV\rho_{12}\right]\qquad(6.3)$$

where $U=\dfrac{\rho_{01}\rho_{12}-\rho_{02}}{(1-\rho_{12}^2)}$ and $V=\dfrac{\rho_{02}\rho_{12}-\rho_{01}}{(1-\rho_{12}^2)}$

By adopting the same procedure we can obtain the minimum mean squared error corresponding to $(\bar{y}_{F-T})_2$ and $(\bar{y}_{F-T})_3$ by (5.6) and (5.9).

The information of optimization regarding $(\bar{y}_{F-T})_2$ and $(\bar{y}_{F-T})_3$ is

$$\hat{\delta}_{21}=\hat{\delta}_{11};\ \hat{\delta}_{22}=-\hat{\delta}_{12};\hat{\delta}_{21}=\hat{\delta}_{11}\text{and}\,\hat{\delta}_{22}=-\hat{\delta}_{12} \tag{6.4}$$

Rewriting (6.2), as

$$\left.\begin{aligned}\hat{\delta}_{11}&=\frac{C_0}{C_1}\frac{\rho_{02}\rho_{12}-\rho_{01}}{(1-\rho_{12}^2)}=\Delta_1(say)\\ \hat{\delta}_{12}&=\frac{C_0}{C_2}\frac{\rho_{01}\rho_{12}-\rho_{02}}{(1-\rho_{12}^2)}=\Delta_2(say)\end{aligned}\right] \tag{6.5}$$

From (6.5) we can obtain the relation in the form of characterizing scalar as follows

$$\left.\begin{aligned}&(\Delta_1+1)K_1^3+(f\Delta_1-f-8\Delta_1-9)K_1^2+(23\Delta_1-5f\Delta_1+5f+26)K_1\\ &\qquad\qquad+(4f\Delta_1-22\Delta_1-4f-24)=0\\ &(\Delta_2+1)K_2^3+(f\Delta_2-f-8\Delta_2-9)K_2^2+(23\Delta_2-5f\Delta_2+5f+26)K_2\\ &\qquad\qquad+(4f\Delta_2-22\Delta_2-4f-24)=0\end{aligned}\right] \quad \dots (6.6)$$

Above polynomial (6.6) provides three choices of K_1 and K_2 for the minimum mean squared errors of proposed estimators.

In the similar way $\hat{\delta}_{21}=\Delta_1;\ \hat{\delta}_{22}=-\Delta_2;\hat{\delta}_{21}=-\Delta_1\text{and}\,\hat{\delta}_{22}=\Delta_2$ will also provide the polynomials of degree three *i.e.* in each case we have three different choices of constant $K_i; i=1,2$ to improve the estimator.

7. Empirical study

The target in this section is to evaluate the gain in efficiencies (in terms of mse) obtained by the proposed estimators. To see the performance of the various estimators discussed here, we are considering two different population data used earlier by other researchers. The empirical analysis is discussed below.

Population - 1 [sources: Anderson (1958)]

y *: Head length of second son*

x_1 *: Head length of first son*

x_2 *: Head breadth of first son*

The required information is given in Table 7.1.

Table 7.1. Population – 1 Parameters.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
$\bar{Y}$	183.84	n	7	C_0	0.0546	ρ_{01}	0.7108
$\bar{X}_1$	185.72	N	25	C_1	0.0526	ρ_{02}	0.6932
$\bar{X}_2$	151.12	f	0.28	C_2	0.0488	ρ_{12}	0.7346

Table 7.2. Percent Relative Efficiency of various estimators with respect to mean per unit estimator for Population – 1.

Estimator(s)	$PRE(\bullet)$ with respect to $\bar{y}$		
	$(\bar{y}_{F-T})_1$	$(\bar{y}_{F-T})_2$	$(\bar{y}_{F-T})_3$
$\bar{y}$	100	100	100
t_1	72.29	75.1	62.8
t_2	75.10	72.29	15.15
t_3	145.04	149.41	40.9
t_4	179.03	179.03	30.32
t_5	62.8	15.15	72.29
t_6	15.15	62.8	75.1
t_7	40.9	22.76	145.04
t_8	30.32	30.32	179.03
t_9	151.64	46.43	135.00
t_{10}	46.43	151.64	23.79
t_{11}	211.67	98.12	89.24
t_{12}	164.53	164.53	59.77
t_{13}	178.66	32.91	178.66
t_{14}	32.91	178.66	32.91
t_{15}	156.51	62.39	156.51
$(\bar{y}_{F-T})_1^*$	**231.90468**	101.6880601	83.46525626
$(\bar{y}_{F-T})_2^*$	101.68806	**231.9046782**	36.9499305
$(\bar{y}_{F-T})_3^*$	83.465256	36.9499305	**231.9046782**

Population - 2 [sources: Steel and Torrie (1960, p.282)]

y : *Log of leaf burn in sec*

x_1 : *Potassium percentage*

x_2 : *Chlorine percentage*

The information regarding population -2 is given in Table 7.3.

Table 7.3. Population - 2 Parameters.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
$\bar{Y}$	0.6860	n	6	C_0	0.4803	ρ_{01}	0.1794
$\bar{X}_1$	4.6537	N	30	C_1	0.2295	ρ_{02}	-0.4996
$\bar{X}_2$	0.8077	f	0.20	C_2	0.7493	ρ_{12}	0.4074

Table 7.4. Percent Relative Efficiency of various estimators with respect to mean per unit estimator for Population – 1.

Estimator(s)	$\%RE(\bullet)$ with respect to $\bar{y}$		
	$(\bar{y}_{F-T})_1$	$(\bar{y}_{F-T})_2$	$(\bar{y}_{F-T})_3$
$\bar{y}$	100	100	100
t_1	17.67	75.5	20.89
t_2	75.50	17.67	34.69
t_3	57.125	149.82	55.87
t_4	94.61	94.61	71.44
t_5	20.89	34.69	17.67
t_6	34.69	20.89	75.50
t_7	55.87	76.10	57.12
t_8	71.44	71.44	94.61
t_9	19.54	59.01	20.41
t_{10}	59.01	19.54	47.98
t_{11}	64.46	143.70	64.06
t_{12}	102.94	102.94	94.59
t_{13}	20.02	53.33	20.02
t_{14}	53.33	20.02	53.33
t_{15}	64.858	131.16	64.85
$(\bar{y}_{F-T})_1^*$	**174.04**	40.64	70.53
$(\bar{y}_{F-T})_2^*$	40.64	**174.04**	43.93
$(\bar{y}_{F-T})_3^*$	70.53	43.93	**174.04**

8. Discussion & Conclusion

For population-1 the choices to optimization of mean squared error of $(\bar{y}_{F-T})_1$ can be derived from (6.5) which give a polynomial of degree three (6.6). On solution we have $[K_1]_1 = 6.0098$; $[K_1]_2 = 2.9586$; $[K_1]_3 = 1.6115$; $[K_2]_1 = 5.7733$; $[K_2]_2 = 2.9825$ and $[K_2]_3 = 1.634$. For $(\bar{y}_{F-T})_2$ the values are $[K_1]_4 = [K_1]_1$, $[K_1]_5 = [K_1]_2$; $[K_1]_6 = [K_1]_3$ and $[K_2]_4 = 1.9132$. Similarly for $(\bar{y}_{F-T})_3$ values are $[K_1]_7 = 1.9206$; $[K_2]_7 = [K_2]_1$; $[K_2]_8 = [K_2]_2$ and $[K_2]_9 = [K_2]_3$ whereas other roots are imaginary.

For population-2 the choices of the constant scalar K_i to reduce the mean squared error of $(\bar{y}_{F-T})_1$ are $[K_1]_1 = 39.9225$; $[K_1]_2 = 2.5859$; $[K_1]_3 = 1.0972$ and $[K_2]_1 = 1.939$. For $(\bar{y}_{F-T})_2$ the values are $[K_1]_4 = [K_1]_1$, $[K_1]_5 = [K_1]_2$; $[K_1]_6 = [K_1]_3$; $[K_2]_4 = 5.7698$; $[K_2]_5 = 2.8515$ and $[K_2]_6 = 1.6794$. Similarly for $(\bar{y}_{F-T})_3$ values are $[K_1]_7 = 1.9968$ and $[K_2]_7 = [K_2]_1$. The remaining roots are imaginary. $(\bar{y}_{F-T})_1^*$, $(\bar{y}_{F-T})_2^*$ and $(\bar{y}_{F-T})_3^*$ denotes the optimal efficiency gain with respect to mean per unit estimator in the above mentioned tables.

From these results it is certain that the proposed estimators submit a wide ground for the optimization by multiple choices of the characterizing scalar K_i. Since the generation of the estimators by the proposed classes is easy, a number of estimators can be able to achieve for more study. The proposed estimator proposed a wide choice for the characterizing scalar, which is the beauty of the proposed analysis.

By the compilation of the percentage relative efficiencies corresponding to population-1 and 2 shown in table-7.2 and table-7.4 it is clear that the proposed estimators are more efficient than the other existing estimators as ratio estimator, product estimator, dual to ratio estimator, mean per unit estimator, ratio-cum-product type estimator, etc., and many more chain type estimators which are discussed above, with considerable gain in terms of mean square error. Thus, the proposed estimators are recommended for use in practice.

REFERENCES

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

ANDERSON, T.W. (1958): *An introduction to multivariate statistical analysis,* John Wiley and Sons, Inc., New York.

COCHRAN, W.G. (2005): *Sampling Techniques.* John Wiley and Sons, New York.

COCHRAN, W.G. (1940): *The estimation of the yields of cereal experiments by sampling for the ratio gain to total produce,* Journal of Agricultural Society, 30, 262–275.

COCHRAN, W.G. (1942): *Sampling theory when the sampling units are of unequal sizes,* Journal of American Statistical Association, 37, 119–132.

DALABEHARA, M. and SAHOO, L.N. (1994): *A class of estimators in stratified sampling with two auxiliary variables,* Jour. Ind. Soc. Ag. Stat., 50, 2, 144 - 149.

DALABEHARA, M. and SAHOO, L.N. (2000): *An unbiased estimator in two - phase sampling using two auxiliary variables,* Jour. Ind. Soc. Ag. Stat., 53, 2, 134-140.

DESRAJ (1965): *On a method of using multi-auxiliary information in sample surveys,* Journal of American Statistical Association, 60, 270–277.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953): *Sample survey methods and theory,* John Wiley and Sons, New York.

KADILAR, C. and CINGI, H. (2004): *Estimator of a population mean using two auxiliary variables in simple random sampling,* International Mathematical Journal, 5, 357-360.

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

KHARE, B.B. and SRIVASTAVA, S.R. (1981): *A generalized regression ratio estimator for the population mean using two auxiliary variables,* The Aligarh Journal of Statistics, 1 (1), 43–51.

MUKHOPADHYAY, P. (2000): *Theory and methods of survey sampling,* Prentice Hall of India Pvt. Ltd., New Delhi.

MURTHY, M.N. (1964): *Product method of estimation,* Sankhya, 26, A, 294–307.

MURTHY, M.N. (1976): *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

NAIK, V.D. and GUPTA, P.C. (1991): *A general class of estimators for estimating population mean using auxiliary information*, Metrika, 38, 11-17.

PERRI, P.F. (2007): *Improved Ratio-cum-product type estimator*, Statistics in Transition, 8, 1, 51-69.

RAY, S.K. and SAHAI, A. (1980): *Efficient families of ratio and product type estimators,* Biometrika, 67, 215–217.

SAHOO, J. and SAHOO, L.N. (1993): *A class of estimators in two-phase sampling using two auxiliary variables,* Jour. Ind. Soc. Ag. Stat., 31, 107-114.

SAHOO, L.N., SAHOO, R.K. (2001): *Predictive estimation of finite population mean in two phase sampling using two auxiliary variables,* Jour. Ind. Soc. Ag. Stat., 54, 4, 258-264.

SHUKLA, D. (2002): *F-T estimator under two-phase sampling*, Metron, 59, 1-2, 253-263.

SHUKLA, D., SINGH, V.K. and SINGH, G.N. (1991): *On the use of transformation in factor type estimator,* Metron, 49, 1-4, 359-361.

SHUKLA, D., THAKUR, N.S., PATHAK SHARAD and RAJPUT, D.S. (2009): *Estimation of mean under imputation of missing data using factor-type estimator in two-phase sampling,* Statistics in Transition, 10, 3, 397-414.

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

SINGH, M.P. (1965): *Ratio cum product method of estimation,* Metrika, 12, 34-42.

SINGH, V.K. AND SHUKLA, D. (1987): *One parameter family of factor-type ratio estimator.* Metron, 45, 1-2, 273-283.

SINGH, V.K. AND SHUKLA, D. (1993): *An efficient one parameter family of factor-type estimator in sample survey.* Metron. 51, 1-2, 139-159.

SINGH, V.K. AND SHUKLA, D. (1987): *One parameter family of factor type ratio estimator,* Metron. 45, 1-2, 273-283.

SINGH, V.K. AND SINGH, G.N. (1991): *Chain type estimator with two auxiliary variables under double sampling scheme,* Metron, 49, 279-289.

SINGH, V.K., SINGH, G.N. AND SHUKLA, D. (1994): *A class of chain ratio estimator with two auxiliary variables under double sampling scheme,* Sankhya, Ser. B., 46, 2, 209-221.

SRIVASTAVA, S.K. (1971): A generalized estimator for the mean of a finite population using multi-auxiliary information, Journal of American Statistical Association, 66, 404–407.

SRIVASTAVA, S.K. and JHAJJ, H.S. (1983): A class of estimators of the population mean using multi-auxiliary information, Calcutta Statistical Association Bulletin, 32, 47–56.

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

STEEL, R.G.D and TORRIE J.H.(1960): *Principles and procedures of statistics*, Mc Graw Hill Book Co.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASHOK, C. (1984): *Sampling Theory of Surveys with Applications*, Iowa State University Press, I. S. A. S. Publication, New Delhi.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 37—46

MODIFIED ESTIMATORS OF POPULATION VARIANCE IN PRESENCE OF AUXILIARY INFORMATION

Dr. Rajesh Tailor[1], Balkishan Sharma[2]

ABSTRACT

This paper proposes estimator of population variance using information on known parameters of auxiliary variable. The variances of the proposed estimators are obtained. It has been shown that using modified sampling fraction the proposed estimators are more efficient than the usual unbiased estimator of population variance and usual ratio estimator for population variance under certain given conditions. Empirical study is also carried out to demonstrate the merits of the proposed estimators of population variance over other estimators considered in this paper.

Key words: Finite population variance, Bias, Mean squared error Auxiliary information and Efficiency.

1. Introduction

It is known fact that in many practical situations auxiliary information is available or may be made available in cheap cost in surveys. If this information is used intelligibly, it may give better estimators in terms of efficiency in comparison to the estimators in which auxiliary information is used.

The problem of constructing efficient estimators for the population variance S_y^2 has been widely discussed by various authors such as [3]Das and Tripathi (1978), [13]Srivastava and Jhajj (1980), [14]Upadhyay and Singh (1983), [4]Garcia and Cebrian (1996), [10]Singh S. and Joarder, A. H. (1998), [11]Singh et al. (1998), [1]Cebrian and Garcia (1997) and [12]Singh, H. P. and Singh, R. (2003). Later on [9]Singh and Tailor (2003) defined a generalized class of estimators of variance using population mean, variance, coefficient of variation of

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

2 Assistant Professor in Statistics. Department of Community Medicine. Sri Aurobindo Institute of Medical Sciences, INDORE, (M. P.), India. E-mail: bksnew@rediffmail.com.

auxiliary variate and correlation coefficient between study and auxiliary variate whereas [15]Upadhyaya, L. N. and Singh, H. P. (2006) and [7]Kadilar, C. and Cingi, H. (2006) considered the problem of estimating the variance of the ratio estimator.

These motivate authors to propose modified estimators of population variance based on sampling fraction using auxiliary information.

Let $U=(U_1,U_2,\dots,U_N)$ be the finite population of size N and y be a real valued function, i. e. random variable taking the values y_i (i=1,2, . . . , N) for the i^{th} unit of the population U.

Let $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}y_i$ and $S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$

denote unknown population mean and population mean square of (variance) the study character y. Suppose x is an auxiliary variate which is positively correlated with study variate y taking value x_i on unit U_i. Assuming that population size N is large so that finite population terms are ignored.

The usual unbiased estimator for population variance S_y^2 is given as

$$s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{Y})^2 \tag{1.1}$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ sample mean of y,

$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}y_i$ population mean of y and

$$S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$$

When the population variance of auxiliary variate S_x^2 is known, [5]Isaki (1983) proposed a ratio estimator for population variance S_y^2 of study variate y as

$$\hat{s}_R^2=s_y^2\left(\frac{S_x^2}{s_x^2}\right) \tag{1.2}$$

where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$ is an unbiased estimator for population variance

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$$

The variance of s_y^2 and mean squared error of $\hat{s}_R^2$ up to the first order of approximation are given as

$$V(s_y^2) = \frac{(N-n)}{Nn}S_y^4\left[\beta_2(y) - 1\right], \qquad (1.3)$$

$$MSE(\hat{s}_R^2) = \frac{(N-n)}{Nn}S_y^4\left[\beta_2(y) + \beta_2(x) - 2h\right], \qquad (1.4)$$

where $\beta_2(y) = \frac{\mu_{40}}{\mu_{20}^2}$, $\beta_2(x) = \frac{\mu_{04}}{\mu_{02}^2}$, $h = \frac{\mu_{22}}{\mu_{20}\mu_{02}}$,

$$S_y^4 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^4 \quad \text{and} \quad \mu_{st} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{Y})^s(x_i - \overline{X})^t .$$

Here $\beta_2(y)$ and $\beta_2(x)$ are the population coefficients of kurtosis of the study variate and auxiliary variate respectively.

2. Strategy-I

Using the sampling fraction f, modified ratio type estimator for population variance of study variate y is given as

$$t_1 = f\,s_y^2 + (1-f)\,s_y^2\frac{S_x^2}{s_x^2} \qquad (2.1)$$

To obtain the bias and mean squared error of suggested estimator t_1, we write

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

such that $E(e_i) = 0$, i=0 and 1

$$E(e_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\beta_2(y) - 1\right) ,$$

$$E(e_1^2)=\left(\frac{1}{n}-\frac{1}{N}\right)(\beta_2(x)-1)\ \text{ and}$$

$$E(e_0e_1)=\left(\frac{1}{n}-\frac{1}{N}\right)(h-1)\ .$$

Now, the suggested estimator t_1 in terms of $e_i^{'}$s may be written as

$$t_1=S_y^2\left[f\,(1+e_0)+(1-f)(1+e_0)(1+e_1)^{-1}\right],$$

$$E(t_1-S_y^2)=S_y^2\,E\left\{e_0-e_1(1-f)+e_1^2(1-f)-e_0e_1(1-f)\right\}, \tag{2.2}$$

$$B(t_1)=(1-f)\,S_y^2\ \left\{\beta_2(x)-h\right\}. \tag{2.3}$$

Squaring and taking expectation of both sides of equation (2.2), we get mean squared error of suggested estimator t_1, up to the first degree of approximation as

$$E(t_1-S_y^2)^2=S_y^4\,E\left\{e_0^2+(1-f)^2e_1^2-2(1-f)e_0e_1\right\}, \tag{2.4}$$

$$MSE(t_1)=\left[V(s_y^2)+(1-f)^2V(s_x^2)-2(1-f)Cov(s_y^2,s_x^2)\right], \tag{2.5}$$

where s_y^2 and s_x^2 are the population variances of the study and auxiliary variates respectively and

$$V(s_y^2)=\left(\frac{1}{n}-\frac{1}{N}\right)S_y^4\left[\beta_2(y)-1\right],$$

$$V(s_x^2)=\left(\frac{1}{n}-\frac{1}{N}\right)S_x^4\left[\beta_2(x)-1\right]\ \text{ and}$$

$$Cov(s_y^2,\,s_x^2)=\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2\,S_x^2\left[h-1\right].$$

3. Efficiency comparisons for t_1.

From (8.1.3), (8.1.4) and (8.2.5), it is observed that

(i) Suggested estimator t_1 would be more efficient than usual unbiased estimator for population variance s_y^2, i. e.

$MSE(t_1)$ - $V(s_y^2)$<0 if

$$\frac{1-f}{2} < \frac{Cov(s_y^2, s_x^2)}{V(s_x^2)} \tag{3.1}$$

(ii) Suggested estimator t_1 would be more efficient than [5]Isaki (1983) ratio estimator $\hat{s}_R^2$, i. e.

$MSE(t_1)$ - $MSE(\hat{s}_R^2)$<0 if

$$\frac{f}{2} < \left[1 - \frac{Cov(s_y^2, s_x^2)}{V(s_x^2)}\right] \tag{3.2}$$

4. Strategy-II

Another modified ratio type estimator for population variance of study variate y using the sampling fraction f, is given as

$$t_2 = \left[\left(\frac{1-f}{1+2f}\right)s_y^2 + \left(\frac{3f}{1+2f}\right)s_y^2 \frac{S_y^2}{s_x^2}\right] \tag{4.1}$$

To obtain the bias and mean squared error of suggested estimator t_2, we write

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

such that $E(e_i) = 0$, i=0 and 1

$E(e_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)(\beta_2(y) - 1)$, $E(e_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right)(\beta_2(x) - 1)$ and

$E(e_0 e_1) = \left(\frac{1}{n} - \frac{1}{N}\right)(h-1)$

Now, the suggested estimator t_2 may be written in terms of $e_i^{'}$s as

$$t_2 = S_y^2\left[\frac{1-f}{1+2f}(1+e_0)+\frac{3f}{1+2f}(1+e_0)(1+e_1)^{-1}\right]$$

$$E(t_2 - S_y^2) = S_y^2\, E\left\{e_0 - e_1\frac{3f}{(1+2f)} + e_1^2\frac{3f}{(1+2f)} - e_0e_1\frac{3f}{(1+2f)}\right\}$$

(4.2)

$$B(t_2) = \frac{3f}{(1+2f)}S_y^2\,\{\beta_2(x) - h\} \tag{4.3}$$

Squaring and taking expectation of both sides of equation (4.2), we get Mean squared error of suggested estimator t_2, up to the first degree of approximation as

$$E(t_2 - S_y^2)^2 = S_y^4\, E\left\{e_0^2 + \left(\frac{3f}{1+2f}\right)^2 e_1^2 - 2\left(\frac{3f}{1+2f}\right)e_0e_1\right\},$$

(4.4)

$$MSE(t_2) = \left[V(s_y^2) + \left(\frac{3f}{1+2f}\right)^2 V(s_x^2) - 2\left(\frac{3f}{1+2f}\right)COV(s_y^2, s_x^2)\right],$$

(4.5)

where s_y^2 and s_x^2 are the population variances of the study and auxiliary variates respectively and expressed in previous section.

5. Efficiency comparisons for t_2

From (1.3), (1.4) and (4.5) it is observed that

(i) Suggested estimator t_2 would be more efficient than usual unbiased estimator for population variance s_y^2, i. e.

$MSE(t_2)$ - $V(s_y^2)$<0 if

$$\frac{3f}{1+2f} < \frac{2\,Cov(s_y^2, s_x^2)}{V(s_x^2)} \tag{5.1}$$

(ii) Suggested estimator t_2 would be more efficient than Isaki (1983) ratio estimator $\hat{s}_R^2$, i. e.

$MSE\,(t_2)$ - $MSE\,(\hat{s}_R^2)$ <0 if

$$\frac{1+3f}{1+f} < \frac{2\,Cov\,(s_y^2\,,s_x^2)}{V\,(s_x^2)} \tag{5.2}$$

6. Strategy-III

The suggested modified ratio type estimator for population variance of study variate y is given as

$$t_3 = \left[\left(\frac{1-f}{1+3f}\right)s_y^2 + \left(\frac{4f}{1+3f}\right)s_y^2\,\frac{S_y^2}{s_x^2}\right] \tag{6.1}$$

To obtain the bias and mean squared error of suggested estimator t_3, we write

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

such that $E(e_i) = 0$, i=0 and 1

$E(e_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\beta_2(y) - 1\right)$, $E(e_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\beta_2(x) - 1\right)$ and

$E(e_0 e_1) = \left(\frac{1}{n} - \frac{1}{N}\right)(h-1)$

Now, the suggested estimator t_3 may be written in terms of $e_i^{'}$s as

$$t_3 = S_y^2\left[\frac{1-f}{1+3f}(1+e_0) + \frac{4f}{1+3f}(1+e_0)(1+e_1)^{-1}\right],$$

$$E(t_3 - S_y^2) = S_y^2\,E\left\{e_0 - e_1\frac{4f}{(1+3f)} + e_1^2\frac{4f}{(1+3f)} - e_0 e_1\frac{4f}{(1+3f)}\right\}, \tag{6.2}$$

$$B(t_3) = \frac{4f}{(1+3f)}S_y^2\,\left\{\beta_2(x) - h\right\}. \tag{6.3}$$

Squaring and taking expectation of both sides of equation (6.2), we get mean squared error of suggested estimator t_3, up to the first degree of approximation as

$$E(t_3 - S_y^2)^2 = S_y^4 \, E\left\{ e_0^2 + \left(\frac{4f}{1+3f}\right)^2 e_1^2 - 2\left(\frac{4f}{1+3f}\right) e_0 e_1 \right\}$$

(6.4)

$$MSE(t_3) = \left[V(s_y^2) + \left(\frac{4f}{1+3f}\right)^2 V(s_x^2) - 2\left(\frac{4f}{1+3f}\right) Cov(s_y^2, s_x^2) \right] \quad (6.5)$$

where s_y^2 and s_x^2 are the population variances of the study and auxiliary variates respectively and expressed in previous section.

7. Efficiency comparisons for t_3

From (1.3), (1.4) and (6.5) it is observed that

(i) Suggested estimator t_3 would be more efficient than usual unbiased estimator for population variance s_y^2, i. e.

$MSE\,(t_3)$- $V(s_y^2)$<0 if

$$\frac{4f}{1+3f} < \frac{2\,Cov\,(s_y^2\,,s_x^2)}{V\,(s_x^2)} \quad (7.1)$$

(ii) Suggested estimator t_3 would be more efficient than [5]Isaki (1983) ratio estimator $\hat{s}_R^2$, i. e.

$MSE\,(t_3)$- $MSE\,(\hat{s}_R^2)$<0 if

$$\frac{1+4f}{1+2f} < \frac{2Cov(s_y^2\,,s_x^2)}{V\,(s_x^2)} \quad (7.2)$$

Section 3, 5 and 7 provided the conditions under which proposed estimators t_1, t_2 and t_3 have less mean squared errors in comparison to usual unbiased estimator for population variance and ratio estimator for population variance.

8. Empirical study

To analyze the performance of the proposed estimators t_1, t_2 and t_3 in comparison to other estimators, we consider the data given in [8]Murthy (1967, p.-226). The variates and data set is given as

y : Output and

x : number of workers.

N=80, n=30,

$\beta_2(y)$ =2.2667, $\beta_2(x)$ =3.65 and h=2.3377.

Table 8.1. Percent Relative Efficiencies of $\hat{s}_R^2$, t_1, t_2 and t_3 with respect to s_y^2.

PRE's	Estimators				
	s_y^2	$\hat{s}_R^2$	t_1	t_2	t_3
Percent Relative Efficiencies	100	102.05	150.53	155.75	146.88

It is observed from the table 8.1 that there is a significant gain in efficiency by using proposed variance estimators t_1, t_2 and t_3 in comparison to unbiased estimator for population variance s_y^2 and ratio estimator for population variance $\hat{s}_R^2$ given by [5]Isaki (1983).

Therefore suggested estimators t_1, t_2 and t_3 are recommended for their use in practice.

REFERENCES

CEBRIAN, A.A. and GARCIA, R.M. (1997). Variance estimation using auxiliary information. An almost unbiased multivariable ratio estimator. Metrika, 45, 171-178.

COCHRAN, W.G. (1977). Sampling Techniques. Third U. S. Edition. Wiley Eastern Limited, 325.

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

GARCIA, M. RUEDA and CEBRIAN, A.A. (1996). Repeated substitution method. The ratio estimator of the population variance. Metrika, 43, 101-105.

ISAKI, C.T. (1983). Variance estimation using auxiliary information, Journal of the American Statistical Association 78, 117-123.

KADILAR, C. and CINGI, H. (2006). Ratio estimators for population variance in simple and stratified sampling, Applied Mathematics and Computation 173, 1047–1058, 2006.

KADILAR, C. and CINGI, H. (2006). Improvement in variance estimation using auxiliary information, Hacettepe Journal of Mathematics and Statistics 35 (1), 111–115.

MURTHY, M.N. (1967). Sampling theory and methods, statistical publishing society, Calcutta.

SINGH, H. P. and TAILOR, R. (2003). Use of known correlation coefficient in estimating the finite population mean, “Statistics in Transition”, 6, 555-560.

SINGH, S. and JOARDER, A.H. (1998). Estimation of finite population variance using random non-response in surveys sampling. Metrika,47, 241-249.

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

SINGH, H.P. and SINGH, R. (2003). Estimation of variance through regression approach in two phase sampling. Alig. Jour. Stat., 23, 13-30.

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

UPADHYAYA, L.N. AND SINGH, H.P. (1983). Use of auxiliary information in the estimation of population variance. Mathematical forum, 4, 2, 33-36.

UPADHYAYA, L.N. and SINGH, H.P. (2006). Almost unbiased ratio and product-type estimators of finite population variance in sample surveys, Statist. in Transi.,7 5, 1087–1096.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 47—64

CROP ACREAGE AND CROP PRODUCTION ESTIMATES FOR SMALL DOMAINS - REVISITED

G.C. Tikkiwal[1], Alka Khandelwal[2]

ABSTRACT

For any country advance and final estimates of yield of principle crops, at National and State levels, are of great importance for its macro level planning. But, for decentralized planning and for other purposes like crop insurance, loan to farmers, etc., the reliable estimates of crop production for small domains are also in great demand. This paper, therefore, discusses and review critically the methodology used to provide crop acreage and crop production estimates for small domains, based on indirect methods of estimation, including the SICURE model approach. The indirect methods of estimation so developed use data obtained either through traditional surveys, like General Crop Estimation Surveys (GCES) data, or a combination of the surveys and satellite data.

Key words: Timely Reporting Scheme (TRS); General Crop Estimation Surveys (GCES); Simulation-cum- Regression (SICURE) model.

1. Introduction

The advance and final estimates of crop production of principle crops at national and sub-national level like districts, counties, blocks for any country are of importance for its macro and micro level planning.

In many countries, including India, the yield rate of principle crops are being estimated through crop-cutting experiments. The technique of crop-cutting experiments is mostly developed in India in early seventies. The estimation of crop yield is done under the national programme known as General Crop Estimation Surveys (GCES) using crop-cutting experiments. The GCES are being conducted through survey methodology developed mostly in 1940's [Mahalanobis (1946), Sukhatme and Agarwal (1946-47, 1947-48)].

[1] Department of Mathematics and Statistics, Jai Narain Vyas University, Jodhpur – 342005, India. E-mail: gctikkiwal@yahoo.com.

[2] Department of Mathematics and Statistics, Jai Narain Vyas University, Jodhpur – 342005, India. E-mail: alkakhandelwaljdr@gmail.com.

The estimation of crop yield involves two components viz. the estimation of crop acreage and the estimation of yield rates. As regard crop acreage estimation, a scheme known as Timely Reporting Scheme (TRS) has been in vogue since early seventies in most of the States of India. The TRS has the objective of providing quick and reliable estimates of crop acreage statistics and thereby production of the principle crops during each agricultural season on the basis of 20 percent sample villages, using direct estimators. The performance of direct estimators is satisfactory at national and state level, as the sampling error of the estimators is within 5 percent, but not at lower levels as shown by Tikkiwal and Tikkiwal (1998), Tikkiwal and Ghiya (2000, 2004). The authors developed and used synthetic and composite methods of estimation to provide crop acreage estimation for small domains. Further, it has been observed that composite method of estimation is easy to apply and this approach overcomes the limitation of synthetic estimator to some extent [cf. comments by Francisco (1998, p.254)]. Where this approach does not work, then SICURE model can be tried or other model based estimation methods may provide satisfactory results.

Apart from traditional approach the remote sensing technologies were initiated after launch of many advanced satellites, to provide crop acreage and crop production estimates for major and minor domains [Dadhwal et al. (1985)]. For example, the National Agricultural Statistics Service (NASS) of the United States of America has been using Landsat series of satellites since 1950's, and France entered the field of earth resources satellites in 1986 with the launch of SPOT-I [cf. Bellow et al. (1996)]. In India this work has been entrusted to Indian Space Research Organization. The model based estimation methods for small domains, using survey and satellite data has been developed over a period of time by authors Battese et al. (1988), Singh et al. (1992), Shaible and Casady (1994), Srivastava (2007) and others. These methods may provide efficient estimators provided a suitable model is selected and there should not be problem of mixed cropping.

This paper provides a comprehensive review of the work done on crop production estimates of small domains. In this paper Section 2 describes the methods of estimating crop acreage statistics. The Section 3 describes the crop-cutting experiments and presents method of estimation of crop yield. Section 4 discusses and review the methodology used to provide crop acreage and crop production estimates for small domains, based on survey data, whereas the model based estimation methods for small domains using survey and satellite data are discussed in Section 5.

2. Crop acreage statistics

In Temporarily Settled States (the states, in which land revenue is fixed for a definite period of years and is subject to revision at the end of this period) of India crop acreage statistics are collected on complete enumeration basis, whereas in

Permanently Settled States (the states, in which land revenue is fixed for perpetuity) they are estimated through selection of 20 percent villages. In order to provide quick and reliable advanced estimates of the crop production, in temporarily settled states also crop acreage statistics are estimated under Timely Reporting Scheme (TRS). The TRS has been in vogue since early seventies in these States of India. Under the scheme the Patwari (Village Accountant) is required to collect acreage statistics on a priority basis in a 20 percent sample of villages. These villages are selected by stratified linear systematic sampling scheme, taking Tehsil as a stratum. These statistics are further used to provide state level estimates using direct estimators viz. unbiased (based on sample mean) and ratio estimators.

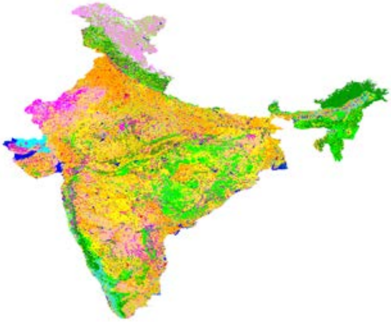
The performance of both direct estimators in the state of Rajasthan, like in other states, is satisfactory at state level, as the sampling error is within 5 percent. However, the sampling error of both direct estimators increases considerably, when they are used for estimating acreage statistics of various principle crops even at district level, what to speak of levels lower than a district. Tikkiwal and Ghiya (2000, 2004) notice that the sampling error of direct ratio estimator for Kharif crops of Jodhpur district (of Rajasthan state) for the agricultural season 1991-92 varies approximately between 6 to 68 percent. Therefore, there is a need to use indirect estimators at district and lower levels for decentralized planning and other purposes like crop insurance, bank loan to farmers. As regards estimation of yield rates, it is being done through crop-cutting experiments.

It may be noted here that for administrative purposes India is divided into the number of states, each state consists of a number of districts, each district consists of a number of tehsils and further each tehsil consists of a number of villages.

The crop acreage statistics are also collected by the Indian Space Research Organization under its Crop Acreage and Crop Production (CAPE) project, through remote sensing technology. But due to mixed cropping pattern, prevailed in India, this technique of the crop acreage statistics are not so reliable.

Land Use and Land Cover statistics of India and the state of Rajasthan are shown here using the satellite data obtained from Regional Remote Sensing Centre-West, Indian Space Research Organization, Department of Space (India).

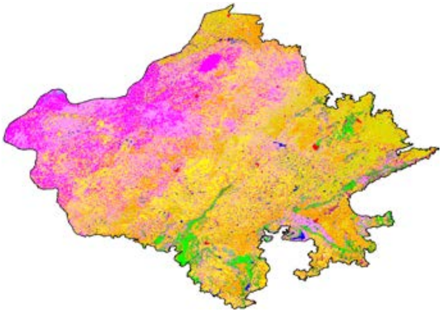
LAND USE / LAND COVER STATISTICS OF INDIA (2010-11)



	Categories	Area (Sq Km)
🟥	Built-up	20705.73
🟨	Kharif only	540441.5
🟨	Rabi only	258442.8
🟩	Zaid only	9387.03
🟧	Double / tripple	544436.1
🟧	Current fallow	385835.11
🟦	Plantation/orchard	67798.33
🟩	Evergreen forest	167870.67
🟩	Deciduous forest	337899.13
🟩	Scrub/Deg. forest	145263.53
🟦	Littoral swamp	4659.24
🟩	Grassland	74915.82
🟪	Other wasteland	289810.76
🟪	Gullied	9986.8
🟪	Scrubland	186970.02
🟦	Water bodies	79362.73
⬜	Snow covered	40645.31
🟫	Shifting Cultivation	1769.02
🟦	Rann	19274.94
	Total	**3185474.57**

Source: AWIFS Satellite data (2010-11)

LAND USE / LAND COVER STATISTICS OF RAJASTHAN (2010-11)



	Categories	Area (Sq Km)
🟥	Built-up	855.74
🟨	Kharif only	46773.01
🟨	Rabi only	40726.94
🟫	Zaid only	869.21
🟧	Double/Triple crop	52818.26
🟧	Current fallow	56693.91
🟩	Deciduous forest	8406.61
🟩	Shrub/scrub/degraded forest	9792.06
🟦	Grassland/grazing land	121.39
🟪	Other wastelands	67727.72
🟪	Gullied/ravines land	1169.01
🟪	Scrub land	39627.74
🟦	Waterbodies	4684.84
	Total	**330266.43**

Source: AWIFS Satellite data (2010-11)

In Indian system there are mainly three agricultural season's viz. Kharif, Rabi and Zaid.

(i) Kharif crops – the crops sown in June-July and harvested in October-November every year.

(ii) Rabi crops – the crops sown in November-December and harvested March-April every year.

(iii) Zaid crops – the crops grown between March and June are known as Zaid.

3. Estimation of crop yield

Final estimates of crop production based on complete enumeration of area and yield become available much after the crops are actually harvested. However, the Government may require advance estimates of production for taking various policy decisions relating to pricing, marketing, export/import, distribution, etc. Considering the genuine requirement of crop estimates much before the crops are harvested for various policy purposes, a time schedule of releasing the advance estimates has been evolved under a national programme known as “General Crop Estimation Surveys (GCES)”. The GCES uses the technique of crop-cutting experiments.

3.1. Crop – cutting experiments under GCES

The most important factor of Crop production statistics is the estimation of yield rates. Presently the yield rates are estimated through crop-cutting experiments under GCES. The GCES covers 68 crops (52 foods and 16 non foods) in 22 states and 04 union territories. Such surveys are conducted twice a year to cover different types of crops.

About five hundred thousand crop-cutting experiments, for major crops throughout the country, are conducted annually under this programme. The sampling design adopted for the GCES is a multistage stratified random sampling with tehsils/inspector land revenue circles/community development blocks, etc. as strata, the villages selected randomly form the primary stage sampling unit, the fields from each selected village formed the second stage sampling unit and the experimental plot within the field form the ultimate stage of sampling. A sample of villages is selected from different strata in proportion to the area under crop. From each selected village, two fields are selected randomly and from each field a plot of fixed shape and size usually measuring (5meter x 5meter) is selected for recording the green yield by actual harvesting the crop.

3.2. Estimation Procedure

Estimation procedure for estimating of crop yield through Crop Estimation Surveys:

The methodology generally adopted for estimating the average yield of crop is as below:

At the stratum (tehsil) level, the estimated average yield of the crop is obtained as a simple arithmetic mean of plot yields. For this, let

Y_{ijk} – The green yield (net in gms/plot) of the k-th plot in the j-th village in the i-th stratum.

n_{ij} – Number of experiments analyzed in the j-th village of i-th stratum.

m_i – Number of villages in which experiments are analyzed in the i-th stratum.

n_i – Number of experiments analyzed in the i-th stratum.

S – Number of strata in a district.

a_i – The area (net) of the crop in the i-th stratum.

f – The conversion factor for converting the green yield per plot into the yield of dry marketable produce per hectare.

Stratum level average of the green yield for the i-th stratum is

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

and further, District level estimated average yield of the dry marketable produce per hectare is given by

$$\bar{Y} = \frac{\sum_{i=1}^{s} a_i.\bar{Y}_i}{\sum_{i=1}^{s} a_i}.\, f$$

Then, $\bar{Y}$ is to be multiplied by the District level crop acreage estimates of that particular crop, to have an estimate of the yield.

4. Estimation for small domains using survey data

For decentralized planning and other purposes like crop insurance, loan to farmers, etc., the Governments need reliable agricultural statistics for small domains like district, CD block, counties, etc. But the estimates provided by National Agencies such as NSSO, TRS and EARAS are generally reliable at the state level and not at district level. In such situation "Small Area Estimation" methods hold out a promising solution.

4.1. Synthetic and composite methods

Tikkiwal, B.D. and Tikkiwal, G.C. (1998) in their invited paper presents an excellent review of the landmarks in the development of crop yield and acreage statistics in India and other developing countries. As regards providing estimates of average crop yield at small area (Assistant Agricultural Officer (AAO) circle) level the authors use direct methods only, because the sample size was sufficiently large. In the absence of such information/sufficient data the SICURE model can be helpful. The authors further demonstrate the use of synthetic and composite estimators to provide reliable acreage statistics at small area levels. The small areas in this study are Inspector Land Revenue Circles (ILRC's), the sub-groups of Tehsils. The study suggests the use of composite estimators, if the synthetic assumption closely meets. When this assumption does not meet, they suggest the use of other types of estimators such as those obtained through the SICURE model (1993). The following discussant Francisco Juvier Gallego's (1998, p. 254) comments on this paper show applicability of the results.

"…The approach might overcome some limitations of synthetic estimators and looks easier to apply than other small area estimation procedures that have been used in agricultural statistics [For example, Battese et al. (1988)]. Some additional clarifications would be of interest on the computation of variance from a single sample. If the results presented are confirmed in other countries, the method would be of interest, and not only for developing countries, as stated in the paper. Actually, India is a developed country if we speak about statistics".

Tikkiwal, G.C. and Ghiya, A. (2000) define and discuss a generalized class of synthetic estimators with application to crop acreage estimation for small domains (ILRC's), using auxiliary information, under different sampling schemes. The generalized class of synthetic estimators, among others, includes the simple, ratio and product synthetic estimators. The proposed class of synthetic estimators gives consistent estimators if the synthetic assumption holds. Further, the authors compare the relative performance of a number of synthetic estimators with direct estimators, empirically, through a simulation study using live data. The study reveals that for the domains where synthetic estimators do not deviate considerably from their corresponding assumption, performance of the synthetic estimator is satisfactory. When the synthetic estimators deviate considerably from their corresponding assumption, then the authors suggest to look for other types of estimators such as those obtained through the SICURE model [Tikkiwal (1993)].

Sisodia and Singh (2001) develop three synthetic estimators of total crop production Y_i of i-th block (small area) level using crop production and other relative information at district level, as given below:

Estimator (1)

$$\hat{Y}_i = \left(\sum_{j=1}^{P} W_j X_j \right) \hat{\bar{Y}} \qquad (4.1.1)$$

where, $\hat{\bar{Y}} = \dfrac{\hat{Y}}{A}$; $\hat{Y}$ is obtained through multiple linear regression model.

A = Area under the crop in a given year

X_j = Value of j-th predictor at the block level in a given year

W_j = Weight assign to each predictor.

Estimator (2) & Estimator (3) are of the form

$$\tilde{Y}_i = b_i \ \hat{Y}_i$$

where, b_i are constants such that $\sum_{i=1}^{a} \tilde{Y}_i = \sum_{i=1}^{a} b_i \hat{Y}_i = Y$

For $b_i = b$ (constant) second estimator of Y_i is, i.e.

Estimator (2)

$$\tilde{Y}_i^{(1)} = \hat{Y}_i \frac{Y}{\sum_{i=1}^{a} \hat{Y}_i} \qquad (4.1.2)$$

Y = actual crop production reported at district level through crop-cutting experiments in a given year.

For $b_i = 1 + \dfrac{\left(Y - \sum_{i=1}^{a} \hat{Y}_i\right)}{a\hat{Y}_i}$ third estimator of Y_i is, i.e.

Estimator (3)

$$\tilde{Y}_i^{(2)} = \hat{Y}_i + \frac{\left(Y - \sum_{i=1}^{a} \hat{Y}_i\right)}{a}$$

(4.1.3)

a = Number of blocks in the district.

Further, the authors carried out an empirical study for rice crop in Faizabad district of Uttar Pradesh during the years 1981-82 and 1982-83 to compare the relative efficiency of these estimators under multiple linear regression model. The relative efficiency of $\tilde{Y}_i^{(1)}$ and $\tilde{Y}_i^{(2)}$ over $\hat{Y}_i$ comes out to be same for all the blocks, i.e. 88.84% and 105.88% respectively during the year 1981-82. Similarly, during the year 1982-83 it comes out to be 110.80% and 105.88% respectively.

Thus $\tilde{Y}_i^{(2)}$ is found to be most efficient when comparing with $\hat{Y}_i$ and $\tilde{Y}_i^{(1)}$ in case 1 when weights are given to be more than 1. In case 2 when weights are less than 1 both estimators $\tilde{Y}_i^{(1)}$ and $\tilde{Y}_i^{(2)}$ are found to be more efficient than $\hat{Y}_i$. But $\tilde{Y}_i^{(2)}$ need not be the most efficient estimator. The results presented in the Table 4 and Table 5 (p. 313 & 315) does not correlate with the findings, when the estimated values are compared with the actual estimates based on crop-cutting experiments.

All the three estimators considered by the authors are nothing but synthetic, regression type, estimators and, therefore, their efficiency depends on the validity of the assumption of the corresponding synthetics estimator under use. Also, the three estimators are design-biased; therefore, ignoring the bias remains a serious limitation. But these estimators can be further improved upon by the technique of composite estimation. [cf. Tikkiwal and Ghiya (2004)].

Tikkiwal and Ghiya (2004) define and discuss a generalized class of composite estimators for small domains, using auxiliary information, under different sampling schemes. The proposed estimator of population mean $\overline{Y}_i$, based on auxiliary variable 'x' under SRSWOR design is defined as:

$$\bar{y}_{c,i} = w_i \bar{y}_i \left(\frac{\bar{x}_i}{\overline{X}_i} \right)^{\beta_1} + (1 - w_i)\bar{y} \left(\frac{\bar{x}}{\overline{X}_i} \right)^{\beta_2} ; \quad (0 \le \mathrm{w_i} \le 1) \qquad (4.1.4)$$

where, β_1 and β_2 are suitably chosen constants.

The estimator $\bar{y}_{c,i}$ is a weighted sum of the generalized direct estimator [Srivastava (1967)] and the generalized synthetic estimator [Tikkiwal and Ghiya (2000)].

The proposed estimator has desirable consistency property (in traditional sense), when the following assumption is satisfied.

$$\overline{Y}_i \left(\overline{X}_i \right)^{\beta_2} \cong \overline{Y} \left(\overline{X} \right)^{\beta_2} \qquad (4.1.5)$$

It is to be noted that the synthetic estimator may be heavily biased unless the above assumption is satisfied [cf. Tikkiwal and Ghiya (2000), Eq. (4.1)].

The proposed generalized class of composite estimators includes a number of direct, synthetic and composite estimators as special cases. Here follows a list of such estimators with corresponding choice of values of the different constants.

Table 4.1. Various Direct, Synthetic and Composite Estimators as Special Cases of the Generalized Composite Estimators

S No	Estimator	w_i	$(1-w_i)$	β_1	β_2
1	Simple Direct $(\bar{y}_i)$	1	0	0	-
2	Simple Synthetic $(\bar{y})$	0	1	-	0
3	Simple Ratio $\left[\left(\bar{y}_i / \bar{x}_i\right)\bar{X}_i\right]$	1	0	-1	-
4	Ratio Synthetic $\left[\left(\bar{y} / \bar{x}\right)\bar{X}_i\right]$	0	1	-	-1
5	Simple Product $\left[\left(\bar{x}_i / \bar{X}_i\right)\bar{y}_i\right]$	1	0	1	-
6	Product Synthetic $\left[\left(\bar{y} / \bar{X}_i\right)\bar{x}\right]$	0	1	-	1
7	Composite : Combining simple direct with simple synthetic $w_i\bar{y}_i + (1-w_i)\bar{y}$	w_i	$(1-w_i)$	0	0
8	Composite: combining simple direct with ratio synthetic $w_i\bar{y}_i + (1-w_i)\frac{\bar{y}}{\bar{x}}\bar{X}_i$	w_i	$(1-w_i)$	0	-1
9	Composite : combining simple ratio with ratio synthetic $w_i\frac{\bar{y}_i}{\bar{x}_i}\bar{X}_i + (1-w_i)\frac{\bar{y}}{\bar{x}}\bar{X}_i$	w_i	$(1-w_i)$	-1	-1

Further, the authors comparing the various empirical results of Absolute Relative Bias (ARB) and Simulated relative standard error (Srse), draw the conclusion that if the synthetic estimators do not deviate considerably from their corresponding assumptions (describe in Eq. 4.1.5), then performance of the composite estimators (given at S.No.9 in the Table 4.1), based on a sample of 20% villages, is satisfactory at the level of ILRCs. Therefore, these estimators will certainly perform better up to the level of district. When the given condition is not satisfied we should look for other methods of estimation. One of such method is to use SICURE model (1993) or the methods presented in Ghosh & Rao (1994).

Sharma, Srivastava and Sud (2004) consider two different synthetic estimators based on auxiliary variables for providing crop yield estimate at Gram Panchayat (small area) level. The proposed estimators for i-th Gram Panchayat (GP) are defined as follows:

$$\hat{T}_i = \frac{\bar{x}_i}{\sum_{i=1}^{a} A_i \bar{x}_i} A\hat{\bar{Y}} , \quad \text{i = 1, 2,..., a} \tag{4.1.6}$$

and,

$$\hat{T}_{\text{i}}^{'} = \bar{x}_{\text{i}} + \hat{\beta}_{iopt}\left[\hat{\bar{Y}} - \frac{\sum_{i=1}^{a} A_i \bar{x}_i}{A}\right] \tag{4.1.7}$$

$$\hat{\beta}_{iopt} = \frac{\left(A_i \big/ A\right)\hat{\sigma}_{x_i}^2}{n_i\left[\hat{V}\left(\hat{\bar{Y}}\right) + \frac{1}{A^2}\sum_{i=1}^{a}\frac{A_i^2 \sigma_{x_i}^2}{n_i}\right]}$$

a = number of GP in a block.

A_i = Area under a particular crop for the i-th GP.

$A = \sum_{i=1}^{a} A_i$ = Total area under the crop in the block.

N_i = Number of farmers in the i-th GP.

n_i = Number of farmers selected in the i-th GP for obtaining information about the expected yield of the crop grown in the field.

x_{ij} = expected yield as obtained from j-th farmer in the i-th GP; j = 1,2,...,n_i .

$\hat{\bar{Y}}$ = block level estimate of crop yield as obtained through the method of crop-cutting experiment.

$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, average of expected yield of i-th GP.

In the empirical study the proposed estimators are based on crop yield estimates obtained through crop-cutting experiments under General Crop Estimation Surveys and estimate of crop production obtained through data collected from a fresh selected random sample of 10 farmers from each of all the GP’s in a district. Analysis of data obtained from a survey carried out on wheat crop in the Basti district of the state of Uttar Pradesh in India in the year 2000 revealed that both the estimators perform satisfactorily in terms of the criterion of percentage root mean squared error as it varies from 1% to 10% in most of the cases. The biases of both the estimators are also negligible. Here, it may be noted that the estimators proposed by Sharma et al. (2004) depend on the estimate of crop production obtained through crop-cutting experiments and on the basis of fresh samples selected independently from each Gram Panchayat (GP) of the block. In the case study, for example, a block roughly consists of 90 GP's which results in a selection of an additional sample of 900 farmers. This method, therefore, does not fall within the domain of small area estimation methods. Also basis of the proposed estimator is the assumption that "over estimation and under estimation", with respect to estimates obtained through method of crop-cutting experiments, behave in similar way in all the domains of a block of interest; which is not realistic. Apart from this there are errors in the formulae of Bias and Mean Square Errors [Eq. (6.2), (6.3) of the paper].

5. Estimation for small domains using satellite and survey data

National Agricultural Statistics Service (NAAS) of the U.S. Department of Agriculture has been a user of remote sensing data since 1950's when it began using mid-altitude aerial photography to construct sampling frames for the 48 states of the continental United States. A new era in remote sensing began in 1972 with the launch of the Landsat I earth-resource monitoring satellite. Four Landsats have been launched since 1972 with Landsat IV and V which are still in operation. A regression estimator was developed which related the ground-gathered area frame data to the computer classification of Landset MSS (multi-spectral scanner images). The basic regression approach used to produce state estimates does not produce reliable county (small area) estimates. Three domain indirect regression estimators have been used or considered for producing small area county estimates using ancillary satellite data by NAAS. From 1972 to 1982 the Huddleston-Ray estimator was used, from 1983 to 1990 Battese-Fuller family of estimators was used and since 1991, the Battese-Fuller model has been used to produce country estimates with Landsat TM (Thematic Mapper) data. The details of these models have been discussed in detail by Bellow et al. (1996).

In India, as mentioned above, at present crop area statistics are based on complete enumeration of all fields and crop yield statistics based on GCES. With the advent of remote sensing technology satellite data has been widely used by many countries including India for obtaining various crop statistics. Several studies have been conducted during the past decade by the India's Department of Space under the Crop Acreage and Production Estimation (CAPE) project for crop acreage and production estimation for various major crops using satellite spectral data. Recently some studies have been taken at the Indian Agricultural Statistics Research Institute, New Delhi, India to develop more efficient estimator of crop yield using satellite data along with survey data of crop yield based on crop-cutting experiments. [cf. Singh et al. (1992), Goel et al. (1994), Shaible and

Casady (1994), Singh et al. (1999), Singh and Goel (2000), Singh (2004), Srivastava (2007)].

Singh and Geol (2000) used synthetic estimators to provide the yield estimates at Tehsil and Block levels, using crop yield data for Rabi crops 1997-98 obtained from GCES and the satellite spectral data of IRS-1D LISS-III. The study shows that the standard error of synthetic estimator is less than the corresponding direct estimator. The study also developed yield estimates at District level, using the direct estimator under post-stratification. The standard error of the direct estimator at district level is very small (around 5%). This confirms the results of earlier study by Singh et al. (1999).

Singh (2003) used the farmer's eye estimate of crop yield corresponding to the crop plots selected for crop-cutting experiment as an auxiliary variable along with the vegetation indices for improving the crop yield forecasting models. The yield data pertains to wheat crop yield data for district Rohtak for the year 1995-96 based on crop-cutting experiments. Spectral data in the form of vegetation indices RVI and NDVI has been obtained from IRS 1B-LISS II dated February 17, 1996 for the region. The farmer's eye estimate is obtained from the selected farmers for the fields in which crop-cutting experiments were conducted. Singh (2004) reviewing the earlier work also developed regression estimates using RVI (x_1), NDVI (x_2) and farmers eye estimate of crop yield of the corresponding plot (x_3) as auxiliary variables for forecasting cop yield at district level.

In all the above studies the performance of the synthetic estimators are measured in terms of standard errors. However, ignoring the bias remains a serious limitation.

In country like India almost 70% population is dependent on agriculture. The farm sizes in India are very small with diversified crops in each season. The practice of mixed cropping is quite dominant. Therefore, it may not be possible to prepare accurate area frame using remote sensing technology due to limitations of satellite sensor in detecting and differentiating small fields and crops grown, both for major as well as for minor domains.

Rao, J. N. K. (2004) provides an appraisal of indirect estimates, both traditional and model based. He provides a brief account of small area estimation in the context of agriculture surveys. He presents model based small area estimation under a basic area level model and a basic unit level model. He reviews work of Fuller (1981), Battese et al. (1988), Stasny et al. (1991) and Singh and Goel (2000).

Fuller (1981) applies the mixed area level model

$$\hat{\theta}_i = z_i^T \beta + v_i + e_i \,, \qquad \text{i=1,2,...,m} \tag{5.1}$$

to estimate mean soybean hectares per segment in 1978 at the county level.

This model is combination of a **basic area level model**

$$\hat{\theta}_i = \theta_i + e_i , \qquad \text{i=1,2,...,m}$$

and a **linear regression model**

$$\theta_i = z_i^T \beta + v_i , \qquad \text{i = 1,2,...m}$$

where, sampling error e$_i$' s are assumed to be independent across area with mean 0 and known variance ψ_i, and

model error $v_i^{'}$s are assumed to be independent and identically distributed with mean 0 and variance σ_v^2, $z_i = (z_{1i},..., z_{Pi})^T$ area specific auxiliary variates.

Using the data $\left\{\left(\hat{\theta}_i , z_i\right), i = 1,..., m\right\}$ we can obtain estimates, θ_i^*, of the realized values of θ_i from the mixed model.

It may be noted that empirical best linear unbiased prediction (EBLUP) method is applicable for mixed linear models and its estimates do not require normality assumption on the random errors v$_i$ and e$_i$. EBLUP estimate of θ_i is a composite estimate of the form

$$\theta_i^* = \hat{w}_i \hat{\theta}_i + \left(1 - \hat{w}_i\right) z_i^T \hat{\beta} ; \ \hat{w}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} \tag{5.2}$$

which is a weighted combination of direct estimate $\hat{\theta}_i$ and a regression synthetic estimate $z_i^T \hat{\beta}$.

$\hat{\beta}$ is the weighted least square estimate of β with weights $\left(\hat{\sigma}_v^2 + \psi_i\right)^{-1}$.

$\hat{\sigma}_v^2$ is an estimate of the variance component σ_v^2.

Fuller obtained model based estimates of the population means, $\overline{Y}_i$ for the sampled county (m=10) as well as the non sampled counties. His model is given by

$$\overline{y}_i - z_{3i} = \beta_0 + \beta_1 z_{2i} + \beta_2 z_{3i} + v_i + e_i \tag{5.3}$$

with known error variance σ_v^2 and σ_e^2 .

z_{2i} = mean number of pixels of soybeans per area segment ascertained by satellite imaginary.

z_{3i} = mean soybean hectares from the 1974 U.S. Agricultural Census, as county (area) level covariates.

Note that z_{2i} and z_{3i} are known for all the 16 counties.

The model (5.3) is a special case of (5.2) with $\hat{\theta}_i = \bar{y}_i - z_{3i}$ and $\psi_i = \psi = \sigma_e^2$. Fuller's estimate of $\bar{Y}_i$ for sampled counties is obtained from (5.2.2) as

$$\bar{y}_{iF}^{*} = g^{-1}\left(\theta_i^{*}\right) = \theta_i^{*} + z_{3i}$$

$$= z_{3i} + z_i^T \hat{\beta} + w\left(\bar{y}_i - z_{3i} - z_i^T \hat{\beta}\right), \quad \text{i}\in\text{s}$$

where, $$w = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2}$$

For the non sampled counties, $\bar{y}_{iF}^{*} = z_{3i} + z_i^T \hat{\beta}$, i∉s

He concludes that the model based estimates, $\bar{y}_{iF}^{*}$, outperform in term of total MSE. They are also better than the direct estimates $\bar{y}_i$ in terms of total MSE for the sampled counties.

Battese et al. (1988) also consider the problem of crop acreage estimation using farm interview data in conjunction with LANDSAT satellite data. The authors use the nested error linear regression model $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$; j = 1,2,..., N$_i$; i = 1,...,m to estimate area under corn and area under soybeans for i-th small area (counties) in north-central Iowa.

y_{ij} is variable under study related to unit-specific auxiliary data $x_{ij} = \left(1, x_{1ij}, x_{2ij}\right)^T$ and normally distributed errors v_i and e_{ij}.

Authors present the EBLUP estimates of small area means for both crops. Estimated standard errors of the EBLUP estimates and the survey regression estimates $\left[\bar{y}_i + \left(\bar{X}_i - \bar{x}_i\right)^T \hat{\beta}\right]$ are also given. The ratio of the estimated standard error of the EBLUP estimate to that of the survey regression estimate decreases as the size of sample decreases.

Rao (2004) further uses Hierarchical Bayes approach to test the fitness of the model given by Battese et al. with auxiliary data x_{ij}. Under the criterion of posterior probabilities use, it is noted that for values of such probabilities close to 0.5 it indicated good fit but for probabilities close to 0 and 1, it suggests poor fit of the model.

The major problem with the model based approach is of selection a suitable model. Therefore, selection and validation play a vital role in model based estimation. If the assumed models do not provide a good fit to the sample data, the model based approach can lead to erroneous estimates.

Acknowledgement

The research was supported by the grant received from Department of Science and Technology, New Delhi under Women Scientist Scheme.

The authors are thankful to the Referee for his constructive suggestions which helped to improve the paper.

REFERENCES

BATTESE, G.E., HARTER, R.M. and FULLER, W.A. (1988): An error-components model for prediction of crop areas using survey and satellite data. J. Amer. Statist. Assoc., 83, 28-36.

BELLOW, M., GRAHAM, M. and IWIG, W. C. (1996): Country estimation of crop acreage using satellite data. Indirect Estimation in U.S. Fedral Programs, Springer, 104-123.

CSO (1990): Working group on small area development programme statistics. Report, Department of Statistics, Ministry of Planning, Government of India, New Delhi.

DADHWAL, V.K. and PARIHAR, J.S. (1985): Estimation of 1983-84 wheat acreage of Karnal district (Haryana) using landsat MSS digital data. Technical note, IRS-UP/SAC/CPF/TN/09, Space Application Centre, Ahmedabad.

FRANCISCO, J.G. (1998): Comments on the invited paper entitled “Small area estimation in India – Crop yield and acreage statistics” by Tikkiwal, B.D. and Tikkiwal, G.C. Proceedings of the International Conference “Agricultural Statistics – 2000”, 238 – 254.

FULLER, W.A. (1981): Regression estimation for small areas. Rural America in Passage: Statistics for Policy, National Academy Press, Washington, D.C., 572-586.

GHOSH, M. and RAO, J.N.K. (1994): Small area estimation: An appraisal. Statist. Sci., 9, 55-93.

MAHALANOBIS, P.C. (1946): Sample surveys of crop yields in India. Sankhya, 269-280.

RAO, J.N.K. (2004): Small area estimation with applications to agriculture. J. Ind. Soc. Agril. Statist., 57, 159-170.

SCHAIBLE, W.L. and CASADY, R.J. (1994): The development, application, and evaluation of small area estimators. Statistics in Transition, 1 (6), 727-746.

SHARMA, S.D., SRIVASTAVA, A.K. and SUD, U.C. (2004): Small area crop estimation methodology for crop yield estimates at Gram Panchayat level. J. Ind. Soc. Agril. Statist., 57, 26-37.

SINGH, D. (1968): Double sampling and its application in agriculture. J. Ind. Soc. Agril. Statist. (Panse Memorial volume).

SINGH, R. (2003): Use of satellite data and farmers eye estimate for crop yield modeling. J. Ind. Soc. Agril. Statist., 56 (2), 166-176.

SINGH, R. (2004): Application of remote sensing technology for crop yield estimation. J. Ind. Soc. Agril. Statist, 57, 226-246.

SINGH, R. and GOEL, R.C. (2000): Use of remote sensing satellite data in crop surveys. Technical Report, Indian Agricultural Statistics Research Institute, New Delhi.

SINGH, R., GOEL, R.C., PANDEY, L.M. and SAHA, S.K. (1999): Use of remote sensing technology in crop yield estimation surveys- II, Project Report, IASRI, New Delhi.

SINGH, R., GOYAL, R.C., SAHA, S.K. and CHHIKARA, R.S. (1992): Use of satellite spectral data in crop yield estimation surveys. Int. J. Rem. Sens., 13 (14), 2583-2592.

SISODIA, B.V.S. and SINGH, A. (2001): On small area estimation – An empirical Study. J. Ind. Soc. Agril. Statist., 54 (3), 303-316.

SRIVASTAVA, A.K. (2007): Small area estimation – A perspective and some applications. J. Ind. Soc. Agril. Statist., 61 (3), 295-309.

STASNY, E.A., GOEL, P.K. and RAMSEY, D.J. (1991): County estimates of wheat production. Survey Methodology, 17, 211-255.

SUKHATME, P.V. and AGARWAL, O.P. (1946-47, 1947-48): Crop-cutting survey on wheat by the random sampling method. Report, Indian Council of Agricultural Research, New Delhi.

TIKKIWAL, B.D. (1991): Modeling through survey data for small domains. Keynote Address at the Symposium on Modeling held at Kurukshetra University, March 7 – 9, 1991.

TIKKIWAL, B.D. (1993): Modeling through survey data for small domains. Proc. Scientific Conference on Small Area Statistics and Survey Design held in September 1992 at Warsaw, Poland.

TIKKIWAL, B.D. and TIKKIWAL, G.C. (1998): Small area estimation in India – Crop yield and acreage statistics. Invited paper in the Proceedings of the International Conference “Agricultural Statistics – 2000” (along with the comments of Dr. J.G. Francisco, the discussant of the paper), 238 – 254. The Conference was held in March 1998 at Washington, D.C., USA.

TIKKIWAL, G.C. and GHIYA, A. (2000): A generalized class of synthetic estimators with application to crop acreage estimation for small domains. Biometrical Journal, 42 (7), 865-876.

TIKKIWAL, G.C. and GHIYA, A. (2004): A generalized class of composite estimators with application to crop acreage estimation for small domains. Statistics in Transition, 1(6), 697-711.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 65—78

ESTIMATION OF POPULATION MEAN IN POST-STRATIFIED SAMPLING USING KNOWN VALUE OF SOME POPULATION PARAMETER(S)

Aloy C. Onyeka[1]

ABSTRACT

Following Khoshnevisan et.al. (2007) and Koyuncu and Kadilar (2009), this paper develops a general family of combined estimators of the population mean in post-stratified sampling (PSS) scheme, using known values of some population parameters of an auxiliary variable. Properties of the proposed family of estimators, including conditions for optimal efficiency, are obtained up to first order approximations. The results are illustrated empirically.

Key words: Auxiliary information, general family of estimators, post-stratified sampling, mean squared errors
2000 AMS Classification: 62D05

1. Introduction

The use of auxiliary information in constructing estimators of population parameters of the variable of interest is highly encouraged in surveys, especially when an auxiliary variable is highly correlated with the study variable. Apart from using the known population mean $\overline{X}$ of an auxiliary variable x, many authors have ventured into the use of other known population parameters of x. Notable studies along this line, under the simple random sampling without replacement (SRSWOR) scheme, include Searls (1964), who used the coefficient of variation (CV) of x in estimating the population mean $\overline{Y}$ of the study variable y, and Sisodia and Dwivedi (1981), who used the CV of x in ratio estimation of $\overline{Y}$. Singh et.al. (1973) used known coefficient of kurtosis in estimating the population variance of y. Sen (1978) and Searls and Intarapanich (1990) also used known coefficient of kurtosis in estimating $\overline{Y}$. Singh and Tailor (2003) used known correlation coefficient in ratio estimation of $\overline{Y}$. Singh (2003) used known

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

standard deviation of the auxiliary variable x in estimating $\overline{Y}$. Khoshnevisan et.al. (2007) proposed a general family of estimators of $\overline{Y}$ under the SRSWOR scheme, which uses known parameters of the auxiliary variable x such as standard deviation, coefficient of variation, skewness, kurtosis and correlation coefficient. Motivated by Khoshnevisan et.al. (2007), the present study intends to develop a general family of estimators of $\overline{Y}$ under the post-stratified sampling scheme.

Under the stratified random sampling scheme, Cochran (1977) discussed the usual stratified sampling estimator, $\overline{y}_{st}$, and also separate and combined ratio-type estimators of $\overline{Y}$. Kadilar and Cingi (2003), motivated by the works done under the SRSWOR scheme by Sisodia and Dwivedi (1981), Singh and Kakran (1993), and Upadhyaya and Singh (1999), proposed some estimators of $\overline{Y}$ in stratified random sampling using known values of population mean $\overline{X}_h$, coefficient of variation C_{xh}, and coefficient of kurtosis $\beta_{2h}(x)$ of the auxiliary variable x in stratum h. Kadilar and Cingi (2003) restricted their work to ratio estimation of $\overline{Y}$ in stratified random sampling. Koyuncu and Kadilar (2009) proposed a more general family of combined estimators of $\overline{Y}$ in stratified random sampling along the line of Khoshnevisan et.al. (2007). Chaudhary et.al. (2009) considered, in a more recent paper, a general family of combined estimators of $\overline{Y}$ in stratified random sampling under non-response. Motivated by Khoshnevisan et.al. (2007) and Koyuncu and Kadilar (2009), we consider in the present study, a general class of combined-type estimators of $\overline{Y}$ in post-stratified sampling scheme, using information on known values of some population parameters of an auxiliary variable.

2. The Proposed Estimators

Let y_{hi} (x_{hi}) denote the ith observation in stratum h for the study (auxiliary) variate in post-stratified sampling scheme. Let a random sample of size n be drawn from a population of N units using SRSWOR method, and let the sampled units be allocated to their respective strata, where n_h (a random variable) is the number of units that fall into stratum h such that $\sum_{h=1}^{L} n_h = n$. We assume that n is large enough such that $P(n_h = 0) = 0, \forall\ h$. Following Khoshnevisan et.al. (2007) and Koyuncu and Kadilar (2009), we propose a general family of combined estimators of the population mean $\overline{Y}$ in post-stratified sampling scheme as

$$\bar{y}_{pss} = \bar{y}_{ps}\left(\frac{a\bar{X}+b}{\alpha(a\bar{x}_{ps}+b)+(1-\alpha)(a\bar{X}+b)}\right)^{g} \tag{2.1}$$

where,

$\bar{y}_{ps} = \sum_{h=1}^{L}\omega_h\bar{y}_h$ is the usual post-stratified estimator of $\bar{Y}$

$\bar{x}_{ps} = \sum_{h=1}^{L}\omega_h\bar{x}_h$ is the usual post-stratified estimator of $\bar{X}$

$\bar{X} = \sum_{h=1}^{L}\omega_h\bar{X}_h$ is the known population mean of the auxiliary variate x.

$a(\neq 0)$, b are either constants or functions of known population parameters of the auxiliary variate, such as Standard deviation (σ_x), Coefficient of variation (C_x), Skewness $(\beta_1(x))$, Kurtosis $(\beta_2(x))$, and Correlation coefficient (ρ_{yx}), and

$\omega_h = N_h/N$ is stratum weight, L is the number of strata in the population, N_h is the number of units in stratum h, N is the number of units in the population, $\bar{X}_h$ is the population mean of the auxiliary variate in stratum h, and $\bar{y}_h$ $(\bar{x}_h)$ is the sample mean of the study (auxiliary) variate in stratum h.

Under the conditional argument, that is, for the achieved sample configuration, $\underline{n} = (n_1, n_2, \cdots, n_L)$, the post-stratified estimator, $\bar{y}_{ps}$ is unbiased for $\bar{Y}$ with variance,

$$V_2(\bar{y}_{ps}) = \sum_{h=1}^{L}\omega_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{yh}^2}{n_h} = \sum_{h=1}^{L}\frac{\omega_h^2 S_{yh}^2}{n_h} - \frac{1}{N}\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{2.2}$$

where V_2 refers to “conditional variance” and S_{yh}^2 is the population variance of y in stratum h.

For repeated samples of fixed size n, we obtain the unconditional variance of $\bar{y}_{ps}$ by taking the expectation of equation (2.2). This gives the unconditional variance of $\bar{y}_{ps}$ as:

$$V(\bar{y}_{ps}) = E(V_2(\bar{y}_{ps})) = \sum_{h=1}^{L}\omega_h^2\, E\left(\frac{1}{n_h}\right)S_{yh}^2 - \frac{1}{N}\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{2.3}$$

Following Stephan (1945), we obtain, to terms of order n^{-2},

$$E\left(\frac{1}{n_h}\right)=\frac{1}{n\omega_h}+\frac{1-\omega_h}{n^2\omega_h^2} \tag{2.4}$$

Consequently, the unconditional variance of $\bar{y}_{ps}$ obtained up to first order approximation is

$$V(\bar{y}_{ps})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{2.5}$$

Similarly, the unconditional variance of $\bar{x}_{ps}$ and the unconditional covariance of $\bar{y}_{ps}$ and $\bar{x}_{ps}$ are obtained respectively as

$$V(\bar{x}_{ps})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{xh}^2 \tag{2.6}$$

and

$$Cov\,(\bar{y}_{ps}\,,\bar{x}_{ps})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{yxh} \tag{2.7}$$

where $f=n/N$ is the population sampling fraction, S_{xh}^2 is the population variance of x in stratum h, and S_{yxh} is the population covariance of y and x in stratum h. Let

$$e_0=\frac{\bar{y}_{ps}-\bar{Y}}{\bar{Y}} \tag{2.8}$$

and

$$e_1=\frac{\bar{x}_{ps}-\bar{X}}{\bar{X}} \tag{2.9}$$

Under the unconditional argument, we obtain

$$E(e_0)=E(e_1)=0 \tag{2.10}$$

$$E(e_0^2)=\frac{V(\bar{y}_{ps})}{\bar{Y}^2}=\frac{1}{\bar{Y}^2}\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{2.11}$$

$$E(e_1^2)=\frac{V(\bar{x}_{ps})}{\bar{X}^2}=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{xh}^2 \tag{2.12}$$

and

$$E(e_0 e_1)=\frac{Cov(\bar{y}_{ps}\,,\bar{x}_{ps})}{\bar{Y}\bar{X}}=\frac{1}{\bar{Y}\bar{X}}\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{yxh} \tag{2.13}$$

We can rewrite equation (2.1) in terms of e_0 and e_1 as

$$\bar{y}_{pss}=\bar{Y}(1+e_0)(1+\alpha\lambda e_1)^{-g} \tag{2.14}$$

where $\lambda=\frac{a\bar{X}}{a\bar{X}+b}$. Assuming $|\alpha\lambda e_1|<1$, so that the series $(1+\alpha\lambda e_1)^{-g}$ converges, and expanding equation (2.14) up to first order approximations in expected value, we obtain the expressions:

$$(\bar{y}_{pss}-\bar{Y})=\bar{Y}(e_0-\alpha\lambda g e_1-\alpha\lambda g e_0 e_1+\tfrac{1}{2}\alpha^2\lambda^2 g(g+1)e_1^2) \tag{2.15}$$

and

$$(\bar{y}_{pss}-\bar{Y})^2=\bar{Y}^2(e_0^2+\alpha^2\lambda^2 g^2 e_1^2-2\alpha\lambda g e_0 e_1) \tag{2.16}$$

To obtain the unconditional bias and mean squared error of the proposed estimators $\bar{y}_{pss}$ we take the unconditional expectations of equations (2.15) and (2.16), and use equations (2.10) – (2.13) to make the necessary substitutions. This gives the unconditional bias and mean squared error of $\bar{y}_{pss}$, respectively as

$$B(\bar{y}_{pss})=\frac{\alpha\lambda g}{2\bar{X}}\left(\frac{1-f}{n}\right)\left(\sum_{h=1}^{L}\omega_h(\alpha\lambda(g+1)RS_{xh}^2-2S_{yxh})\right) \tag{2.17}$$

and

$$MSE(\bar{y}_{pss})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\alpha^2\lambda^2 g^2 R^2 S_{xh}^2-2\alpha\lambda g R S_{yxh}) \tag{2.18}$$

where $R=\bar{Y}/\bar{X}$.

3. Special Cases

The proposed estimator, $\bar{y}_{pss}$ is a general class of estimators capable of generating an infinite number of combined estimators of $\bar{Y}$ by making appropriate choices of the values of α, g, a and b in equation (2.1). The following are some special cases of the proposed estimators, $\bar{y}_{pss}$, of $\bar{Y}$ in post-stratified sampling scheme.

Estimator	Values of			
	α	g	a	b
1. Usual stratified estimator, $\bar{y}_{pss}(1) = \bar{y}_{ps} = \sum_{h=1}^{L} \omega_h \bar{y}_h$	–	0	–	–
2. Usual Combined ratio-type estimator, $\bar{y}_{pss}(2) = \bar{y}_{psRC} = \frac{\bar{y}_{ps}}{\bar{x}_{ps}} \bar{X}$	1	1	1	0
3. Sisodia-Dwivedi (1981) estimator, $\bar{y}_{pss}(3) = \bar{y}_{psSD} = \bar{y}_{ps} \frac{\bar{X} + C_x}{\bar{x}_{ps} + C_x}$	1	1	1	C_x
4. Singh-Kakran (1993) estimator (1), $\bar{y}_{pss}(4) = \bar{y}_{psSK1} = \bar{y}_{ps} \frac{\bar{X} + \beta_2(x)}{\bar{x}_{ps} + \beta_2(x)}$	1	1	1	$\beta_2(x)$
5. Upadhyaya-Singh (1999) estimator (1), $\bar{y}_{pss}(5) = \bar{y}_{psUS1} = \bar{y}_{ps} \frac{\bar{X}\beta_2(x) + C_x}{\bar{x}_{ps}\beta_2(x) + C_x}$	1	1	$\beta_2(x)$	C_x
6. Upadhyaya-Singh (1999) estimator (2), $\bar{y}_{pss}(6) = \bar{y}_{psUS2} = \bar{y}_{ps} \frac{\bar{X}C_x + \beta_2(x)}{\bar{x}_{ps}C_x + \beta_2(x)}$	1	1	C_x	$\beta_2(x)$

Estimator	Values of			
	α	g	a	b
7. Singh-Tailor (2003) estimator (1), $\bar{y}_{pss}(7)=\bar{y}_{psST1}=\bar{y}_{ps}\dfrac{\bar{X}+\rho_{yx}}{\bar{x}_{ps}+\rho_{yx}}$	1	1	1	ρ_{yx}
8. The usual combined product-type estimator, $\bar{y}_{pss}(8)=\bar{y}_{psPC}=\dfrac{\bar{y}_{ps}\bar{x}_{ps}}{\bar{X}}$	1	–1	1	0
9. Pandey-Dubey (1988) estimator, $\bar{y}_{pss}(9)=\bar{y}_{psPD}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}+C_x}{\bar{X}+C_x}$	1	–1	1	C_x
10. Upadhyaya-Singh (1999) estimator (3), $\bar{y}_{pss}(10)=\bar{y}_{psUS3}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}\beta_2(x)+C_x}{\bar{X}\beta_2(x)+C_x}$	1	–1	$\beta_2(x)$	C_x
11. Upadhyaya-Singh (1999) estimator (4), $\bar{y}_{pss}(11)=\bar{y}_{psUS4}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}C_x+\beta_2(x)}{\bar{X}C_x+\beta_2(x)}$	1	–1	C_x	$\beta_2(x)$
12. G.N. Singh (2003) estimator (1), $\bar{y}_{pss}(12)=\bar{y}_{psGNS1}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}+\sigma_x}{\bar{X}+\sigma_x}$	1	–1	1	σ_x
13. G.N. Singh (2003) estimator (2), $\bar{y}_{pss}(13)=\bar{y}_{psGNS2}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}\beta_1(x)+\sigma_x}{\bar{X}\beta_1(x)+\sigma_x}$	1	–1	$\beta_1(x)$	σ_x
14. G.N. Singh (2003) estimator (3), $\bar{y}_{pss}(14)=\bar{y}_{psGNS3}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}\beta_2(x)+\sigma_x}{\bar{X}\beta_2(x)+\sigma_x}$	1	–1	$\beta_2(x)$	σ_x
15. Singh-Tailor (2003) estimator (2), $\bar{y}_{pss}(15)=\bar{y}_{psST2}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}+\rho_{yx}}{\bar{X}+\rho_{yx}}$	1	–1	1	ρ_{yx}
16. Singh-Kakran (1993) estimator (2), $\bar{y}_{pss}(16)=\bar{y}_{psSK2}=\bar{y}_{ps}\dfrac{\bar{x}_{ps}+\beta_2(x)}{\bar{X}+\beta_2(x)}$	1	–1	1	$\beta_2(x)$

Notice that the usual post-stratified estimator $\bar{y}_{ps}$ is a special case of the proposed family of estimators $\bar{y}_{pss}$ if and only if we choose $g=0$, no matter the values of α, a and b. Again, we observe that the next six (6) special cases, $\bar{y}_{pss}(i)$, $i=2,\cdots,7$ are ratio-type estimators, while the remaining nine (9) special cases $\bar{y}_{pss}(i)$, $i=8,\cdots,16$ are examples of product-type estimators of $\bar{Y}$ in post-stratified sampling scheme.

4. Efficiency Comparisons

Using equation (2.18), we obtain the unconditional variance/mean squared errors of the estimators, $\bar{y}_{pss}(i)$, $i=1,2,\cdots,16$ as follows:

$$V(\bar{y}_{pss}(1))=V(\bar{y}_{ps})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{4.1}$$

$$MSE(\bar{y}_{pss}(2))=MSE(\bar{y}_{psRC})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+R^2S_{xh}^2-2RS_{yxh}) \tag{4.2}$$

$$MSE(\bar{y}_{pss}(3))=MSE(\bar{y}_{psSD})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_1^2S_{xh}^2-2\theta_1S_{yxh}) \tag{4.3}$$

$$MSE(\bar{y}_{pss}(4))=MSE(\bar{y}_{psSK1})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_2^2S_{xh}^2-2\theta_2S_{yxh}) \tag{4.4}$$

$$MSE(\bar{y}_{pss}(5))=MSE(\bar{y}_{psUS1})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_3^2S_{xh}^2-2\theta_3S_{yxh}) \tag{4.5}$$

$$MSE(\bar{y}_{pss}(6))=MSE(\bar{y}_{psUS2})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_4^2S_{xh}^2-2\theta_4S_{yxh}) \tag{4.6}$$

$$MSE(\bar{y}_{pss}(7))=MSE(\bar{y}_{psST1})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_5^2S_{xh}^2-2\theta_5S_{yxh}) \tag{4.7}$$

$$MSE(\bar{y}_{pss}(8))=MSE(\bar{y}_{psPC})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+R^2S_{xh}^2+2RS_{yxh}) \tag{4.8}$$

$$MSE(\bar{y}_{pss}(9))=MSE(\bar{y}_{psPD})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_1^2S_{xh}^2+2\theta_1S_{yxh}) \tag{4.9}$$

$$MSE(\bar{y}_{pss}(10))=MSE(\bar{y}_{psUS3})=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2+\theta_3^2S_{xh}^2+2\theta_3S_{yxh}) \tag{4.10}$$

$$MSE(\bar{y}_{pss}(11)) = MSE(\bar{y}_{psUS4}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_4^2 S_{xh}^2 + 2\theta_4 S_{yxh}) \tag{4.11}$$

$$MSE(\bar{y}_{pss}(12)) = MSE(\bar{y}_{psGNS1}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_6^2 S_{xh}^2 + 2\theta_6 S_{yxh}) \tag{4.12}$$

$$MSE(\bar{y}_{pss}(13)) = MSE(\bar{y}_{psGNS2}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_7^2 S_{xh}^2 + 2\theta_7 S_{yxh}) \tag{4.13}$$

$$MSE(\bar{y}_{pss}(14)) = MSE(\bar{y}_{psGNS3}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_8^2 S_{xh}^2 + 2\theta_8 S_{yxh}) \tag{4.14}$$

$$MSE(\bar{y}_{pss}(15)) = MSE(\bar{y}_{psST2}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_5^2 S_{xh}^2 + 2\theta_5 S_{yxh}) \tag{4.15}$$

$$MSE(\bar{y}_{pss}(16)) = MSE(\bar{y}_{psSK2}) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h(S_{yh}^2 + \theta_2^2 S_{xh}^2 + 2\theta_2 S_{yxh}) \tag{4.16}$$

where

$$R = \frac{\bar{Y}}{\bar{X}}, \qquad \theta_1 = \frac{\bar{Y}}{\bar{X} + C_x}, \qquad \theta_2 = \frac{\bar{Y}}{\bar{X} + \beta_2(x)}, \qquad \theta_3 = \frac{\bar{Y}\beta_2(x)}{\bar{X}\beta_2(x) + C_x},$$

$$\theta_4 = \frac{\bar{Y}C_x}{\bar{X}C_x + \beta_2(x)},$$

and

$$\theta_5 = \frac{\bar{Y}}{\bar{X} + \rho_{yx}}, \; \theta_6 = \frac{\bar{Y}}{\bar{X} + \sigma_x}, \; \theta_7 = \frac{\bar{Y}\beta_1(x)}{\bar{X}\beta_1(x) + \sigma_x}, \; \theta_8 = \frac{\bar{Y}\beta_2(x)}{\bar{X}\beta_2(x) + \sigma_x}$$

Applying the least squares method, the (optimum) choice of α that minimizes equation (2.18), is obtained as

$$\alpha_{opt} = \frac{\beta_0}{\lambda g R} \tag{4.17}$$

and the resulting optimum unconditional mean squared error of $\bar{y}_{pss}$ is obtained as

$$MSE_{opt}(\bar{y}_{pss}) = \left(\frac{1-f}{n}\right)(1-\rho_0^2)\sum_{h=1}^{L}\omega_h S_{yh}^2 \tag{4.18}$$

where

$$\beta_0 = \frac{\sum_{h=1}^{L} \omega_h S_{yxh}}{\sum_{h=1}^{L} \omega_h S_{xh}^2} \ , \quad \rho_0 = \frac{\sum_{h=1}^{L} \omega_h S_{yxh}}{\sqrt{\left(\sum_{h=1}^{L} \omega_h S_{yh}^2\right)\left(\sum_{h=1}^{L} \omega_h S_{xh}^2\right)}} \tag{4.19}$$

Notice that equation (4.18) is the same as the unconditional variance of the usual combined post-stratified regression estimator, $\bar{y}_{psREG} = \bar{y}_{ps} - \beta_0(\bar{x}_{ps} - \bar{X})$. This implies that the efficiency of the proposed general family of estimators may not be improved beyond the efficiency of the customary combined regression-type estimator in post-stratified sampling. However, using equations (4.1) and (4.18), we observe that:

$$V(\bar{y}_{ps}) - MSE_{opt}(\bar{y}_{pss}) = \left(\frac{1-f}{n}\right)\rho_0^2 \sum_{h=1}^{L} \omega_h S_{yh}^2 > 0 \tag{4.20}$$

This shows that under optimum conditions, the proposed family of estimators is more efficient than the usual post-stratified estimator, $\bar{y}_{ps}$ in terms of having a smaller mean squared error.

Again, let $A_0 = \sqrt{\sum_{h=1}^{L} \omega_h S_{yh}^2}$ and $A_1 = \sqrt{\sum_{h=1}^{L} \omega_h S_{xh}^2}$. Then, using equations (4.2) and (4.18), we observe that

$$MSE(\bar{y}_{pss}(2)) - MSE_{opt}(\bar{y}_{pss}) = \left(\frac{1-f}{n}\right)(\rho_0 A_0 - RA_1)^2 > 0 \tag{4.21}$$

This shows that the optimum estimator in the proposed family of estimators is more efficient than the usual combined ratio-type estimator $\bar{y}_{psRC}$. Similarly, it could be shown that none of the special cases of the proposed family of estimators is more efficient than the optimum estimator in the proposed general family of estimators.

5. Empirical Illustration

We have applied the proposed general family of estimators on the data on the academic performance of 96 students of Statistics department, Federal University of Technology, Owerri, 2008/2009 academic session. (Source: Department of Statistics, Federal University of Technology, Owerri, Nigeria). Here, we used the cumulative grade point average (CGPA) as the study variate, and performance in a general (pretest) Statistics examination as the auxiliary variate. Stratification was carried out by gender, and we assumed, hypothetically, that the number of male and female students to be included in the sample might not be determined until after sample selection. Consequently, we first took a random sample of size $n = 20$, which showed the distribution of male and female students, after sample selection, as 8 males and 12 females. The data statistics, consisting mainly of population parameters, are shown in Table 1, while Table 2 shows the percentage relative efficiencies (PRE) of the proposed estimators over the customary post-stratified estimator $\bar{y}_{ps}$ of the population mean, $\bar{Y}$ in post-stratified sampling scheme.

Table 1. Data Statistics

POPULATION	MALES = STRATUM 1	FEMALES = STRATUM 2
$N = 96$	$N_1 = 72$	$N_2 = 24$
$n = 20$	$n_1 = 8$	$n_2 = 12$
$\bar{X} = 68.13$	$\bar{X}_1 = 68.11$	$\bar{X}_2 = 68.17$
$\bar{Y} = 2.44$	$\bar{Y}_1 = 2.44$	$\bar{Y}_2 = 2.46$
$S_x = 7.03$	$S_{x1} = 7.28$	$S_{x2} = 6.36$
$S_x^2 = 49.37$	$S_{x1}^2 = 52.97$	$S_{x2}^2 = 40.41$
$S_y = 0.57$	$S_{y1} = 0.60$	$S_{y2} = 0.50$
$S_y^2 = 0.33$	$S_{y1}^2 = 0.35$	$S_{y2}^2 = 0.25$
$S_{yx} = 3.26$	$S_{yx1} = 3.43$	$S_{yx2} = 2.75$
$\rho_{yx} = 0.82$	$\rho_{yx1} = 0.80$	$\rho_{yx2} = 0.90$
$\rho_{yx}^2 = 0.67$	$\rho_{yx1}^2 = 0.64$	$\rho_{yx2}^2 = 0.80$
$C_x = 0.10$	$C_{x1} = 0.11$	$C_{x2} = 0.09$
$C_y = 0.23$	$C_{y1} = 0.24$	$C_{y2} = 0.20$
$\beta_1(x) = -1.10$	$\beta_{11}(x) = -1.23$	$\beta_{12}(x) = -0.50$
$\beta_1(y) = -0.11$	$\beta_{11}(y) = -0.14$	$\beta_{12}(y) = 0.14$
$\beta_2(x) = 3.83$	$\beta_{21}(x) = 4.33$	$\beta_{22}(x) = 1.34$
$\beta_2(y) = 1.27$	$\beta_{21}(y) = 1.40$	$\beta_{22}(y) = 0.31$
$\gamma = 0.04$	$\gamma_1 = 0.05$	$\gamma_2 = 0.16$
–	$\omega_1 = 0.75$	$\omega_2 = 0.25$
–	$\omega_1^2 = 0.56$	$\omega_2^2 = 0.06$

$R = 0.035814,\quad \theta_1 = 0.035761,\quad \theta_2 = 0.033908,$

$\theta_3 = 0.035800,\quad \theta_4 = 0.022926,\quad \theta_5 = 0.035388,$

$\theta_6 = 0.032464,\quad \theta_7 = 0.039521,\quad \theta_8 = 0.034874,$

Table 2. PRE of some post-stratified combined estimators of $\bar{Y}$ over $\bar{y}_{ps}$

Estimator	Variance / MSE	PRE over $\bar{y}_{st}$
$\bar{y}_{pss}(1)$	0.012864	100
$\bar{y}_{pss}(2)$	0.006151	209.14
$\bar{y}_{pss}(3)$	0.006158	208.90
$\bar{y}_{pss}(4)$	0.006381	201.60
$\bar{y}_{pss}(5)$	0.006153	209.07
$\bar{y}_{pss}(6)$	0.007984	161.12
$\bar{y}_{pss}(7)$	0.006202	207.42
$\bar{y}_{pss}(8)$	0.024637	52.21
$\bar{y}_{pss}(9)$	0.024616	52.26
$\bar{y}_{pss}(10)$	0.024632	52.22
$\bar{y}_{pss}(11)$	0.019818	64.91
$\bar{y}_{pss}(12)$	0.023322	55.16
$\bar{y}_{pss}(13)$	0.026145	49.20
$\bar{y}_{pss}(14)$	0.024264	53.02
$\bar{y}_{pss}(15)$	0.024468	52.57
$\bar{y}_{pss}(16)$	0.023883	53.86
$\bar{y}_{pss}(opt)$	0.004422	290.91

From Table 2, we observe that not every estimator in the proposed general family of estimators performed better than the usual post-stratified combined estimator $\bar{y}_{ps}$. The table also confirms that the optimum estimator $\bar{y}_{pss}(opt)$ is the most efficient estimator in the proposed family of combined estimators of $\bar{Y}$ in post-stratified sampling scheme. Again, we observe that for the given data set, the combined ratio-type estimators, $\bar{y}_{ps}(i)$, $i = 2, 3, \cdots, 7$ performed better than the customary post-stratified estimator $\bar{y}_{ps}$ in terms of having smaller mean squared errors, while the combined product-type estimators, $\bar{y}_{ps}(i)$, $i = 8, 9, \cdots, 16$ did not perform better than $\bar{y}_{ps}$. This is expected since for the given data set, there is a strong positive correlation (0.82) between the study and auxiliary variables. The product-type estimators would perform better than $\bar{y}_{ps}$ and the ratio-type estimators when there is a strong negative correlation between the study and auxiliary variables.

6. Concluding Remark

We have proposed a general family of combined estimators of $\overline{Y}$, in post-stratified sampling (PSS) scheme, which is found, under some optimum conditions, to be as efficient as the post-stratified regression estimator $\overline{y}_{psREG}$, but more efficient, in terms of having a smaller mean squared error, than the usual post-stratified combined estimator, $\overline{y}_{ps}$ Properties of the proposed general family of estimators are obtained up to first order approximations and illustrated empirically.

REFERENCES

CHAUDHARY, M.K., SINGH, R., SHUKLA, R.K., KUMAR, M. and SMARANDACHE, F. (2009): A family of estimators for estimating population mean in stratified sampling under non-response. Pak. J. Stat. Oper. Res. V(1), 47-54.

COCHRAN, W.G. (1977): Sampling Techniques, John Wiley and Sons, New York.

KADILAR, C. and CINGI, H. (2003): Ratio Estimators in Stratified Random Sampling. Biometrical Journal 45(2), 218-225.

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

KOYUNCU, N. and KADILAR, C. (2009). Ratio and product estimators in stratified random sampling, Journal of Statistical Planning and Inference 139 (8), 2552-2558.

PANDY, B.N. and DUBEY, VYAS (1988): Modified product estimator using coefficient of variation of auxiliary variate, Assam Statistical Rev., 2(2), 64-66.

SEARLS, D.T. (1964): The utilization of known coefficient of variation in the estimation procedure. Journal of American Statistical Association, 59, 1125-1126.

SEARLS, D.T. and INTARAPANICH, P. (1990): A note on an estimator for the variance that utilizes the kurtosis. The American Statistician, 44(4), 295-296.

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

SINGH, G.N. (2003): On the improvement of product method of estimation in sample surveys. Jour. Ind. Soc. Agric. Statistics, 56(3), 267-275.

SINGH, H.P. AND KAKRAN, M.S. (1993): A Modified Ratio Estimator Using Known Coefficients of Kurtosis of an Auxiliary Character (unpublished).

SINGH, H.P. AND TAILOR, R. (2003): Use of known correlation coefficient in estimating the finite population mean. Statistics in Transition, 6(4), 555-560.

SINGH, J., PANDEY, B.N. and HIRANO, K. (1973): On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. Ann. Inst. Stat. Math., 25, 51-55.

SISODIA, B.V.S. and DWIVEDI, V.K. (1981): A Modified Ratio Estimator Using Coefficient of Variation of Auxiliary Variable. Journal of Indian Society Agricultural Statistics 33, 13-18.

STEPHAN, F. (1945): The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate. Ann. Math. Stat., 16, 50-61.

UPADHYAYA, L.N. and SINGH, H.P. (1999): Use of Transformed Auxiliary Variable in Estimating the Finite Population Mean. Biometrical Journal 41(5), 627-636.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 79—94

NONRESPONSE BIAS IN THE SURVEY OF YOUTH UNDERSTANDIG OF SCIENCE AND TECHNOLOGY IN BOGOTÁ

Edgar Mauricio Bueno Castellanos[1]

ABSTRACT

The Colombian Observatory of Science and Technology -OCyT- developed, in 2009, a survey about understanding of Science and Technology in students of high school in Bogotá, Colombia. The sampling design was stratified according to the nature of school (public or private). Two sources of unit nonresponse were detected. The first one corresponds to schools that did not allowed to collect information. The second source corresponds to students who did not assist during the days when survey was applied. Estimates were obtained through two different approaches. Results obtained in both cases do not show visible differences when estimating ratios; even though, some great differences were observed when estimating totals. Results obtained using the second approach are believed to be more reliable because of the methodology used to handle item nonresponse.

Key words: Sampling design; nonresponse bias; calibration.

1. Introduction

In 2009, the Colombian Observatory of Science and Technology -OCyT- developed the *Survey of Youth Understanding of Science and Technology in Bogotá*, which inquires about topics related to understanding about scientist, engineers and benefits and risks of science and technology. Results and analysis of the survey are presented by Daza et. al (2011).

As expected, on the data collecting process, there were students who were not possible to contact (unit nonresponse) and others that did not fulfill some of the questions in the questionnaire (item nonresponse). As a consequence, arises the need to use methodologies that allows to obtain estimations taking into account the presence of nonresponse.

[1] Colombian Observatory of Science and Technology –OCyT-. Bogotá Colombia.
E-mail: embuenoc@ocyt.org.co.

Initially, item nonresponse was considered as a new category and the unit nonresponse was handled by conforming Response Homogeneity Groups (Särndal, Swensson and Wretman, 1992). After that, it was proposed to obtain estimations through other methodology: to impute missing values corresponding to item nonresponse and to use the calibration estimator for unit nonresponse.

The second section of this document describes the methodology used for design and development of the survey. The third section describes the causes of nonresponse in the survey and the two methodologies proposed to handle it. In order to compare these methodologies, a Monte Carlo simulation was carried out, its results are described in the fourth section. This simulation allowed to see the behavior of estimators under different cases. In the last section conclusions and suggestions are presented based on the experience achieved through the survey.

2. Methodology

The survey target population was conformed to students of the last two years of high school of all the schools in Bogotá, Colombia. The sampling frame used to identify the schools was the educational establishment register from the Secretaría de Educación de Bogotá (bureau of education), which includes, besides identification and contact variables, the nature of school (public or private) and information about the number of students registered in year 2008 in every grade. The register includes all the educational establishments in the city, therefore, it was necessary to eliminate the institutions that does not offer the grades defined for the study and those that offer them but have an approach on adult education. Finally, it was obtained a sampling frame with 1073 schools, 715 of these are private and reported 59984 students in 2008, the remaining 358 are public and reported 112830 students in the same year.

Once conformed the final frame, the sample was drawn. The design was a Stratified one-stage cluster sampling.

- The nature of school was used as stratification variable.
- In each stratum a sample of schools was drawn using a probability proportional to size -*pps*- design. The size variable used to assign probabilities to schools was the number of students reported for 2008 according to the frame, incremented in one unity.
- The questionnaire was applied to every student in the last two years in selected schools.

A sample of 31 private and 16 public schools was drawn (ordered sample). One public and two private schools were reselected in the sample, obtaining a set-sample of 29 private and 15 public schools, which have, respectively, 6231 and 7498 students in 2008. Throughout this document and unless otherwise is specified, *sample* will make reference to the ordered sample.

When the data collection stage ended, paper questionnaires were transcribed, conforming a data set that was validated and then estimations were carried out. In a first moment, it was planned to obtain estimations using the estimator proposed by Hansen and Hurwitz (1943), also known as with-replacement sampling estimator –pwr estimator-.

This estimator could not be used because it was not possible to obtain information from all individuals expected in the sample. For this reason it was necessary to identify other alternatives to obtain estimations in the presence of nonresponse. The next section describes the alternatives used for the survey.

3. Dealing with nonresponse

As usual when developing a survey, in the understanding survey both types of nonresponse arises: item nonresponse and unit nonresponse.

Two unit nonresponse sources were identified. The first one, due to directives that deny data collection: the survey was implemented in the 16 public schools, but only in 13 out of 31 private schools drawn in the sample; this case will be referred as *cluster nonresponse*. The second source corresponds to students who belong to schools in which access was allowed but did not assist during the days when survey was applied, this case will be referred as *element nonresponse.*

Given that all the questions in the survey are categorical, in a first moment estimations were obtained by considering item nonresponse as a new category for every variable. Unit nonresponse was handled by modifying expected sample sizes by those observed. This approach will be referred as *Approach 1*.

Later, it was decided to obtain new estimations by the use of methods allowing to control the nonresponse effects: the *nearest neighbor* methodology was used to impute values belonging to item nonresponse and the calibration estimator was used to handling unit nonresponse. This approach will be referred as *Approach 2* and is described in Section 3.2.

a. Approach 1

In this approach the nonresponse was handled according to:

Item nonresponse: Missing values due to item nonresponse was considered as a new category. By doing this, a rectangular data set is obtained, in which missing values are replaced by a code representing its absence. One advantage of the methodology is that allows to obtain a completely rectangular data set allowing to make cross tabulation of variables in survey; on the other hand, some disadvantages are the arising of meaningless cross-classified cells and that nothing is done in order to control the bias due to nonresponse.

Unit nonresponse: Element nonresponse was handled by assuming that, in every school, students who participated in the survey conform a simple random sample of students. The bias generated by this assumption is expected to be small given that the nonresponse rates within the schools that participate in the survey were low.

Cluster nonresponse was handled by assuming a *response homogeneity group* model with groups given by the nature of school. This means that is assumed that the response probability, θ_i, in each group of schools (public or private) is fixed and estimated by $\hat{\theta}_i = m_h^*/m_h$. In this case, *pwr* estimator takes the form

$$\hat{t}_y^* = \sum_h \left(\frac{1}{m_h^*}\sum_{r_h}\frac{\hat{t}_{yi}}{p_i}\right), \quad h = 1{,}2;\ \text{with } \hat{t}_{yi} = \frac{N_i}{n_i}\sum_{r_i} y_k \qquad (1)$$

where

- $\hat{t}_{yi}$ is the estimation of the total of variable y in the ith school,
- N_i is the number of students in the ith school. This number was recorded for every school in the response set,
- n_i is the number of students who answered the questionnaire in the ith school,
- r_i is the response set of students belonging to the ith school,
- r_h is the response set of schools in the stratum h,
- m_h is the number of selected schools in stratum h,
- m_h^* is the number of schools in the response set in stratum h,
- y_k is the value of y for the kth,
- p_i is the selection probability of ith school.

At first glance, the estimator in (1) does not control the bias or the variance increments that may be generated as a consequence of the nonresponse. Even so, this estimator satisfies the desirable property of reproducing totals for the size variable used to obtain the selection probabilities of individuals:

For the case of element sampling from a population U, which counts with values of y for every element in the sample s of size m, let $\mathrm{t_x} = \sum_{\mathrm{U}} \mathrm{x_k}$ the total of size variable x and $\mathrm{x_k}$ the value of x associated to the kth individual. Selection probability for kth individual is defined as $\mathrm{p_k} = \mathrm{x_k}/\mathrm{t_x}$. When applying the pwr estimator to values of x in the sample s, we obtain

$$\hat{t}_x = \frac{1}{m}\sum\nolimits_s \frac{x_k}{p_k} = t_x$$

This result, obtained for element sampling, works also for the modification proposed for handling nonresponse, equation (1),

$$\hat{t}_{h,x}^* = \frac{1}{m_h^*}\sum\nolimits_{r_h} \frac{\hat{t}_{x_i}}{p_i} = t_{h,x}$$

This property indicates that, if y = αx, the estimator given in (1) will obtain perfect estimations for $\mathrm{t_y}$ for every sample, no matters the nonresponse, this property resembles the calibration estimator. It is clear that is impossible that the proportionality be satisfied in practice, even so, the property suggest that while there exists a high correlation between y and x, both, variance and bias due to nonresponse will be small. For the understanding survey it is required that totals in schools (t_{y_i}) to be proportional to the number of students reported for 2008 (t_{x_i}).

b. Approach 2

The nonresponse was handled according to:

Item nonresponse: Missing values due to item nonresponse were imputed using the *nearest neighbor* methodology: For every variable in the questionnaire, y_j, a set of G variables W, which is expected to be related to y_j, is identified and then sorted according to *explanatory power* expected with y_j, this is, the first variable in W will be the one that explain the most of y_j, the second will be the one following this rule, and so on. It is important to clarify that every variable in the study was qualitative and that the choice of the variables in W and its order were due to subjective criteria.

Individuals in the data set were divided in two groups according to the values for y_j: the response set r_j and the nonresponse set $r_u - r_j$, where $r_u = \bigcup_{j=1}^{q} r_j$ is the set of individuals having information for at least one of the q variables in the questionnaire. The value y_{kj} (in $r_u - r_j$) is imputed as follows:

- The matrix Z is created from W as: $z_{lg} = \begin{cases} 1, & \text{if } w_{kg} = w_{lg} \\ 0, & \text{if } w_{kg} \neq w_{lg} \end{cases}$, $g = 1,2,\cdots,G$; $l = 1,2,\cdots,n_{rj}$, where n_{rj} is the number of individuals in r_j.
- For every individual, l, in Z we calculate $D_k(l) = 2^{G-1}z_{l1} + 2^{G-2}z_{l2} + \cdots + 2^0 z_{lG}$.
- Individual that maximizes $D_k(l)$ is identified and its value y_{lj} is assigned to y_{kj}.
- When there are ties, y_{kj} is obtained as the mode of the y values associated to those individuals that maximizes $D_k(l)$.
- If there is not a unique mode, a random value of y is chosen from the set of modes.

It is clear that the distance metric $D_k(l)$ is such that matching in z_{l1} dominates matching in the remaining variables z_{l2} to z_{lG}; if z_{l1} does not match, z_{l2} dominates matching in the remaining z_{l3} to z_{lG}; and so on. This situation was decided in order of reducing the burden of calculations that would imply the assignation of different weights to every variable in W for every variable in the questionnaire.

Unit nonresponse: Element nonresponse was handled in the same fashion that in *Approach 1:* it was assumed that, in every school, students who participated in the survey conform a simple random sample from the total of students.

Särndal and Lundström (2005) proposed the calibration estimator for the Horvitz and Thompson estimator (1952) at the level of individuals. For cluster nonresponse a variation of this estimator was used. Due to the absence of auxiliary information at the level of students, calibration was carried out at the level of schools using one quantitative and two categorical variables for classification.

Quantitative variable, $\mathrm{t}_{\mathrm{x}_\mathrm{i}}$, is the total of students in the ith school during 2008. The first classification variable, γ, is the same variable used for stratification, the nature of school (public or private), $\gamma_\mathrm{i} = (\gamma_{1\mathrm{i}}, \gamma_{2\mathrm{i}})'$, where

$$\gamma_{1i} = \begin{cases} 1, & \text{if school } i \text{ is private} \\ 0, & \text{if school } i \text{ is public} \end{cases} \text{ and } \gamma_{2i} = \begin{cases} 1, & \text{if school } i \text{ is public} \\ 0, & \text{if school } i \text{ is private} \end{cases}$$

The second classification variable, δ, is an indicator of the size of school, defined as $\delta_\mathrm{i} = (\delta_{1\mathrm{i}}, \delta_{2\mathrm{i}}, \delta_{3\mathrm{i}})'$, where

$$\delta_{1i} = \begin{cases} 1, \text{ if } t_{x_i} \leq 100 \\ 0, \text{ otherwise} \end{cases}, \delta_{2i} = \begin{cases} 1, \text{ if } 100 < t_{x_i} \leq 400 \\ 0, \text{ otherwise} \end{cases} \text{ and } \delta_{3i} = \begin{cases} 1, \text{ if } t_{x_i} > 400 \\ 0, \text{ otherwise} \end{cases}$$

The auxiliary vector associated to the ith school, t_{x_i}, is conformed as

$$\boldsymbol{t}_{x_i} = \left(\gamma_{1i} t_{x_i}, \gamma_{2i} t_{x_i}, \delta_{1i} t_{x_i}, \delta_{2i} t_{x_i}\right)'$$

and the input vector required is the total of students in every group in 2008:

$$\boldsymbol{X} = \left(\sum\nolimits_U \gamma_{1i} t_{xi}, \sum\nolimits_U \gamma_{2i} t_{xi}, \sum\nolimits_U \delta_{1i} t_{xi}, \sum\nolimits_U \delta_{2i} t_{xi}\right)'$$

δ_3 is not included in order to avoid singularities in the matrix to be inverted to obtain the calibrated selection probabilities.

Once defined the auxiliary vector and the input vector, the calibrated selection probabilities, w_i, are calculated as

$$w_i = p_i / v_i \text{ with } v_i = 1 + \boldsymbol{\lambda}_r' \mathbf{t}_{x_i} \text{ and}$$

$$\lambda_r' = \left(X - \sum\nolimits_h \left(\frac{1}{m_h} \sum\nolimits_{r_h} \frac{t_{x_i}}{p_i} \right) \right) \left(\sum\nolimits_h \left(\frac{1}{m_h} \sum\nolimits_{r_h} \frac{t_{x_i} t_{x_i}'}{p_i} \right) \right)^{-1}$$

and then, the total of y is estimated as

$$\hat{t}_y^c = \sum_h \left(\frac{1}{m_h}\sum_{r_h}\frac{\hat{t}_{y_i}}{w_i}\right),\ h = 1{,}2;\ \text{with } \hat{t}_{y_i} = \frac{N_i}{n_i}\sum_{r_i} y_k \qquad (2)$$

A comparison between the estimators $\hat{t}_y^*$ and $\hat{t}_y^c$ is presented in the next section.

4. A Monte Carlo simulation study

In order to compare the bias and variance of $\hat{t}_y^*$ and $\hat{t}_y^c$, defined in equations (1) and (2), respectively, a Monte Carlo simulation study was carried out. This process took into account only cluster nonresponse; element nonresponse and item nonresponse were ignored.

A population of $N = 148245$ individuals in 1073 schools was created. The number of schools was fixed to match the number of schools in the sampling frame, while the number of individuals was fixed to match the estimated number of students according to *Approach 1*.

Three auxiliary variables at the level of schools (x_1, x_2 and x_3), one *exogenous* variable (z) and three study variables at the level of students (y_1, y_2 and y_3) were generated as follows:

x_{i1}: Number of students in the ith school according to the sampling frame,

x_{2i}: The nature of ith school (public or private), $x_2 = \begin{cases} 1, & \text{if school is private} \\ 0, & \text{if school is public} \end{cases}.$

This variable is used also for conforming strata.

x_{3i}: The size of ith school, $x_3 = \begin{cases} 1, \text{if } x_1 \leq 100 \\ 2, \text{if } 100 < x_1 \leq 400. \\ 3, \text{if } x_1 > 400 \end{cases}$

z: A dichotomous exogenous variable related to the nature of school. By exogenous variable I mean a variable that is completely unknown in the survey: it is not an auxiliary variable known beforehand, and also is not measured in the questionnaire as a study variable:

$$P(z = 1|x_2 = 1) = 0.7 \qquad \text{and} \quad P(z = 1|x_2 = 0) = 0.4$$

y_1: A dichotomous variable that takes value 1 with different probabilities according to the nature of school:

$$P(y_1 = 1|x_2 = 1) = 0.8 \text{ and} \quad P(y_1 = 1|x_2 = 0) = 0.5$$

y_2: A dichotomous variable that takes value 1 depending on strata (x_2), and the value of z:

$P(y_2 = 1|x_2 = 1, z = 1) = 0.9$, $P(y_2 = 1|x_2 = 1, z = 0) = 0.8$, $P(y_2 =$ *1|x2=0,z=1=0.5*
and $P(y_2 = 1|x_2 = 0, z = 0) = 0.2$

y_3: A dichotomous variable that takes value 1 depending only on the value of z:
$P(y_3 = 1|z = 1) = 0.9$ and $P(y_3 = 1|z = 0) = 0.45$

It is clarified that the auxiliary variables x_1, x_2 and x_3 are present at the level of schools, while the study variables y_1, y_2 and y_3 are present at the level of students and they are related to the auxiliary variables through its totals within schools.

The idea behind the setup for the study variables and the response distribution (step 4 of the simulation process) will be described below.

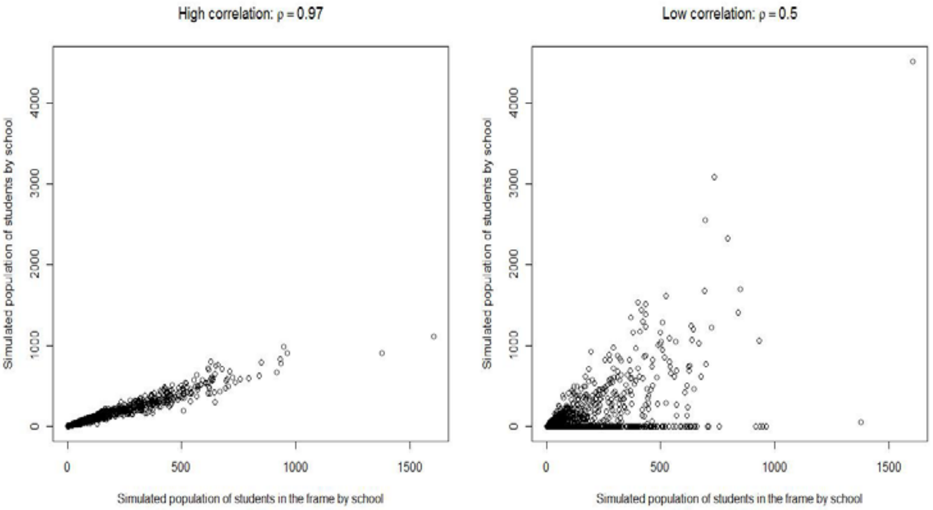


Figure 1. Simulated individuals by school: x_{1i} vs. N_i

The number of individuals in the ith school, N_i, was defined in order that the correlation coefficient between the size in the frame, x_{1i}, and N_i was (approximately) equal to ρ_0:

$$N_i = \hat{B}_0 x_{1i} + e_i \ ,$$

where $e_i = \varepsilon_i x_{1i}$ and ε_i is an observation of a random variable $N(0, a)$, with $a > 0$ chosen properly and $\hat{B}_0 = 0.32$ is the slope of the regression line of x_1 on N_i.

Two *correlation levels* between N_i and x_1 were generated: high ($\rho(N_i, x_{1i}) \approx 0.97$) and low ($\rho(N_i, x_{1i}) \approx 0.50$). The left panel of Figure 1 shows the scatter plot between the number of students for school according to the sampling frame and the number of students observed for the case $\rho(N_i, x_{1i}) \approx 0.97$. The right panel shows the case $\rho(N_i, x_{1i}) \approx 0.50$. It is clarified that the minimum value for the simulated populations is equal to 2.

With each of these populations the following process was carried out:

1. A stratified (with replacement) *pps* of 31 private and 16 public schools was drawn. The selection probability for the ith school was defined as $p_i = x_{1i}/\sum_{U_h} x_{1i}$. All students in selected schools were selected.
2. The totals by stratum of y_1, y_2, y_3 and the population size, N, from the full sample using the pwr estimator were estimated: $\hat{t}_{yh}^{(1)} = \frac{1}{m_h}\sum_{s_h}\frac{t_{y_i}}{p_i}$.
3. The totals by stratum of y_1, y_2, y_3 and the population size, N, using the calibration estimator including x_2 and x_3 as classification variables and x_1 as quantitative were estimated: $\hat{t}_{yh}^{(2)} = \frac{1}{m_h}\sum_{s_h}\frac{t_{y_i}}{w_i}$.
4. A school *response distribution* was generated by a fixed response probability θ_i depending on the strata and the school total of z. θ_i was defined in order that the response probability in Stratum 1 and Stratum 2, was 0.42 and 0.94, respectively.
5. Once defined the response set, totals for the four already mentioned variables were estimated using the estimator (1): $\hat{t}_{yh}^{(3)} = \frac{1}{m_h^*}\sum_{r_h}\frac{t_{y_i}}{p_i}$.
6. Totals of the four variables were estimated using the calibration estimator: $\hat{t}_{yh}^{(4)} = \frac{1}{m_h}\sum_{r_h}\frac{t_{y_i}}{w_i}$.
7. A stratified simple random sample -*srs*- of the full population of schools was drawn. This procedure was carried out in order to compare a design that includes auxiliary information (*pps*) against one that does not include it (*srs*). The number of schools, (m_h), was chosen with the goal that the number of individuals expected under the *srs* sample was (approximately) equal to the number of individuals expected under the *pps* sample.
8. Totals of y_1, y_2, y_3 and the population size, N, were estimated by using the Horvitz-Thompson estimator (also known as π-estimator) (1952): $\hat{t}_{yh}^{(5)} = \frac{M_h}{m_h}\sum_{s_h} t_{yi}$, with M_h the number of schools in stratum h.
9. In addition, with every estimator, the ratio $R_1 = t_{y1}/N$ was estimated. Also $R_2 = t_{y2}/N$ and $R_3 = t_{y3}/N$ were estimated. The results obtained are similar to those obtained for R_1.

The procedure described in numerals 1 to 9 is repeated $I = 10000$ times. Every time the estimations obtained through the five estimators are recorded. The (simulated) expectation of each estimator is obtained as

$$E_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{1}{I}\sum_{i=1}^{I}\hat{t}_{y_i}^{(j)}$$

and the (simulated) variance is obtained as

$$V_{SIM}\left[\hat{t}_y^{(j)}\right] = S^2_{\hat{t}_y^{(j)}} = \frac{1}{I-1}\sum_{i=1}^{I}\left(\hat{t}_{yi}^{(j)} - E_{SIM}\left[\hat{t}_y^{(j)}\right]\right)^2.$$

Table 1 shows the parameters to estimate: totals of variables y_1, y_2 and y_3, total of individuals in the population and the ratio R_1.

Table 1. Population totals and ratios

	Total	Private schools	Official schools
N	148245	54417	93828
t_{y1}	90524	43664	46860
t_{y2}	77488	47452	30036
t_{y3}	100553	41370	59183
R_1	0,61	0,80	0,50

Tables 2 and 3 shows (simulated) relative bias and (simulated) coefficient of variation for cases $\rho(N_i, x_{1i}) \approx 0.97$ and 0.50, respectively. The (simulated) relative bias, RB_{SIM}, of the jth estimator for total t_y is calculated as

$$RB_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{E_{SIM}\left[\hat{t}_y^{(j)}\right] - t_y}{t_y}$$

and the (simulated) coefficient of variation, CV_{SIM}, of the jth estimator for total t_y is calculated as

$$CV_{SIM}\left[\hat{t}_y^{(j)}\right] = \frac{\sqrt{V_{SIM}\left[\hat{t}_y^{(j)}\right]}}{E_{SIM}\left[\hat{t}_y^{(j)}\right]}$$

A few words on the response distribution, the variables y_1, y_2 and y_3, and the auxiliary vector $\boldsymbol{x}_k$: According to Särndal and Lundström (2005) there is a triple $(\theta_k, y_k, \boldsymbol{x}_k)$ associated to every individual in the population. It is clear that, by construction, θ_k depends partially on the known $\boldsymbol{x}_k$ and partially on the unknown z. Given that y_{1k} depends only on x_{2k} not on z, in this case the nonresponse is completely explained by the auxiliary vector $\boldsymbol{x}_k$, so this case can be considered as Missing at Random -MAR-. y_{2k} depends on both x_{2k} and z, so the nonresponse is partially explained by x_{2k}. Finally, y_{3k} depends only on the unknown variable z, so the auxiliary vector is unable to explain the nonresponse distribution, at least directly.

Table 2 shows the results for the case in which the correlation between the number of individuals by school in the frame and the number of individuals observed by school is 0.97. About the bias, Table 2 suggests the following results:

- It is known that $\hat{t}_1$ and $\hat{t}_5$ are unbiased in total estimation and $\hat{t}_2$ is *asymptotically unbiased*. The simulation allows to see these facts. The bias of $\hat{t}_2$ and $\hat{t}_3$ is also small.
- In the stratum of high nonresponse (stratum 1) the bias for the calibration estimator under nonresponse ($\hat{t}_4$) although small, is notably greater than the bias for the *pwr* estimator under nonresponse ($\hat{t}_3$). Meanwhile, in the stratum 2, there is a reverse situation: bias of $\hat{t}_4$ is smaller than the bias of $\hat{t}_3$.
- The bias of the five estimators for the ratio R_1 are small.

Table 2. Simulated relative bias and simulated coefficient of variation (as a percentage) of five estimators for the case $\rho(N_i, x_{1i}) \approx 0.97$.

		Relative bias					Coefficient of variation				
Strata	Parameter	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$
Stratum 1	*N*	0,02	0,00	-0,03	-0,21	-0,14	3,43	3,58	5,53	6,19	12,74
	t_{y1}	0,02	-0,01	-0,05	-0,21	-0,13	3,62	3,78	5,86	6,42	12,81
	t_{y2}	0,04	0,01	0,00	-0,14	-0,14	3,53	3,68	5,67	6,31	12,82
	t_{y3}	0,01	0,00	-0,07	-0,27	-0,13	3,67	3,82	5,90	6,48	12,71
	R_1	0,00	-0,01	-0,02	0,00	0,00	1,01	1,05	1,64	1,61	0,67
Stratum 2	*N*	-0,05	-0,10	-0,06	0,01	-0,13	5,18	5,27	5,37	5,50	14,67
	t_{y1}	-0,04	-0,10	-0,05	0,02	-0,12	5,57	5,66	5,72	5,84	14,68
	t_{y2}	-0,05	-0,13	-0,05	0,04	-0,15	5,80	5,90	6,00	6,18	14,82
	t_{y3}	-0,05	-0,10	-0,06	0,02	-0,15	5,46	5,55	5,62	5,76	14,72
	R_1	0,01	0,00	0,01	0,01	0,02	1,72	1,74	1,76	1,82	1,18

Some comments on the coefficients of variation in Table 2:

- The variance of the estimators for totals is highly reduced when including auxiliary information: CV of $\hat{t}_5$ are clearly greater than those of the other four estimators.
- The CV of the calibration estimator are slightly greater than those for the *pwr* estimator: the CV of $\hat{t}_2$ is slightly greater than the CV of $\hat{t}_1$, for the case of full response; and, the CV of $\hat{t}_4$ is slightly greater than the CV of $\hat{t}_3$, for the case of nonresponse.
- The CV of the estimators that works in the presence of nonresponse in the stratum 1 are clearly higher than those of the estimators that works under full response; on the other hand, in stratum 2 this difference is small. This is a consequence of the response probabilities in each stratum.
- When estimating the ratio R_1, the results differs from those for totals: in this case the smallest CV corresponds to the strategy (*srs*, π-estimator), a strategy that does not includes auxiliary information in the design or the estimation stage. This is due to the fact that when estimating totals, the size variable used in $\hat{t}_1$ to $\hat{t}_4$ is more or less related to the totals within schools, so reducing the variance; whereas, when estimating a ratio, the size variable does not explain the variation in the variable of interest, and can even cause a loss of efficiency with regard to a strategy that does not include auxiliary information.

Table 3. Simulated relative bias and simulated coefficient of variation (as a percentage) of five estimators for the case $\rho(N_i, x_{1i}) \approx 0.50$.

		Relative bias					Coefficient of variation				
Strata	Parameter	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$
Stratum 1	N	-0,12	0,49	-17,53	-20,92	0,18	23,73	24,53	44,04	47,79	21,52
	t_{y1}	-0,12	0,48	-17,51	-20,91	0,18	23,78	24,58	44,12	47,87	21,57
	t_{y2}	-0,12	0,48	-17,55	-20,90	0,18	23,70	24,50	44,01	47,76	21,54
	t_{y3}	-0,13	0,48	-17,63	-21,02	0,18	23,72	24,52	44,07	47,78	21,46
	R_1	-0,01	-0,01	0,02	0,04	0,00	1,05	1,08	2,17	2,16	0,65
Stratum 2	N	0,34	0,24	1,34	1,43	0,30	31,23	32,28	32,11	33,97	32,31
	t_{y1}	0,35	0,24	1,35	1,44	0,30	31,11	32,15	31,99	33,83	32,21
	t_{y2}	0,33	0,21	1,33	1,44	0,27	31,21	32,25	32,10	34,00	32,46
	t_{y3}	0,34	0,25	1,35	1,45	0,30	31,27	32,32	32,15	34,04	32,40
	R_1	0,06	0,06	0,07	0,07	0,03	1,85	1,88	1,90	1,97	1,16

Table 3 show the results for the case when correlation between the number of individuals by school in the frame and the number of individuals observed is 0.50. In reference to the bias, it is observed that, in this case there is a strong bias in the two estimators that works under nonresponse ($\hat{t}_3$ and $\hat{t}_4$) in the stratum of high nonresponse (stratum 1). This is due to the fact that neither the design nor the auxiliary variables were (enough) correlated to the study variables, so they were unable to control the bias generated by the nonresponse. Even so, the bias for R_1 is still small.

Some comments on the variances in Table 3:

- The increment in the variance of all the estimators for totals when comparing with those in Table 2 is clear.
- The variance of the two estimators that works with auxiliary information and counts with a full sample, $\hat{t}_1$ and $\hat{t}_2$, is similar. This variance is, indeed, similar with that obtained for the estimator $\hat{t}_5$. This result suggest that in this case, the gain obtained by the size variable x_1 (at the design stage) and by the auxiliary vector (at the estimation stage) is *negligible* as a consequence of the low correlation between these and the study variables.
- The effect of nonresponse in the variance is visibly greater in the first stratum: compare $\hat{t}_3$ and $\hat{t}_4$ with $\hat{t}_1$ and $\hat{t}_2$, respectively. This result is a consequence of the low response probability in stratum 1 and the low correlation between the auxiliary variables and the response distribution.
- The variance of the estimators when estimating a ratio does not show an increment when comparing with the results in Table 2; once more, $\hat{t}_5$ is the estimator with the smaller variance.
- It is interesting that although y_1, y_2 and y_3 were generated under different conditions, the results are not affected by this fact. The explanation is that although z is not included directly in the survey, it is explained indirectly by the size of the schools.

In my opinion, the most interesting result that is obtained from the simulation study already described is that, although the nonresponse have visible effects in bias and variance of the estimators when estimating totals (effects that becomes even bigger when auxiliary information is not highly correlated with the survey variables), this weakness does not seem to be *inherited* when estimating a ratio: estimations are still reliable, no matter the presence or absence of *powerful* auxiliary information or the patterns imposed on the nonresponse distribution.

This is important for the understanding survey given that ratios (proportions) are the most important parameters to be estimated in it.

Two aspects were taken into account in order to make a choice on one of the approaches: handling of unit nonresponse (in terms of the bias and the variance in the simulation study) and the handling of item nonresponse. Finally, it was decided to choose the *Approach 2.* The reasons to make this choice are:

- With regard to the bias, both estimators have a similar behavior: when estimating totals the bias is small if there is a high correlation between the expected and observed number of students; on the other hand, the bias is equally great for both estimators when the correlation is low. When estimating a ratio (proportion) the bias is negligible for both estimators.
- With regard to the variance, again both estimators have a similar behavior: a small CV when there is a high correlation between the expected and observed number of students; a greater CV when this correlation is low and an even greater CV when there is a low nonresponse.
- With regard to the handling of item nonresponse, it is strongly believed that the methodology used in Approach 2 overcomes to that used in Approach 1, where little is done, while in Approach 2 the relation between study variables is used to impute the missing values.
- Handling of item nonresponse in *Approach 1* creates a new category for every variable, this category does not correspond to the original questionnaire, it is a consequence of an unlucky –although common- event: partially incomplete information on the responses of an individual. This new category is not a problem in *Approach 2,* in which final tables keeps the structure expected at the moment of the questionnaire design. This fact facilitates the results interpretation.

5. Conclusions

- Although it is clear that nonresponse is an undesirable, but almost inevitable event in any survey, in the understanding survey developed by the OCyT there was the fortune of identifying a variable associated with its occurrence: the nature of school. Given that this variable was considered since the design stage as a stratification variable, the effect that nonresponse could have on bias and variance was reduced.
- Estimations for totals yields clear differences between both approaches, moreover, *Approach 1* yields lower estimates than *Approach 2*. Even so, these

differences are reduced when estimating proportions. This result is consistent with the simulations carried out, where proportions were less sensitive to the estimator. Furthermore, estimations for ratios happened to be *insensitive* to design, estimator or response distribution.

- Although there was not auxiliary information available at the level of individuals, available variables at the level of school (nature of school, number of student in the last year) allowed to build estimators that reduced the effect of nonresponse on the final estimations.
- The *pwr* estimator for a probability proportional to size *-pps-* design resembles the calibration estimator in the sense that it *reproduces exactly* the total of the size variable. A consequence of this property is that, when the selection probabilities are highly correlated to the study variables, a reduction in bias and variance generated by nonresponse is obtained.
- The results from the simulation study shows that both estimators have similar behaviors and that achieves satisfactorily the goal of controlling bias and variance generated by nonresponse when there is a high correlation between the expected and the observed number of students in schools.
- Simulations shown in section 4 allow to see the behavior of both proposed estimators in a set of cases. These cases were proposed in the context of the understanding survey and they were useful to make decisions on the estimators. Even so, it is important to recall that results must not be generalized, since they depends on the simulated population, considered designs, auxiliary variables included, response distribution, and so on.

REFERENCES

DAZA, S. ED, (2011). Entre datos y relatos: percepciones de jóvenes escolarizados sobre la Ciencia y la Tecnología. Observatorio Colombiano de Ciencia y Tecnología, Bogotá.

HANSEN, M.H., and HURWITZ W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics 41, 517-529.

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

R Development Core Team, (2011). A Language and environment for Statistical Computing, http_//www.r-project.org.

SÄRNDAL, C.E. and SWENSSON, B. and WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer.

SÄRNDAL, C.E. and LUNDSTRÖM, S., (2005). Estimation in Surveys with Nonresponse. Wiley.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 95—106

CUMULATIVE SUM CONTROL CHARTS FOR TRUNCATED NORMAL DISTRIBUTION UNDER MEASUREMENT ERROR

R. Sankle[1], J.R. Singh[2], I.K. Mangal[3]

ABSTRACT

In the present paper Cumulative Sum Control Chart (CSCC) for the truncated normal distribution under measurement error (r) is discussed. The sensitivity of the parameters of the V-Mask and the Average Run Length (ARL) is studied through numerical evaluation for different values of r.

Key words: Truncated Normal Distribution, Measurement Error, ARL and CSCC.

1. Introduction

The purpose of this paper is to show how truncated data affect the field performance of controlling manufacturing process. Failure to account for truncation can lead to biased inferences. Cumulative Sum (CUSUM) Control Charts are widely used instead of standard Shewhart charts when detection of small changes in a process parameter is important. Patel and Gajjar (1994) provide references on research performed on CUSUM charts. Errors of measurements are the differences between observed values recorded under “identical” conditions and some fixed true values. Hunter (1986) showed that finally all measurement should be traceable to nation standard the most important area for studying the effect of measurement error and misclassification in the sampling inspection plan and control charts.

Most of the standard statistical tests are derived on the assumption that the sample is combined from a “complete” population. But this assumption appears to be an unrealistic one. Truncations in the parents distributions at one or both the

[1] School of Studies in Statistics, Vikram University, Ujjain (M.P.), India. E-mail: rajshreesankle@gmail.com.

[2] School of Studies in Statistics, Vikram University, Ujjain (M.P.), India. E-mail: jrsinghstat@gmail.com.

[3] Madhav Science College, Ujjain (M.P.), India.

tails may have considerable effect on the CUCUM control charts. Woodroofe (1985) gave a comprehensive discussion of the exact estimation theory as well as consistency and asymptotic normality of the product limit estimator. Woodroofe also surveyed the history and application of this theory within astronomy but did not put it into the context of survival analysis. This was done by Wang et al. (1986) and by Keiding and Gill (1987) who went on to show how the exact and asymptotic properties of the estimators may be obtained as corollaries from statistical theory of counting processes and Markov processes, and Nelson (1990) stresses the distinction between truncation and censoring. Chang (1990) and Schneider (1986) contain some basic results concerning the properties of the truncated normal distribution.

Johnson and Leone (1962) considered mathematical procedure for construction of CUSUM Control Chart for Poisson variable using the relationship between Wald's Sequential Probability Ratio Test (SPRT) and CUSUM on the assumption that the probability of the second kind of error is small. They used this relationship to construct CUSUM charts for the mean and standard deviation of a normal distribution. Yeh et al. (2004) gave a unified CUSUM charts for monitoring process mean and variability. Cox (2009) studied control charts for monitoring observations from a truncated normal distribution. Nenes and Tagaras (2010) evaluated the properties of a CUSUM chart designed for monitoring the process mean in short production runs. Grigg and Spiegelhalter (2008) developed an empirical approximation to the null steady-state distribution of CUSUM statistics. Ryu et al. (2010) used ARL-based performance measure and proposed a method to optimally design a CUSUM chart based on expected weighted run length.

In this paper we have constructed CUSUM chart for mean under truncated normal distribution and measurement error. For different truncation points and different sizes of measurement error tables have been prepared for the average run length, lead distance and the angle of mask.

2. Determination of mask parameters under measurement error

Assuming that the true measurement x and the random error of measurement e are additive, we can write the observed measurement X as

$$X = x + e \tag{2.1}$$

where x and e are independent. The constants θ (unknown) and σ_p (known) are the mean and the standard deviation of the true quality measurement x and e has the normal distribution, N (0, σ_e^2). The correlation coefficient ρ between the true and the observed measurement can be written as

$$\rho = \frac{\sigma_p}{\sigma_X} \tag{2.2}$$

where

$\sigma_X = \frac{\sigma_p}{\rho} = \acute{\sigma}$ (say) is the standard deviation of X,

and the size of the measurement error

$$r = \frac{\sigma_p}{\sigma_e} \tag{2.3}$$

After some mathematical manipulation Singh (1984) showed that

$$\rho = \frac{r}{\sqrt{1+r^2}} \tag{2.4}$$

Let x be the true value of the variable which is distributed as

$$\begin{aligned} f(X, \theta, \sigma_p) &= \frac{c(\theta, \sigma_p)}{\sigma_p \sqrt{2\pi}} exp\left[\frac{-1}{2}\left(\frac{x-\theta}{\sigma_p}\right)^2\right] \quad ; as\ a \leq x \leq b \\ &= 0 \qquad ;\ otherwise \end{aligned} \tag{2.5}$$

where $c(\theta, \sigma_p) = \frac{1}{\Phi\left(\frac{b-\theta}{\sigma_p}\right) - \Phi\left(\frac{a-\theta}{\sigma_p}\right)}$ (2.6)

and a and b are the points of truncation with

$$\Phi(t) = \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} exp\left(-\frac{1}{2}y^2\right) dy$$

If $x_1, x_2, x_3, \ldots, x_m$ are m independent random variables whose probability density function (p.d.f.) is given by (2.5). The likelihood ratio of the hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta = \theta_0 + \delta\sigma_p = \theta_1 (\delta > 0)$ is given by

$$\frac{f(x_1, x_2, \ldots, x_m \mid \theta_1, \sigma_p)}{f(x_1, x_2, \ldots, x_m \mid \theta_0, \sigma_p)} = R^m exp\left[-\frac{1}{2\sigma_p^2} \sum_{i=1}^{m} \{-2\delta\sigma_p(x_i - \theta_0) + \delta^2\sigma_p^2\}\right], \tag{2.7}$$

where

$$R = \frac{c(\theta_1, \sigma_p)}{c(\theta_0, \sigma_p)}$$

The continuation region of SPRT discriminating between the two hypotheses, $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$ is given by

$$\frac{1}{\delta} \ln\left(\frac{\alpha_1}{1-\alpha_0}\right) + m\left[\frac{\delta}{2} - \left(\frac{\ln R}{\delta}\right)\right] < \frac{1}{\sigma_p} \sum_{i=1}^{m} (x_i - \theta_0) < \frac{1}{\delta} \ln\left(\frac{1-\alpha_1}{\alpha_0}\right) + m\left[\frac{\delta}{2} - \left(\frac{\ln R}{\delta}\right)\right] \tag{2.8}$$

where

$\alpha_i = Pr[Accept\ H_{1-i} \mid H_i]$, $(i = 0{,}1)$.

For very small value of α_1 the right hand inequality of (2.8) reduces to

$$\frac{1}{\sigma_p}\sum_{i=1}^{m}(x_i-\theta_0) < -\frac{1}{\delta}\ln\alpha_0 + m\left[\frac{\delta}{2}-\left(\frac{\ln R}{\delta}\right)\right] \qquad (2.9)$$

But for the observed values

$\sigma_X^2 = \sigma_p^2 + \sigma_e^2$

$$\frac{1}{\acute{\sigma}}\sum_{i=0}^{m}(x_i-\theta_0) = \frac{\rho}{\sigma_p}\sum_{i=0}^{m}(x_i-\theta_0) < -\frac{1}{\delta}\ln\alpha_0 + m\left[\frac{\delta}{2}-\left(\frac{\ln R}{\delta}\right)\right]$$
$$= \frac{1}{\sigma_p}\sum_{i=0}^{m}(x_i-\theta_0) < -\frac{1}{\rho\delta}\ln\alpha_0 + \frac{m}{\rho}\left[\frac{\delta}{2}-\left(\frac{\ln R}{\delta}\right)\right]$$

For constructing a CUSUM Chart for the mean under observed measurements, we plot the sum

$S_m = \frac{1}{\acute{\sigma}}\sum_{i=1}^{m}(x_i-\theta_0)$ against the number of observations m.

3. The effect on the dimentions and the ARL of CUSUM chart for mean of a truncated normal distribution under measurment error

A narrow V-mask will detect change more quickly but it will give more frequent false alarms. On the other hand, we could reduce the frequency of false alarms by widening the angle of mask, but the average run length for real changes would be increased.

Under the measurement error the parameter of the mask, namely the angle of the mask ϕ and the lead distance d are given by

$$tan\phi = \frac{1}{\rho}\left[\frac{\delta}{2}-\left(\frac{lnR}{\delta}\right)\right] \qquad (3.1)$$

where

$$R = \frac{c(\theta_{1,}\acute{\sigma})}{c(\theta_{0,}\acute{\sigma})} \qquad (3.2)$$

and

$$d = \frac{(-\ln\alpha_0)}{\left(\frac{\delta^2}{2} - \ln R\right)} \tag{3.3}$$

For ARL we consider the situation where the true mean θ has shifted from θ_0 to θ_1. For very small value of α_1 the ARL that is expected number of observation before the change from θ_0 to θ_1 is detected (see Mood and Graybill 1963) is approximately

$$(-\ln\alpha_0)E_{\theta_1}^{-1},$$

where

$$E_{\theta_1} = E\left[\ln\frac{f(X \mid \theta_1, \sigma_p)}{f(X \mid \theta_0, \sigma_p)} \mid \theta_1, \sigma_p\right]$$

For observed values

$$E_{\theta_1} = \int_a^b \left\{\left[\ln R - \frac{\delta^2}{2} + \frac{\delta}{\sigma_p}(x - \theta_0)\right]\frac{C(\theta_1, \sigma_p)}{\sigma_p\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x - \theta_1}{\sigma_p}\right)^2\right]\right\}dx,$$

$$= \ln R + \frac{\delta^2}{2} + \frac{\delta c(\theta_1, \sigma_p)}{\sqrt{2\pi}}\left[exp\left\{-\frac{1}{2}\left(\frac{a - \theta_1}{\acute{\sigma}}\right)^2\right\} - exp\left\{-\frac{1}{2}\left(\frac{b - \theta_1}{\acute{\sigma}}\right)^2\right\}\right] \tag{3.4}$$

4. Tabulation of results and conclusions

In Table-1, Table-2 and Table-3 the angle of the mask, lead distance and ARL for the mean chart are given, assuming $\theta_0 = 0, \sigma_p = 1$ and the size of the measurement errors *r*=2, 4, 6 and ∞. The truncation points have been taken as $\pm 2.5, \pm 2.25, \pm 2, (0,2.5), (0,2)$. The values of δ considered are 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 2.5 and 3.0 . The values of α_0 are taken to be 0.05, 0.025, 0.01 and 0.005.

It is evident from the Table-1 that for fixed range of truncation with increasing magnitude of error term and δ, an angle of mask increases. But for one sided truncation angle of mask deceases for increase in error term and increases with increase in δ. From tables for lead distance and ARL it is seen that as the size of measurement error increases the lead distance and ARL increases for fixed range of truncation in case of symmetrical truncation. But for one sided truncation it is observed that there is less increase in lead distance as compared to symmetrical truncation and reverse is the case for ARL i.e. the decrease in ARL is seen with increase error term.

Table 1. Angle of the Mask for Mean for different Truncation Points (a, b)

(a, b)	δ↓/r→	2	4	6	∞
(-2.5,2.5)	0.2	5.613	5.308	5.249	5.201
	0.4	11.091	10.493	10.377	10.283
	0.6	16.318	15.448	15.279	15.142
	0.8	21.207	20.091	19.874	19.697
	1.0	25.705	24.369	24.110	23.898
	1.5	35.166	33.381	33.031	32.747
	2.0	42.324	40.185	39.760	39.412
	2.5	47.701	45.258	44.762	44.355
	3.0	51.793	49.071	48.507	48.041
(-2.25,2.25)	0.2	5.250	4.966	4.914	4.872
	0.4	10.383	9.822	9.718	9.634
	0.6	15.297	14.468	14.314	14.190
	0.8	19.916	18.831	18.629	18.466
	1.0	24.189	22.864	22.614	22.414
	1.5	33.273	31.406	31.047	30.757
	2.0	40.258	37.921	37.460	37.085
	2.5	45.586	42.834	42.276	41.818
	3.0	49.696	46.571	45.921	45.382
(-2,2)	0.2	4.796	4.508	4.457	4.417
	0.4	9.497	8.923	8.821	8.741
	0.6	14.021	13.162	13.009	12.887
	0.8	18.302	17.162	16.956	16.793
	1.0	22.296	20.881	20.622	20.416
	1.5	30.922	28.850	28.457	28.140
	2.0	37.708	35.038	34.513	34.085
	2.5	42.994	39.787	39.135	38.599
	3.0	47.144	43.459	42.687	42.046
(0,2.5)	0.2	38.654	38.804	38.818	38.826
	0.4	40.338	40.106	40.045	39.991
	0.6	41.972	41.406	41.279	41.171
	0.8	43.552	42.695	42.510	42.356
	1.0	45.072	43.965	43.731	43.536
	1.5	48.598	47.003	46.673	46.400
	2.0	51.711	49.764	49.362	49.031
	2.5	54.415	52.184	51.722	51.340
	3.0	56.739	54.257	53.736	53.304
(0,2)	0.2	36.106	36.638	36.734	36.810
	0.4	37.660	37.751	37.762	37.769
	0.6	39.165	38.854	38.786	38.729
	0.8	40.618	39.940	39.801	39.685
	1.0	42.018	41.005	40.799	40.630
	1.5	45.277	43.546	43.194	42.906
	2.0	48.190	45.868	45.392	45.000
	2.5	50.767	47.943	47.355	46.867
	3.0	53.035	49.768	49.073	48.494

Table 2. Lead Distance for Mean for different Truncation Points(a , b) and Measurement Error(r)

r	$\delta\downarrow/\alpha_0\rightarrow$	(a=-2.5,b=2.5)				(a=-2.25,b=2.25)			
		0.05	0.025	0.01	0.005	0.05	0.025	0.01	0.005
2	0.2	170.40	209.83	261.95	301.38	182.27	224.44	280.19	322.36
	0.4	42.71	52.60	65.66	75.55	45.70	56.27	70.25	80.83
	0.6	19.07	23.48	29.31	33.72	20.41	25.13	31.37	36.10
	1.0	6.96	8.57	10.70	12.31	7.46	9.18	11.46	13.19
	2.0	1.84	2.26	2.83	3.25	1.98	2.44	3.04	3.50
	3.0	0.88	1.08	1.35	1.55	0.95	1.17	1.46	1.67
4	0.2	166.18	204.63	255.46	293.91	177.68	218.79	273.14	314.25
	0.4	41.68	51.32	64.07	73.71	44.59	54.91	68.55	78.87
	0.6	18.62	22.93	28.63	32.94	19.95	24.56	30.66	35.28
	1.0	6.82	8.39	10.48	12.06	7.32	9.02	11.26	12.95
	2.0	1.83	2.25	2.81	3.23	1.98	2.44	3.05	3.51
	3.0	0.89	1.10	1.37	1.58	0.97	1.20	1.50	1.72
6	0.2	165.29	203.54	254.09	292.34	176.62	217.49	271.51	312.38
	0.4	41.46	51.05	63.74	73.33	44.34	54.60	68.16	78.41
	0.6	18.53	22.82	28.48	32.77	19.84	24.43	30.50	35.09
	1.0	6.79	8.36	10.43	12.00	7.29	8.98	11.21	12.89
	2.0	1.83	2.25	2.81	3.23	1.98	2.44	3.05	3.51
	3.0	0.90	1.10	1.38	1.58	0.98	1.21	1.51	1.73
∞	0.2	164.56	202.64	252.97	291.05	175.73	216.39	270.14	310.80
	0.4	41.28	50.83	63.46	73.01	44.12	54.33	67.82	78.03
	0.6	18.45	22.72	28.36	32.63	19.75	24.31	30.35	34.92
	1.0	6.76	8.33	10.39	11.96	7.26	8.94	11.17	12.85
	2.0	1.82	2.24	2.80	3.22	1.98	2.44	3.05	3.50
	3.0	0.90	1.11	1.38	1.59	0.99	1.21	1.51	1.74

Table 2. Lead Distance for Mean for different Truncation Points(a , b) and Measurement Error(r) (cont.)

r	$\delta\downarrow/\alpha_0\rightarrow$	(a=-2,b=2)				(a=0,b=2.5)				(a=0,b=2)			
		0.05	0.025	0.01	0.005	0.05	0.03	0.01	0.01	0.05	0.03	0.01	0.01
2	0.2	199.61	245.80	306.85	353.04	20.94	25.78	32.19	37.03	22.96	28.27	35.30	40.61
	0.4	50.05	61.63	76.94	88.52	9.86	12.14	15.16	17.44	10.85	13.36	16.68	19.19
	0.6	22.35	27.53	34.36	39.53	6.21	7.64	9.54	10.98	6.85	8.44	10.53	12.12
	1.0	8.17	10.06	12.56	14.45	3.34	4.11	5.14	5.91	3.72	4.58	5.71	6.57
	2.0	2.17	2.67	3.33	3.83	1.32	1.63	2.03	2.34	1.50	1.84	2.30	2.65
	3.0	1.04	1.28	1.59	1.83	0.73	0.90	1.13	1.30	0.84	1.03	1.29	1.49
4	0.2	195.84	241.16	301.06	346.37	19.20	23.64	29.52	33.96	20.76	25.56	31.91	36.72
	0.4	49.17	60.55	75.59	86.97	9.17	11.29	14.09	16.21	9.97	12.28	15.33	17.63
	0.6	22.01	27.10	33.83	38.92	5.84	7.19	8.97	10.32	6.39	7.87	9.82	11.30
	1.0	8.09	9.97	12.44	14.32	3.20	3.94	4.92	5.66	3.55	4.37	5.46	6.28
	2.0	2.20	2.71	3.38	3.89	1.31	1.61	2.01	2.31	1.50	1.84	2.30	2.65
	3.0	1.09	1.34	1.67	1.92	0.74	0.91	1.14	1.31	0.87	1.07	1.34	1.54
6	0.2	194.81	239.89	299.48	344.55	18.87	23.24	29.01	33.38	20.35	25.06	31.28	35.99
	0.4	48.93	60.25	75.21	86.53	9.03	11.12	13.89	15.98	9.80	12.07	15.07	17.34
	0.6	21.91	26.98	33.68	38.75	5.77	7.10	8.86	10.20	6.30	7.76	9.68	11.14
	1.0	8.07	9.94	12.41	14.27	3.17	3.91	4.88	5.61	3.52	4.33	5.41	6.22
	2.0	2.21	2.72	3.39	3.91	1.30	1.60	2.00	2.30	1.50	1.84	2.30	2.65
	3.0	1.10	1.35	1.69	1.94	0.74	0.91	1.14	1.31	0.88	1.08	1.35	1.55
∞	0.2	193.90	238.77	298.08	342.94	18.61	22.92	28.61	32.92	20.01	24.65	30.77	35.40
	0.4	48.71	59.98	74.88	86.15	8.93	10.99	13.72	15.79	9.67	11.90	14.86	17.10
	0.6	21.82	26.87	33.55	38.60	5.71	7.03	8.78	10.10	6.23	7.67	9.57	11.01
	1.0	8.05	9.91	12.37	14.23	3.15	3.88	4.85	5.58	3.49	4.30	5.37	6.18
	2.0	2.21	2.73	3.40	3.91	1.30	1.60	2.00	2.30	1.50	1.84	2.30	2.65
	3.0	1.11	1.36	1.70	1.96	0.74	0.92	1.14	1.32	0.88	1.09	1.36	1.56

Table 3. ARL for Mean for different Truncation Points (a ,b) and Measurement Error(r)

r	δ↓/α_0→	(a=-2.5,b=2.5)				(a=-2.25,b=2.25)			
		0.05	0.025	0.01	0.005	0.05	0.025	0.01	0.005
2	0.2	176.48	217.31	271.29	312.12	192.53	237.07	295.96	340.50
	0.4	44.54	54.85	68.47	78.78	48.65	59.91	74.78	86.04
	0.6	20.11	24.76	30.91	35.56	22.00	27.09	33.82	38.92
	1.0	7.60	9.35	11.68	13.44	8.36	10.29	12.84	14.78
	2.0	2.31	2.84	3.55	4.08	2.57	3.17	3.96	4.55
	3.0	1.31	1.62	2.02	2.32	1.47	1.82	2.27	2.61
4	0.2	167.67	206.47	257.76	296.55	180.22	221.91	277.03	318.73
	0.4	42.35	52.15	65.10	74.90	45.61	56.16	70.11	80.67
	0.6	19.14	23.57	29.42	33.85	20.68	25.47	31.79	36.58
	1.0	7.26	8.94	11.16	12.84	7.92	9.75	12.17	14.00
	2.0	2.26	2.78	3.47	3.99	2.52	3.10	3.87	4.46
	3.0	1.33	1.64	2.04	2.35	1.50	1.85	2.31	2.66
6	0.2	166.13	204.57	255.39	293.83	177.98	219.17	273.60	314.79
	0.4	41.96	51.67	64.50	74.21	45.06	55.48	69.26	79.69
	0.6	18.97	23.35	29.16	33.54	20.44	25.17	31.42	36.15
	1.0	7.20	8.86	11.07	12.73	7.84	9.65	12.05	13.86
	2.0	2.25	2.77	3.45	3.97	2.51	3.09	3.86	4.44
	3.0	1.33	1.64	2.05	2.36	1.51	1.86	2.32	2.67
∞	0.2	164.93	203.09	253.54	291.70	176.22	216.99	270.89	311.66
	0.4	41.66	51.30	64.04	73.67	44.62	54.94	68.59	78.91
	0.6	18.83	23.19	28.95	33.30	20.25	24.93	31.13	35.81
	1.0	7.15	8.80	10.99	12.64	7.77	9.57	11.95	13.74
	2.0	2.24	2.76	3.44	3.96	2.50	3.08	3.85	4.43
	3.0	1.34	1.65	2.06	2.36	1.52	1.87	2.33	2.68

Table 3. ARL for Mean for different Truncation Points (a ,b) and Measurement Error(r) (cont.)

r	$\delta\downarrow/\alpha_0\rightarrow$	(a=-2,b=2)				(a=0,b=2.5)				(a=0,b=2)			
		0.05	0.025	0.01	0.005	0.05	0.03	0.01	0.01	0.05	0.03	0.01	0.005
2	0.2	217.11	267.35	333.75	383.99	136.66	168.28	210.08	241.70	156.88	193.18	241.17	277.47
	0.4	54.89	67.59	84.37	97.07	52.71	64.91	81.03	93.23	62.73	77.24	96.42	110.94
	0.6	24.84	30.59	38.19	43.94	28.32	34.87	43.53	50.08	34.68	42.70	53.31	61.33
	1.0	9.45	11.64	14.53	16.72	12.06	14.84	18.53	21.32	15.43	19.00	23.72	27.29
	2.0	2.93	3.60	4.50	5.18	3.52	4.33	5.40	6.22	4.79	5.90	7.36	8.47
	3.0	1.68	2.07	2.58	2.97	1.80	2.21	2.77	3.18	2.49	3.06	3.82	4.40
4	0.2	200.20	246.52	307.76	354.08	271.73	334.60	417.71	480.59	324.08	399.07	498.20	573.18
	0.4	50.74	62.48	77.99	89.73	82.80	101.96	127.29	146.44	103.11	126.97	158.51	182.37
	0.6	23.06	28.39	35.44	40.78	38.89	47.89	59.78	68.78	49.89	61.43	76.69	88.23
	1.0	8.88	10.93	13.65	15.70	14.28	17.58	21.95	25.25	19.14	23.57	29.42	33.85
	2.0	2.87	3.53	4.41	5.07	3.67	4.52	5.64	6.49	5.23	6.44	8.04	9.25
	3.0	1.72	2.12	2.65	3.05	1.87	2.30	2.87	3.31	2.69	3.31	4.14	4.76
6	0.2	197.04	242.63	302.90	348.49	347.05	427.34	533.49	613.79	425.28	523.68	653.76	752.16
	0.4	49.96	61.52	76.80	88.36	94.59	116.48	145.41	167.29	120.23	148.05	184.82	212.64
	0.6	22.72	27.98	34.93	40.18	42.31	52.10	65.04	74.83	55.17	67.93	84.81	97.57
	1.0	8.77	10.80	13.48	15.51	14.85	18.29	22.83	26.26	20.16	24.83	30.99	35.66
	2.0	2.86	3.52	4.39	5.05	3.70	4.55	5.68	6.54	5.32	6.56	8.18	9.42
	3.0	1.73	2.13	2.66	3.06	1.88	2.32	2.90	3.33	2.74	3.37	4.21	4.84
∞	0.2	194.51	239.52	299.01	344.02	452.67	557.41	695.87	800.61	578.01	711.75	888.54	1022.28
	0.4	49.34	60.76	75.85	87.26	107.45	132.31	165.17	190.03	139.81	172.16	214.92	247.27
	0.6	22.45	27.65	34.52	39.71	45.67	56.24	70.21	80.78	60.56	74.57	93.09	107.11
	1.0	8.68	10.69	13.35	15.35	15.36	18.92	23.62	27.17	21.10	25.98	32.43	37.31
	2.0	2.85	3.51	4.38	5.04	3.72	4.58	5.72	6.58	5.40	6.65	8.30	9.55
	3.0	1.74	2.14	2.68	3.08	1.90	2.33	2.91	3.35	2.77	3.41	4.26	4.90

Also it is seen that the angle of mask increases as the range of the truncation is increased, and is bigger in case of one sided truncation. The ARL and lead distance deceases with increase in range of truncation. With increase in δ the angle of mask increases, while the ARL and the lead distance decreases. The ARL and lead distance increases as α_0 is deceased and as δ is deceased.

For one sided truncation we see that lead distance is less and angle of mask is greater compared to symmetrical truncation. This shows that in case of symmetrical truncation, as angle of mask is smaller, it will detect change more quickly, but it will give more frequent false alarms, as seen by comparing table for ARL for error free case for both symmetrical truncation and one sided truncation.

This paper considered the normal distribution case. For other symmetrical distribution this model will not be suitable. We have to develop another model for other symmetrical distribution.

From the above discussion it is clear that truncated normal distribution under measurement error appears frequently in quality control problems. In order to meet certain specification the supplier uses some technique to ensure that the quality characteristic must satisfy error free observations, error prone data create a serious problem for determining the value of the mask parameter and the ARL. This paper may be *practically implicated* within engineering, finance, medicine, environmental statistics, and many other fields.

REFERENCES

A.B. YEH, D.K. LINAND and C. VENKATARAMANI. (2004).Unified CUSM Charts for Monitoring Process Mean and Variability, Quality Technology and Quantitative Management, Vol.1, No.1, pp.65-86.

CHANG, M.N. (1990). Weak Convergence of a Self Consistent Estimator of the Survival Function with Doubly Censored Data, Annals of Statistics, 18: 391-404.

G.NENES and G. TAGARAS. (2010). Evaluation of CUSUM charts for Finite-Horizon Processes, Communication in Statistics-Simulation and Computation, Vol.39, Issue 3, pp.578-597.

HUNTER, J.S. (1986). The Exponentially Weighted Moving Average ,Journal of Quality Technology, 18:203-210.

J. H. RYU, H. WAN and S.KIM. (2010). Optimal Design of a CUSUM Chart for a Mean Shift Of Unknown Size, Journal of Quality Technology, Vol.42, No.3, pp.1-16.

JOHNSON, N.L. and LEONE, F.C. (1962). Cumulative Sum Control Charts: Mathematical Principles Applied to their Construction and Use Part II, Industrial Quality Control XIV(2), pp.22-28.

KEIDING, N. and GILL, R.D. (1987). Research Report No. 87/3, Statistical Research Unit, University Copenhagen, Denmark.

M.A.A. COX. (2009). Control charts for monitoring observations from a truncated normal distribution, Journal of Risk Finance, Vol. 10 Iss: 3, pp.288 – 304.

MOOD, A.M. & GRAYBILL, F.A. (1963). Introduction to the Theory of Statistics; Mc Graw Hill Book Co .Inc. second Edition.

NELSON, W. (1990). Hazard Plotting of Left Truncated Life data, Journal of Quality Technology, 22:230-238.

O,A.GRIGG and D.J. SPIEGELHALTER. (2008). An Empirical Approximation to the Null Unbounded Steady-State Distribution of the Cumulative Sum Statistic, Technometrics, 50(4):501-511.

PATEL, M.N. and GAJJAR, A.V. (1994). Cumulative Sum Control Charts for Intervened Geometric Distribution, International Journal of Management and Systems,10(2):181-188.

SCHEIDER, H. (1986). Truncated and Censored Samples from Normal Distribution, Marcel Dekker, New York.

WOODROOFE, M. (1985). Estimating a Distribution Function with Truncated Data, Annals of Statistics, 13: 163-177.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 107—142

DATA INTEGRATION AND SMALL DOMAIN ESTIMATION IN POLAND – EXPERIENCES AND PROBLEMS

Elżbieta Gołata[1]

ABSTRACT

The aim of the study could be identified twofold. On the one hand, it was a presentation of Polish experiences as concerns the most important methodological issues of contemporary statistics. These are the problems of data integration (DI) and statistical estimation for small domains (SDE).On the other hand, attempts to determine relationship between these two groups of methods were undertaken. Given convergence of the objectives of both SDE and DI, that is: striving to increase efficiency of the use of existing sources of information, simulation study was conducted. It was aimed at verifying the hypothesis of synergies referring to combined application of both groups of methods: SDE and DI.

Keywords: Small domain estimation, data integration.

1. Aim of the study

The study was aimed at presentation of Polish experiences in Small Domain Estimation (SDE) and Data Integration (DI). This goal will be achieved in an indirect way. First, some basic remarks concerning both methods will be discussed pointing out similarities and dissimilarities, especially in such dimensions as: purpose, methods and techniques, data sources, evaluation and other problems and threats that appear with practical application.

In general, both methods are used to improve the quality of the statistical estimates, to increase their substantive range and precision using all available sources of information. It can be assumed that combined application of both methods will result in synergy effects on the quality of statistical estimates.

Small Domain Estimation are techniques aimed to provide estimates for subpopulations (domains) for which sample size is not large enough to yield direct estimates of adequate precision. Therefore, it is often necessary to use

[1] Poznan University of Economics, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland, e-mail: elzbieta.golata@ue.poznan.pl.

indirect estimates that ’borrow strength’ by using values of variables of interest from related areas (domains) or time, and sometimes of both: time and domains. These values are brought into the estimation process through a model. Availability of good auxiliary data and suitable linking models are crucial to indirect estimates (Rao 2005). Review of small area estimation methods is included, among others, in such works as Gosh and Rao (1994), Rao (1999, 2003), Pfeffermann (1999) and Skinner C. (1991).

Data Integration could be understood as a set of different techniques aimed to combine information from distinct sources of data which refer to the same target population. Moriarity and Scheuren (2001, p.407) indicated that practical needs formed the basis for the development of statistical methods for data integration (Scheuren 1989). Among the basic studies in this subject, the following should be mentioned Kadane (2001), Rogers (1984) Winkler (1990, 1994, 1995, 1999, 2001), Herzog T. N., Scheuren F. J., Winkler W.E. (2007), D’Orazio M., Di Zio M., Scanu M. (2006) and Raessler (2002). Because of the growing need for complex, multidimensional information for different subsets or domains, in times of crisis and financial constraints, data integration is becoming a major issue. The problem is to use information available from different sources efficiently so as to produce statistics on a given subject while reducing costs and response burden and maintaining quality (Scanu 2010).

Both groups of techniques refer to additional data sources that are specifically exploited. These can be two data sets that are obtained from independent sample surveys. Another, often encountered situation refers to the use of administrative data resources as registers. In this case data from registers are linked to survey data. Via data integration process we can extend - enrich the information available from a sample survey with data from administrative registers. In this way we enable 'borrowing strength’ from other data sources at individual level, which, assuming a strong correlation, allows for estimating from the sample for domains at lower aggregation level than the one resulting from the original sample size. This seems to be the most important connection between SDE and DI and the main advantage of the joint implementation of both techniques.

For this reason, an attempt was made to determine relationship between these two groups of methods. Given convergence of the objectives of both SDE and DI, that is: striving to increase efficiency of the use of existing sources of information, simulation study was conducted. It was aimed at verifying the hypothesis of synergies in data quality and availability resulting from combined application of both groups of methods: SDE and DI. The structure of the paper reflects studies which have been taken to achieve the above target.

First basic characteristics of both groups of methods will be presented in the context of Polish experiences which are shortly described in Section 2 of the paper. Special attention was given to the use of alternative data sources in Polish official statistics, especially administrative registers in the context of population census 2011. This census was the first survey designed to integrate administrative registers and data from a 20% sample. Next, two simulation studies which attempt to apply the indirect estimation methodology for databases resulting from the integration of different sources will be discussed. In section 3 estimation is conducted for linked data from sample survey and administrative records. This case is illustrated with the experiences from MEETS1 Project. Second case study presented in Section 4 refers to linked data from two surveys. Procedures used in simulation studies are discussed in more detail with references to the literature. An empirical assessment of the simulation studies will form the basis for final conclusions discussed in Section 5.

2. Data Integration and Small Domain Estimation in Poland

For a long time the need to use alternative sources of information in Polish public statistics was not conscious. Exception may constitute such fields which traditionally made use of administrative resources as justice statistics. But on the other hand even in such basic areas as vital statistics, the administrative records were not fully accepted. For example, the Central Population Register PESEL, over the years was not used for constructing population projections (Paradysz 2010). Significant differences were observed in the population structure by age and place of residence according to official statistics estimates based on census structure and the register (fig. 1). The divergence measured by the relative difference W_{L_t/P_t} in the number of population estimates by official statistics (Lt) and Population Register (Pt) for the city of Poznan at the end of 2000 (cf. formula (1)), amount to even more than 30% .

$$W_{L_t/P_t} = \frac{(L_t - P_t)\cdot 100}{P_t} \tag{1}$$

Three highest relative differences deserve particular attention. The first is almost 8 percentage of the surplus of population estimates in comparison with the registered for those at zero years of age (children before first year). As this difference relates to the same degree for both sexes, it can be assumed that it stems from the delay in births register. Another characteristic is the excess in population estimates for age 18 - 25 years. The reason for this is probably due to recognition by the census of young people (students or working in Poznan) as permanent residents, although they do not have such status. But population register refers to legal status notified by permanent residence. For people over 25 years, a systematic decrease in the relative differences can be noticed. This may indicate a return of persons to their place of permanent residence, or legalization of their residence because of work or marriage.

[1] The MEETS project was conducted under Grant Agreement No. 30121.2009.004-2009.807 signed on 31.10.2009 between the European Commission and the Central Statistical Office of Poland between 01.11.2009 and 28.02.2011. The Project was aimed at Modernisation of European Enterprise and Trade Statistics, especially to examine the possibilities of using administrative register to estimate enterprise indicators.

Also, a significant negative difference could be noticed between population estimates and register for population aged about 85 years and more. This is probably related to the under-coverage of the elderly in the National Census of Population and Housing in 2002. Confirmation of this hypothesis can be found in population tables for subsequent years after the census, in which negative numbers of people aged over 90 should be observed, if death by age would be considered for various levels of spatial aggregation. It follows that the dying person were not included in the census (Multivariate analysis of errors …, 2008, p.13-14).

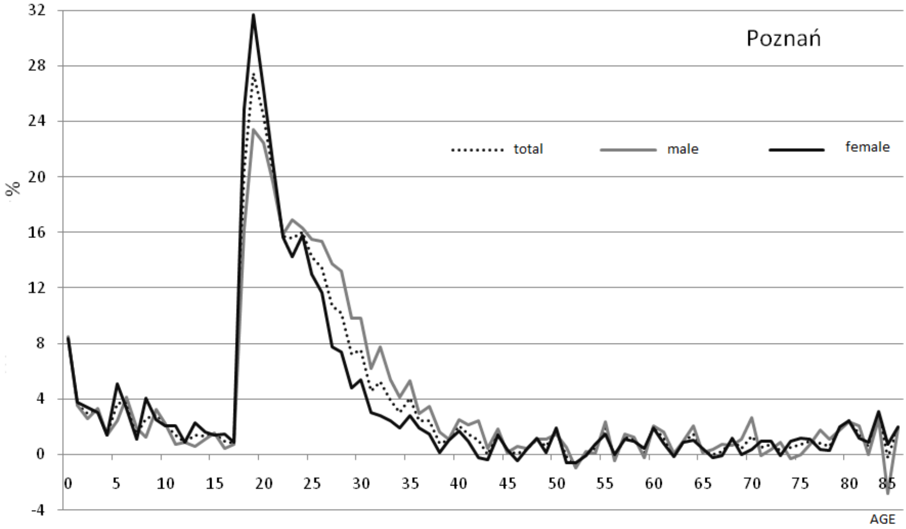


Figure 1: *Relative differences between population estimates by official statistics (L_t) and Population Register (P_t), city of Poznan, 31.12.2000.*

Source: Tomasz Józefowski, Beata Rynarzewska-Pietrzak, 2010.

Changes in the intensity of use of administrative records took place within the last five years, during preparations for the National Census of Population and Housing which was conducted from April to June 2011. This census was based on the population register but used data from about 30 other registers. In addition, a survey on a 20% sample allowed collection of detailed information on demographic and social structures as well as economic activity. Among Polish main experiences in SAE and DI one should mention:

1. EURAREA – Enhancing Small Area Estimation Techniques to meet European needs, IST-2000-26290, Poznan University of Economics, 2003 – 2005
2. ESSnet on Small Area Estimation – SAE 61001.2009.003-2009.859, Statistical Office in Poznan, 2010 – 2011
3. ESSnet on Data Integration – DI 61001.2009.002-2009.832, Statistical Office in Poznan, 2010 – 2011

4. Modernisation of European Enterprise and Trade Statistics – MEETS 30121.2009.004-2009.807, Central Statistical Office, 2010 – 2011
5. Experimental research conducted by Group for mathematical and statistical methods in : Polish Agriculture Census PSR 2010 and National Census of Population and Housing NSP 2011
   - Data Integration of Central Population Register PESEL and Labour Force Survey, July 2009
   - Nonparametric matching: datasets from a micro-census and Labour Force Survey, 2011
   - *Propensity scores matching:* Labour Force Survey and Polish General Social Survey PGSS to enlarge the information scope of the social data base, May 2011

Both groups of methods: Data Integration as well as Small Domain Estimation refer to additional data sources. In SDE auxiliary data is needed to 'borrow strength'. To meet this requirement, the additional, external data source should be a reliable one. Typically, due to specified by law, rules regulating organization of the registers, administrative records data should satisfy this requirement[1]. It is also important, that in many cases, registers provide population data and population total (though the population in task might be differently defined). On the other hand, there are some small area estimators that require domain totals. Thus, in the estimation procedure individual data is not always necessary. To resume, we begin with applying small domain estimation methodology with area level models. Firstly, we use integrated data from sample and register, and secondly the case of two integrating samples is considered. In each of the two cases a simulation study was conducted and small domain estimators: GREG, SYNTHETIC and EBLUP were applied to integrated data.

In the next section presentation of experiences in integration sample data with registers refer to results obtained within the MEETS project. In the following section, study on integration of two samples was based on a pseudo-population data from Polish micro-census 1995. The process of estimating statistics for small domains applied in both sections relied on findings of the EURAREA2 project. The main task of the project was to popularize indirect estimation methods and to assess their properties with respect to complex sampling designs used in statistical practice. In addition to conducting a detailed analysis of the research problem, the project participants created specialist software designed to implement estimation

[1] Of course each register needs special evaluation. For example, analysis conducted by Młodak and Kubacki (2010) showed that matching data on individual farms for the needs of Agricultural Census 2010 showed large discrepancies between various registers and 'borrowing strength' was seriously disturbed.

[2] The European project entitled EURAREA IST-2000-26290 *Enhancing Small Area Estimation Techniques to meet European needs* was part of the Fifth framework programme of the European Community for research, technological development and demonstration activities. The project was coordinated by ONS – Office for National Statistics, UK) with the participation of six countries: The United Kingdom, Finland, Sweden, Italy, Spain and Poland.

techniques developed in the project. The software, with associated theoretical and technical documentation, was published on the Eurarea project website[1] (Eurarea_Project_Reference_Volume, 2004). Estimation within both sections was conducted using the EBLUPGREG program[2]. Detail description of the estimation techniques used in the study is given in R. Chambers and A. Saei (2003).

3. Empirical evaluation of SDE for linked data - integrating sample data with register - MEETS

One of the goals of the MEETS project was to highlight possibilities of using administrative resources to estimate enterprise[3] indicators in twofold way (*Use of Administrative Data for Business Statistics* (2011):

- to increase the estimation precision
- to increase the information scope by providing estimates taking into account kind of business activity (PKD classification) at regional level.

Data Integration

The following administrative systems constituting potential sources for short-term and annual statistics of small, medium and big enterprises were identified, described and used as auxiliary data source in the estimation process:

1) Tax system – information system conducted by the Ministry of Finance – fed with data from tax declarations and statements as well as identification request forms in the field of:
   - database on taxpayers of the personal income tax – PIT
   - database on taxpayers of the corporate income tax – CIT
   - database on taxpayers of the value added tax – VAT
   - National Taxable Persons Records – KEP.
2) System of social insurance – information system conducted by the Social Insurance Institution, the so-called Comprehensive IT System of the Social Insurance Institution (KSI ZUS) fed with data from insurance documents concerning contribution payers and the insured Central Register of the Insured (CRU) and Central Register of Contribution Payers (CRPS):
   - register of natural persons (GUSFIZ)
   - register of legal persons (GUSPRA).

The primary source of data on companies in Poland is the DG1 survey carried out by Central Statistical Office. This survey covers all large companies (of more

[1] The Eurarea_Project_Reference_Volume (2004) can be downloaded from http://www.statistics.gov.uk/eurarea.

[2] Veijanen A., Djerf K., Sőstra K., Lehtonen R., Nissinen K., 2004, EBLUPGREG.sas, program for small area estimation borrowing Strength Over Time and Space using Unit level model, Statistics Finland, University of Jyväskylä.

[3] The project covered enterprises employing more than 9 persons.

than 50 employees) and 20% sample of medium-sized enterprises (the number of employees from 10 to 49 people). In the research the following data referring to DG1 survey were used:

- The DG-1 database directory - list of all small, medium and large economic units used as a frame
- DG-1 survey for 2008.

The data available constituted of over 180 files of different size and structure. For purposes of the study December 2008 was treated as a reference period, as for this period most information from administrative databases was available. To match the records from different datasets, two primary keys were used: NIP and REGON identification numbers. The purpose of integration was to create a database, in which an economic entity would be described by the largest possible number of variables. The DG-1 directory from December 2008 was used as a starting point. This data set was combined with information from the administrative databases and DG-1 reporting. The main obstacle to matching records were missing identification numbers[1].

Table 1. Results of integrating datasets from statistical reporting and administrative databases

Voivodships	Number of matched records				Percentage of unmatched records	Number of records with NIP duplicates
	all sections		4 sections *			
	DG-1 directory	DG-1	DG-1 directory	DG-1		
Dolnoslaskie	6044	2176	4561	1601	2,7	37
Kujawsko-pomorskie	4018	1694	3331	1392	2,2	13
Lubelskie	3040	1217	2485	961	1,4	2
Lubuskie	2278	944	1789	733	1,4	7
Lodzkie	5666	2153	4707	1744	2,1	56
Malopolskie	6844	2402	5314	1860	2,6	45
Mazowieckie	15059	4783	11172	3578	13,5	167
Opolskie	1912	852	1519	654	1,7	7
Podkarpackie	3543	1529	2925	1239	1,3	16
Podlaskie	1892	774	1540	614	1,9	7
Pomorskie	5220	1744	3906	1347	4,2	16
Slaskie	11066	3970	8728	3049	2,5	47
Swietokrzyskie	2131	902	1730	687	1,8	24

[1] It should be stressed that the REGON number is used as the main identification number for statistical sources, while institutions such as the Ministry of Finance or the Social Insurance Institution rely mostly on the NIP number.

Table 1. Results of integrating datasets from statistical reporting and administrative databases (cont.)

Voivodships	Number of matched records				Percentage of unmatched records	Number of records with NIP duplicates
	all sections		4 sections *			
	DG-1 directory	DG-1	DG-1 directory	DG-1		
Warminsko-mazurskie	2932	1093	2159	847	5,7	7
Wielkopolskie	10553	3256	8460	2724	11,2	57
Zachodniopomorskie	3270	1209	2324	911	4,7	32

Remark: * The study was restricted to the following four biggest PKD sections: *processing industry, manufacturing, trade, transport.*

Source: Use of Administrative Data for Business Statistics, GUS, US Poznan 2011.

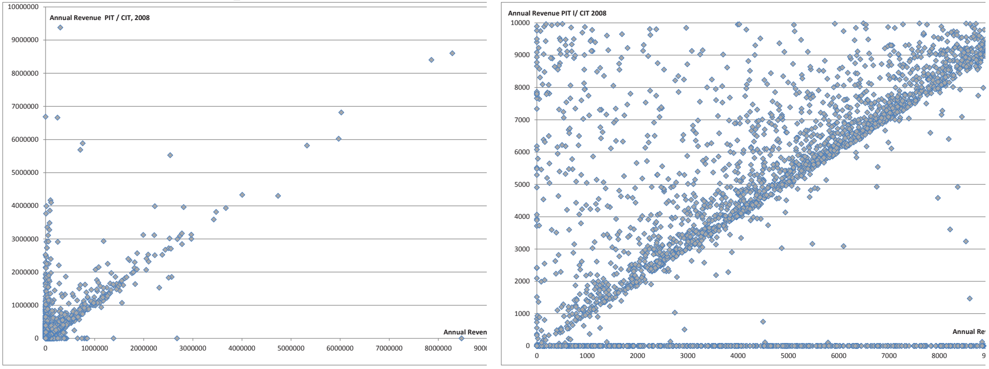
In the process of database integration a special MEETS real data set was created. It contained records about economic entities representing the four PKD sections of economic activity (*manufacturing, construction, trade, transport*), which participated in the DG-1 survey in December 2008 and which were successfully combined with information from the the KEP, CIT, PIT and ZUS databases (tab. 1). The database was treated as the population in the simulation study.

There were various reasons for multiple matching of NIP numbers. In the case of some enterprises, the ZUS register contained 2 or more NIP numbers for one REGON number[1]. The majority of records that couldn't be matched were those relating to small entities. For example, out 1,183 records of the DG-1 directory for the Wielkopolska voivodship that couldn't be matched with register records, 1,173 were small entities. This indicates that the DG-1 directory is largely out of date with respect to enterprises employing from 10 to 49 persons[2]. In the case of medium and big enterprises, which are all subject to the DG-1 reporting, the data

[1] This situation occurred when the activity of a given enterprise was carried out by more persons, each identified by a separate NIP number. In the case of the parent business unit and its local units, the first 9 digits of 14-digit REGON numbers were identical. As DG-1 directory contains only 9-digit numbers, identifying the parent business unit, data integration resulted in combining information about the parent business unit as well as other related local units present in the databases.

[2] In Polish official statistics the category of medium enterprises comprises economic entities employing from 10 to 49 persons and those employing more than 49 persons are referred to as big. Accession to the EU caused the necessity of adjustment of national regulations concerning the division of entrepreneurs to the Union's legal articles, i.e. Recommendation of the Commission of 6 May 2003 concerning the definition of micro-, small and medium enterprises (Recommendation 2003/361/EC). Basing on legal definitions the set of entities is divided into the following groups: (i) micro-enterprises – employing not more than 9 persons, (ii) small enterprises – employing from 10 to 49 persons, (iii) medium enterprises – employing from 50 to 249 persons, (iv) big enterprises – employing more than 249 persons.

are regularly updated. In contrast, only 10% of small enterprises are subject to DG-1 reporting. Consequently, it is impossible to update the DG-1 directory for this section of enterprises[1].



A. Scale fitted to units with the highest revenue (limited to PLN 10 000 000)

B. Scale not fitted to units with the highest revenue (limited to PLN 10 000)

Figure 2: Relationship between the values of accumulated revenue - from DG-1, PIT or CIT register, all units together 2008.

Source: G. Dehnel (2011), pp.58-64.

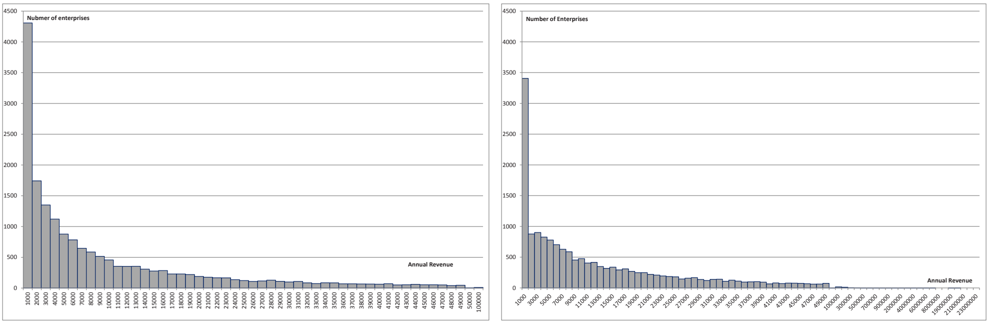
Following the integration of databases it was possible to assess the quality of information provided by the statistical reporting. One noteworthy fact was a considerable number of economic entities with the null value for revenue in the DG-1 survey and positive values of revenue in the PIT and CIT databases (fig. 2.A and 2.B). Most discrepancies between values in the databases and those in the DG-1 survey could be accounted for by a certain terminological incompatibility between the definition of *revenue* in each of the data sources. In the DG-1 survey the variable *revenue* comprises only sales of goods and services produced by the enterprise. Consequently, if an enterprise doesn’t produce anything but acts only as a sales agent, it earns no revenue according to this definition.

Scatterplot presenting DG1 and PIT data (fig. 2.A) seem to centre around the identity line. However closer analysis reveals that the line is formed largely by relatively numerous units characterized by extreme values of revenue. If these units were omitted by limiting revenue to the level of PLN 10,000, the resulting picture is significantly different (fig. 2.B). In addition to units, for which revenue reported in the DG-1 survey coincides with the value reported in tax return forms ($y_1=y_2$), one can see two other patterns. First, there is a large group of units reporting positive revenue in the DG-1 survey while displaying missing or zero values in the tax register (represented by dots lying on the X-axis). This

[1] Statistical offices have only registration information at the start of economic activity – when REGON number is assigned. Information about the activity closure has only been systematically available since the introduction of new regulations in 31 March 2009.

phenomenon can partly be accounted for by the terminological discrepancy between the definition of revenue in the DG-1 survey and the PIT/CIT tax register. Another, equally large group, is made up of units whose revenue reported in tax return forms considerably exceeded values reported in the DG-1 survey (represented by dots lying above the identity line (y_1=y_2). It’s worth noting that there were virtually no cases of units reporting lower revenue in tax return forms than in the DG-1 survey.

In order to estimate selected variables of economic entities their specific characteristics should be taken into account. One of the major challenges are non-homogenous distributions[1]. This refers both to variables estimated on the basis of sample surveys and those coming from administrative databases, which are used as auxiliary variables in the estimation process (fig. 3.A and 3.B). The distribution of revenue shows that a relatively large percentage of economic entities display zero values. For example 9% of entities that participated in the DG-1 survey reported no revenue. On the other hand, many entities in PIT and CIT register didn’t have information about revenue. Businesses with missing or zero values accounted for 14% of all units contained in the MEETS real data set.



A. DG1 data B. PIT or CIT Register data

Figure 3: *Distribution of enterprises by annual revenue, 2008.*

Source: G. Dehnel (2011), pp.57.

The effect of outliers on estimation can be significant, since in such situations estimators don’t retain their properties such as resistance to bias or efficiency. Outliers, non-typical data or null values, however, are an integral part of each population and cannot be dismissed in the analysis. For this reason, in addition to using the classic approach, work is being done to develop more robust methods2. Such methods could be mentioned as GREG estimation, the model of Chambers

[1] Some basic statistical description might be additionally given by the following characteristics:
Annual revenue DG-1, 2008: mean 41572, median 5914, std. 357783, CV 861%;
Annual revenue PIT or CIT, 2008: mean 71947, median 11154, std. 596294, CV 829%.

[2] Robust estimation methodology, as more complicated and challenging to use, will be dealt with in more detailed in further studies.

or Winsor estimation (R. Chambers, 1996, R. Chambers, H. Falvey, D. Hedlin, P. Kokic, 2001 and Dehnel, 2010).

All variables from the DG-1 survey and administrative databases were taken into account in modelling and correlation analysis. Despite of certain discrepancies between variable values in the two sources correlation was regarded as strong. Simulation study was conducted on 1000 samples drawn from the MEETS real data set according to the sampling design as the one used by GUS. For each sample 'standard'[1] SDE estimators: GREG, SYNTHETIC and EBLUP were applied to estimate revenue and other economic indicators in the breakdown of PKD sections at country and at regional level[2].

Estimation of *revenue* by PKD section

The results of estimating *revenue* at the level of selected PKD sections are presented in Tables 2 – 4. Table 2 contains expected values obtained in the simulation study after 1000 replications. The last column contains mean revenue within each section in the MEETS real data set. It is used as the benchmark to assess the convergence of estimates. The actual assessment of estimation precision and bias is possible using information presented in tables 3 and 4.

Table 2: *The expected value of estimators for revenue, 2008*

PKD Section	Estimator				Population MEAN
	DIRECT	GREG	SYNTHETIC	EBLUP	
Manufacturing	54585.85	54625.55	54768.17	54661.80	54576.28
Construction	34855.68	34836.24	34559.73	34703.67	34898.88
Trade	80320.49	80244.88	79884.69	80201.53	80280.19
Transport	63016.47	63255.07	63625.85	63386.54	63028.05

Source: Golata (2011).

Table 3: *REE of estimators for revenue, 2008*

PKD Section	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	0.55	0.37	0.49	0.31
Construction	2.47	0.78	1.14	0.84
Trade	2.17	0.60	1.50	0.66
Transport	1.28	1.73	1.02	1.43

Source: Golata (2011).

[1] The estimators referred to as 'standard' in terms of EURAREA project are: direct (Horvitz-Thompson), GREG (Generalised REGression), regression synthetic and EBLUP (Empirical Best Linear Unbiased Predictor) estimators.

[2] All programming and estimation work was carried out in the Centre for Small Area Estimation at the Statistical Office in Poznan.

The Mean Squared Error (MSE) of an estimator is a measure of the difference between values implied by an estimator and the true values of the quantity being estimated. MSE is equal to the sum of the variance and the squared bias of the estimator[1]. The Relative Error of the Estimate (REE) was calculated on the basis of the MSE as a percentage of the 'true' population value of the task variable (*revenue*). The absolute bias of the estimator (tab. 4) was defined as the difference between the expected and real value.

Table 4: *Absolute bias of estimators for revenue, 2008*

PKD Section	Absolute bias of estimators			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	9.57	49.26	191.88	85.52
Construction	43.20	62.65	339.15	195.21
Trade	40.30	35.30	395.50	78.65
Transport	11.58	227.02	597.80	358.49

Source: Golata (2011).

To assess the composite estimation one can use REE. This measure is based on estimates of MSE, which can be compared with its 'real' value, thus accounting for estimation precision and bias. The GREG and EBLUP estimators yielded similar estimates for each of the PKD sections. A significant improvement in estimation precision was observed. For *manufacturing*, where the best results were obtained, REE is at 0.3 % of the 'real' value. The bias of the GREG estimator is considerably lower than that of the EBLUP estimator, which often yields better general results owing to its lower variance. In the case of the *transport* section, however, none of the estimators used produced better results than those obtained by means of direct estimation.

Estimation of *revenue* by PKD section and regions (64 domains in all)

Owing to limited space, the results were confined to the expected value of revenue for two PKD sections. Additionally, Figures 4 (*manufacturing*) and 5 (*construction*) depict differences in the expected value of estimators and the 'real' values. The resulting discrepancies are obvious, given the nature of available data and the method used, but they are largely compatible with the 'real' values.

[1] In simulation survey the approximate value of MSE estimate was computed using the following formula presented by Choudhry, Rao, 1993 p. 276).

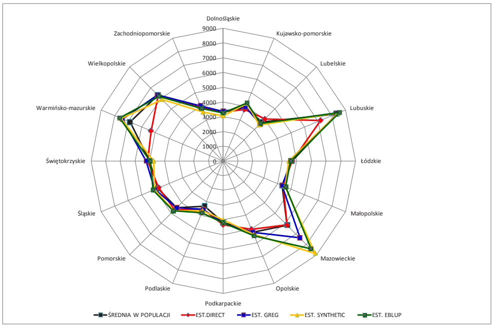


Figure 4: *Expected value of estimators for revenue, manufacturing by voivodship, 2008 Source: Golata (2011).*

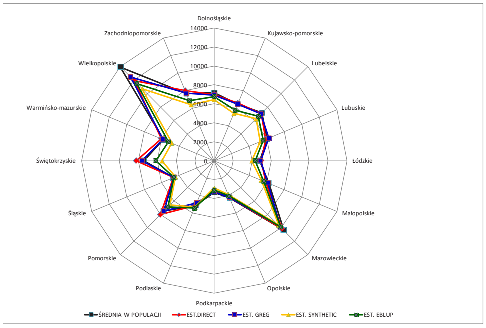


Figure 5: *Expected value of estimators for revenue, construction by voivodship, 2008 Source: Golata (2011).*

Table 5. *REE of estimators for revenue in the construction section by voivodship, 2008*

Voivodship	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Dolnośląskie	32,09	19,79	17,02	9,25
Kujawsko-pomorskie	40,01	15,49	23,71	14,08
Lubelskie	42,32	18,34	20,47	13,85
Lubuskie	70,40	21,34	21,93	11,31
Łódzkie	42,68	18,56	28,84	14,56
Małopolskie	53,21	14,27	22,15	12,68
Mazowieckie	54,81	20,02	13,77	9,01
Opolskie	56,66	22,50	30,17	17,60
Podkarpackie	39,10	18,79	39,15	23,01
Podlaskie	58,30	73,16	22,77	19,41
Pomorskie	91,56	19,28	24,54	18,47
Śląskie	29,52	17,92	24,65	11,71
Świętokrzyskie	136,00	34,22	29,27	25,34
Warmińsko-mazurskie	43,70	12,70	25,19	14,78
Wielkopolskie	106,50	27,77	24,94	24,76
Zachodniopomorskie	54,24	19,28	21,37	13,22

Source: Golata (2011).

Measures of precision in tab. 5 show an evident improvement in efficiency due to the use of indirect estimation and auxiliary data from administrative databases.

Synthetic assessment of estimates for all domains by section

When the Relative Estimation Error (REE, tab. 6) is chosen as a measure of precision, accounting for both precision and bias with respect to the 'real' values in the MEETS real dataset, one can observe an interesting tendency. The use of indirect estimation based on auxiliary information from administrative databases contributes significantly to the improvement in estimation precision in the case of such variables as *revenue*, *number of employees* and *wages*. This improvement can be as much as 50% of the REE obtained by applying direct estimation.

Table 6. *Mean REE for all domains by section, 2008*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Mean REE for all domains (%)				
Revenue	1.62	0.87	1.04	0.81
Number of employees	0.73	0.23	0.34	0.23
Wages	0.70	0.43	0.49	0.39
weighted mean REE for all domains (%)				
Revenue	1.30	0.57	0.90	0.55
Number of employees	0.51	0.18	0.30	0.18
Wages	0.55	0.37	0.50	0.37

Source: Golata (2011)

Synthetic assessment of estimates for all domains by section and voivodship

When estimation is conducted at a lower level of aggregation, one can generally expect a decrease in estimation precision. That was also the case this time. Values of REE, used as a measure of precision with respect to such variables as *revenue*, *number of employees* and *wages*, indicate a significant improvement in comparison with direct estimation (tab. 7). The lower values of REE (a decrease from 35.5% to 13.6% (*Wages*) or from 24.7% to 6.6% (*Number of employees*) obtained as a result of using administrative register data is promising.

Table 7. *Mean REE for all domains by section and voivodship, 2008*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Mean REE for all domains (%)				
Revenue	64.25	54.63	37.14	41.87
Number of employees	24.66	12.14	6.27	6.59
Wages	35.54	25.73	14.38	13.60

Table 7. *Mean REE for all domains by section and voivodship, 2008 (cont.)*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	EBLUP
weighted mean REE for all domains (%)				
Revenue	53.66	26.26	25.73	19.30
Number of employees	15.64	7.50	4.37	4.50
Wages	24.89	17.50	13.00	11.35

Source: Golata (2011)

Finally, the use of weights accounting for the significance of large and medium enterprises has an evident effect on the combined assessment of estimation precision.

4. Empirical evaluation of SDE for linked data - integrating two sample data – simulation study

The second simulation study referred to situation when data from two samples were integrated. It was based on a realistic population. A pseudo-population using real data form Polish micro-census 1995 was constructed. The pseudo-population was called POLDATA and consists of 2 000 000 individuals 15 years or older grouped into 16 strata[1]. But due to time-consuming calculations, for the purpose of this experiment, the pseudo-population was restricted only to three strata, which refer to the following three voivodships: Dolnoslaskie, Kujawsko-pomorskie and Wielkopolskie thus finally consisted of 374 374 individuals. This pseudo-population was the basis on which the sampling procedure was applied.

The study was aimed at estimation of labour market status for NTS3 as domains. Precisely the characteristics to be estimated was the employment rate defined as the percentage of employed population 15 years and older. Therefore dataset ***A*** could be compared to Labour Force Survey[2] (LFS), which due to small sample size does not yield estimates for local labour market (NTS3). Dataset ***B*** is much larger in terms of the number of records, but unfortunately does not include all variables important in the labour market analysis. Lack of these variables prevents construction of the model, which according to previous experience, could be used to estimate the necessary characteristics. This scarcity can be removed by adding variables observed in dataset ***A*** to dataset ***B***.

The decision as to which file should be the donor or the recipient depends on the character of the study. In one approach, the file with more records is treated as a recipient, to prevent a loss of information (Raessler, 2002). Other Authors have pointed out that duplication of information from a smaller set to larger raises

[1] The number of voivodships in Poland.

[2] The Labour Force Survey was not used in the experiment, but sample **A** was constructed to resemble the LFS and the sample was drawn in a similar way. Sample of type A, though small, containing data for many variables, represents relatively comprehensive characteristics of the population in task. It resembles Polish LFS, in which samples cover about 0,05% of the population aged 15 years and more.

risk of duplication, and thus distorts the distribution (Scanu, 2010). Both situations could be considered. The smaller dataset being the recipient file and the larger as donor, seems even more realistic in SDE, especially when making use of administrative records.

The study was conducted according to the following schema:

1. Two types of random samples were drawn from the POLDATA in 100 replicates:
   a. Sample type ***A*** were drawn using two stage stratified sampling design with proportional allocation[1]. The strata were defined as voivodships (NTS2) **-** according to the territorial division of the country. The primary stage units were defined as communes – gminas (NTS5) and on second stage individuals were chosen. On the second stage the simple random sampling without replacement (SRS) was applied. The overall sample size equalled to about 1%.
   b. Sample type ***B*** were drawn with stratified proportional sampling. Similarly as for sample type ***A***, voivodships were defined as strata and then 5% SRS was implemented.
2. The following variables were considered:

   AREA VARIABLES:

   i. NUTS 2 – Voivodship – 3 categories
   ii. NUTS 3 – 11 units

   AGE – 3 categories:

   0 = less than 30 1 = 30 **-** 44 2 = 45 and over

   GENDER – 2 categories:

   0 = male 1= female

   CIVIL STATUS – 3 categories:

   0 = divorced or widowed 1= married 2 = single

   PLACE OF RESIDENCE – 3 categories

   0 = rural areas and towns of less than 2 thousands 1= town 2 – 50 thousands 2 = town 50 thousands and over

   EDUCATION LEVEL – 4 categories:

   0 = university 1= elementary 2 = vocational 3 = secondary

   LABOUR MARKET STATUS – 4 categories:

   0 = unemployed 1= employed 2 = economically inactive

   a. Samples of type A contained all the variables listed above
   b. Samples of type B missed information about education level

[1] The sampling procedure was not exactly the same as in case of LFS, but also follows the two-stage household sampling. Sampling scheme for the LFS defines census units called census clusters in towns or enumeration districts in rural areas, as the primary sampling units subject to the first stage selection. Second stage sampling units are dwellings.

3. Beginning with this step, the following estimation procedures were conducted:
    a. The two random samples ***A*** and ***B*** were matched. One of the simplest but also most frequently used nonparametric procedure for statistical matching based on *k* nearest neighbours[1] was applied (*k*NN). And the estimation procedure used weights according to Rubin (1986)
    b. The two random samples ***A*** and ***B*** were matched using the *k*NN and the estimation procedure applied special weights calibrated according to domains defined for estimation
4. To the linked data the EBLUPGREG program was applied and in each run the following estimates of economic activity for local labour market (domains defines as NTS3) were obtained:
    a. DIRECT
    b. GREG
        i. upon Sample B with no education - referred to as 'no education' approach 'NE'
        ii. upon Sample B with education matched and Rubin's weights approach - referred to as 'imputed education' approach 'IE'
        iii. upon Sample B with education matched and calibration weights approach - referred to as 'imputed education and calibration' approach 'CIE'
    c. SYNTHETIC
        i. upon Sample B with no education - 'NE'
        ii. upon Sample B with education matched and Rubin's weights approach - 'IE'
        iii. upon Sample B with education matched and calibration weights approach - 'CIE'
    d. EBLUP
        i. upon Sample B with no education - 'NE'
        ii. upon Sample B with education matched and Rubin's weights approach - 'IE'
        iii. upon Sample B with education matched and calibration weights approach - 'CIE'
5. The estimates obtained in each run were used to provide the empirical evaluation of the estimation precision with reference to real 'population' value:
    a. Empirical variance
    b. Empirical bias
    c. Empirical REE

The integration algorithm

Since both databases were samples, they most probably did not contain data about the same person, nor they had a unique linkage key. Consequently, such

[1] As *k = 1*, the imputation method was reduced to distance hot deck.

data sources could not be integrated using the deterministic approach. In order to achieve the desired objective, statistical matching was implemented. The integrating algorithm usually may be broken down into 6 basic steps (D'Orazio, Di Zio, Scanu (2006)):

1. Variable harmonisation
2. Selection of matching variables and their standardization or dichotomization
3. Stratification
4. Calculation of distance
5. Selection of records in the recipient and donor datasets with the least distance
6. Calculation of the estimated value of variables

The harmonization of variables involves adjusting of definitions and classifications used in both 'surveys': dataset ***A*** and dataset ***B***. The fact that in the simulation conducted both samples were drawn from the same pseudo-population, allowed us to skip the harmonization step. But the importance of these procedures should be stressed.

The second stage was selecting the matching variables to estimate the measure of similarity between records. In our case the following variables were selected: gender, age, marital status and place of residence. As this set of variables includes categorical as well as quantitative variables their standardization and dichotomization was necessary. So the qualitative variables were transformed into binary ones. The quantitative variable: age was categorized and dichotomized as well.

The third step was to stratify. The strata was created on the basis of two variables: NUTS3 and labour market status. There was eleven NUTS3 subregions in the population but due to small number of units two of them were merged. Altogether there were 27 strata created: 9 subregions (NUTS3 regions 3 and 4, and also 41 and 42 were merged) x 3 attributes of the employment status (employed, unemployed, economically inactive). An important reason for stratifying the dataset was to optimize the computing time [1].

The measure of record similarity used in the integration was the Euclidean Squared Distance given by the formula:

$$d_{A,B} = \sum_{i=1}^{N} \sum_{k=1}^{K_i} (a_{Aik} - a_{Bik})^2 \qquad (2)$$

where:

a_{Aik} – binary variables created in the process of dichotomization of qualitative variables (*i*-th category of *k*-th variable).

[1] In spite of dividing the data set into strata, duration of the integration process amounted to about 6 hours (Intel Core i5 processor, 4 GB RAM).

For a given record in recipient file, the algorithm searches for a record in donor file for which the distance measure is the smallest. The choice of Euclidean Squared Distance was motivated by the use of the integration algorithm developed by Bacher (2002). The algorithm was modified and adjusted for purposes of the simulation. The study was performed under conditional independence assumption (CIA). The integration algorithm yielded a dataset containing 18 715 records (the number of records in Sample B - the larger one) and 7 variables describing the demographic and economic characteristics of Polish population as listed above[1].

Rubin approach

Survey data for estimation or integration process generally are drawn from population according to complex sampling schema. When this is the case, it is necessary to adjust sampling weights in estimation process. There are three different approaches: file concatenation proposed by Rubin (1986), case weights calibration (Renssen, 1998) and Empirical Likelihood according to Wu (2004).

Rubin (1986) suggested to combine the two files ***A*** and ***B*** into ***AB*** and calculate new weight $\boldsymbol{w_{AB}}$ for each *i*th unit in the new file (with some corrections). If the *i*th unit in the sample ***A*** is not represented in sample ***B***, than its inverse probability equals to zero (under sampling schema ***B***). In such case weight of this unit in the concatenated file ***AB*** is simply its weight from sample ***A*** **-** $\boldsymbol{w_{Ai}}$**.** This means not only that the population in task is the union of $\boldsymbol{A \cup B}$, but also that the estimated distributions are conditional of ***Y*** given $\mathbf{(w_{AB}\ ;\ Z)}$ and **Z** given $\boldsymbol{(w_{AB}\ ;\ Y)}$**.**

In our study the file ***A*** was not concatenated to file ***B***. The integration process to join ***A*** and ***B*** was to impute in ***B*** originally unobserved variables ***Z*** that characterize the level of education by using the values of ***X***, which were observed in both files. Thus, as suggested by Rubin, the weight of each observation in the set ***B*** remained unchanged.

Calibration approach

When samples are drawn according to different complex survey designs it is important to consider the weights to preserve the distribution of the variable in task. Especially when the survey is originally planned for the whole population and finally the estimation is conducted for unplanned domains.

The impact of sampling designs for the efficiency in small area estimation is a question difficult to answer due to many optimisation problems. According to Rao J.N.K., (2003) most important design issues for small domain estimation are such as: number of strata, construction of strata, optimal allocation of a sample, selection probabilities. This list can be enlarged by definition of optimisation criteria, availability of strongly correlated auxiliary information, choice of

[1] All programming and calculations was made by W. Roszka in the Department of Statistics at the Poznan University of Economics.

estimators and so on. In practice it is not possible to anticipate and plan for all small areas. As a result indirect estimators will always be needed, given the growing demand for reliable small area statistics. However, it is important to consider design issues that have an impact on small area estimation, particularly in the context of planning and designing large-scale surveys (Sarndal et al 1992).

According to Särndal (2007) calibration is a method of estimating the parameters for the finite population, which applies new "calibration" weights. The calibration weights need to be close to the original ones and satisfy the so-called calibration equation. Applying calibration weights to estimate parameters of the target variable is especially needed in case of no occurrence, no response or other non-sampling errors to provide unbiased estimates[1]. These weights may also take into account relation between the target variable and an additional one to adjust the estimates to the relation observed at global level. Therefore the GREG estimator is widely used in SDE. Additionally we proposed to verify the impact of calibration weights taking into account all the matching variables to adjust the estimates for domains.

Suppose that the objective of the study is to estimate the total value of a variable, defined by the formula (Szymkowiak 2011):

$$Y = \sum_{i=1}^{N} y_i , \tag{3}$$

where y_i denotes the value of variable y for i - th unit, $i = 1, \dots, N$.

Let us assume that the whole population $U = \{1, \dots, N\}$ consists of N elements. From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements. Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = \frac{1}{\pi_i}$ the design weight. The Horvitz-Thompson estimator of the total is given by:

$$\hat{Y}_{HT} = \sum_{s} d_i y_i = \sum_{i=1}^{n} d_i y_i . \tag{4}$$

Small sample size might cause unsufficient representation[2] of particular domains in the sample, and therefore enable direct estimates. If information for the variable y is not known for some domains then the Horvitz-Thompson estimator would be characterised of high variance.

[1] Calibration approach as a method of nonresponse treatment is described in detail in Särndal C–E., Lundström S. (2005) *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd.

[2] In practice it might occur, that the domain is even not represented in the sample. In our simulation study such situation was not considered.

Proper choice of the distance function is essential for constructing calibration weights and the results obtained. In our study the distance function was expressed by the formula which allows to find the calibration weights in an explicit form:

$$D(\mathbf{w},\mathbf{d})=\frac{1}{2}\sum_{i=1}^{m}\frac{(w_i-d_i)^2}{d_i}, \tag{5}$$

Effective use of calibration weights w_i depends on the vector of auxiliary information. Let $x_1,\dots,x_k$ denote auxiliary variables which will be used in the process of finding calibration weights. In our simulation study we used calibration weights obtained for each domain using additional information from the pseudo-population. As auxiliary data the following variables were used: gender, KLM, education, age, marital status and labour market status. Let:

$\mathbf{X}_j=\sum_{i=1}^{N}x_{ij}$, denote total value for the auxiliary variable x_j, $j=1,\dots,k$, (6)

where x_{ij} is the value of j-th auxiliary variable for the i-th unit

$$\mathbf{X}=\left(\sum_{i=1}^{N}x_{i1},\sum_{i=1}^{N}x_{i2},\dots,\sum_{i=1}^{N}x_{ik}\right)^T \tag{7}$$

is known vector of population totals for of auxiliary variables.

The vector of calibration weights $\mathbf{w}=(w_1,\dots,w_m)^T$ is obtained as the following minimization problem:

$$\mathbf{w}=\text{argmin}_{\mathbf{v}}D(\mathbf{v},\mathbf{d}), \tag{8}$$

subject to the calibration constraints

$$\mathbf{X}=\tilde{\mathbf{X}}, \tag{9}$$

where

$$\tilde{\mathbf{X}}=\left(\sum_{i=1}^{m}w_ix_{i1},\sum_{i=1}^{m}w_ix_{i2},\dots,\sum_{i=1}^{m}w_ix_{ik}\right)^T . \tag{10}$$

If the matrix $\sum_{i=1}^{m}d_i\mathbf{x}_i\otimes\mathbf{x}_i^T$ is nonsingular then the solution of minimization problem (8), subject to the calibration constraint (9) is a vector of calibration weights $\mathbf{w}=(w_1,\dots,w_m)^T$, whose elements are described by the formula:

$$w_i=d_i+d_i\left(\mathbf{X}-\hat{\mathbf{X}}\right)^T\left(\sum_{i=1}^{m}d_i\mathbf{x}_i\otimes\mathbf{x}_i^T\right)^{-1}\mathbf{x}_i \tag{11}$$

where

$$\hat{\mathbf{X}}=\left(\sum_{i=1}^{m}d_ix_{i1},\sum_{i=1}^{m}d_ix_{i2},\dots,\sum_{i=1}^{m}d_ix_{ik}\right)^T \tag{12}$$

and

$$\mathbf{x}_i = \left(x_{i1},\ldots,x_{ik}\right)^T \tag{13}$$

is the vector consisting of values of all auxiliary variables for the i-th respondent $i = 1,\ldots,m$.

Assessment of data integration

In the literature there are different approaches to assess matching quality. Raessler (2002) proposed to assess the two files as well matched if they meet the criteria for the distribution compliance and preservation of relations between variables in the initial and matched files. The four criteria specified by Rassler are: (i) the true, unknown distribution of matched variables Z is reproduced in the newly created, synthetic file; (ii) the real, unknown cumulative distribution of the variables (X,Y,Z) is maintained in the newly created, synthetic dataset; (iii) correlation and higher moments of the cumulative distribution of (X,Y,Z) and a marginal distribution of $(X-Y)$ and $(X-Z)$ are preserved; (iv) at least marginal distributions of Z and $(X-Z)$ in the fused file are preserved. In practice it might be difficult, or sometimes even impossible to verify all those criteria (D'Orazio,2010). Also the statistical inference methods are not always suitable, especially in case of administrative data.

Table 8. Statistical **c**haracteristics of the number of matches

Statistical characteristics of the number of matches (together with no-matched records)						
Over all samples	Mean	Std	Median	Mode	Min	Max
MIN	3,80	5,12	2	0	0	49
Q1	4,48	6,12	2	0	0	78
Q2	4,95	6,80	3	0	0	115
Q3	5,35	7,68	3	0	0	171
MAX	6,39	9,48	4	0	0	288
Statistical characteristics of the number of matches (no-matched records omitted)						
Over all samples	Mean	Std	Median	Mode	Min	Max
MIN	5,64	5,34	4	1	1	49
Q1	6,60	6,41	5	1	1	78
Q2	6,99	7,18	5	1	1	115
Q3	7,53	8,21	5	2	1	171
MAX	8,54	10,63	6	4	1	288

Source: own calculations.

In the simulation process the mean number of matches over all samples equalled to 3,8 for all records and 5,64 if the no-matched records were omitted (tab. 8). And the highest number of matches amounted to 8,54 (no-matched records omitted).

In the study the following quality assessment measures were used:

- total variation distance (D'Orazio, Di Zio, Scanu, 2006):

$$\Delta(p_f, p_d) = \frac{1}{2}\sum_{i=1}^{I}|p_{f,i} - p_{d,i}| \qquad (14)$$

- Bhattacharyya coefficient (Bhattacharyya, 1943):

$$BC(p_f, p_d) = \sum_{i=1}^{I}\sqrt{p_{f,i} \times p_{d,i}} \qquad (15)$$

where:

$p_{f,i}$ – proportion of i-th category of a variable in the fused file,

$p_{d,i}$ - proportion of i-th category of a variable in the donor file.

Both of these coefficients are in the range of $\langle 0,1 \rangle$. In case of total variation distance, the lower Δ coefficient, the greater distribution compatibility is achieved. The value indicating the acceptable similarity of distributions is commonly assumed as $\Delta \leq 3\%$. Conversely, the lower the value of the Bhattacharyya coefficient, the lower the compatibility of distributions achieved. As the coefficient proposed by Bhattacharyya generally takes high value, two other measures of structure similarity were applied:

$$W_{p1} = \sum_{i=1}^{k}(\min_{fd} p_i) \text{ and } W_{p2} = \frac{\sum_{i=1}^{k}(\min_{fd} p_i)}{\sum_{i=1}^{k}(\max_{fd} p_i)}, \qquad (16)$$

where:

$\min_{fd} p_i$ the minimum proportion of i-th category in the fused and donor file,

$\max_{fd} p_i$ the maximum proportion of i-th category in the fused and donor file.

These coefficients take values from the interval $\langle 0\%, 100\% \rangle$ and W_{p1} is generally greater than W_{p2}. The greater the value of any of these coefficients, the greater the compatibility of the distributions. Values that indicate the acceptable similarity of distributions are usually assumed to be $W_{p1} \geq 97\%$ and $W_{p2} \geq 95\%$ (Roszka 2011).

Table 9. Total variation distance as matching quality measure

Matching variable	Place of residence	Gender	Marital Status	Source of maintenance
MIN	0,0830	0,0000	0,0070	0,0040
Q1	0,1528	0,0030	0,0129	0,0150
Q2	0,1790	0,0050	0,0160	0,0198
Q3	0,2201	0,0100	0,0221	0,0245
MAX	0,2920	0,0270	0,0405	0,0370

Source: own calculations.

Table 10. Bhattacharyya coefficient as matching quality measure

Matching variable	Place of residence	Gender	Marital Status	Source of maintenance
MIN	0,9355	0,9996	0,9976	0,9978
Q1	0,9607	0,9999	0,9988	0,9991
Q2	0,9691	1	0,9993	0,9995
Q3	0,9769	1	0,9996	1
MAX	0,9916	1	1	1

Source: own calculations.

Very good matching quality coefficients were achieved for the variables “gender”, “marital status” and “source of maintenance”. Much worse quality measures were obtained for the variable “place of residence (tab. 9 and 10). This results from the fact that “class of place of residence” variable was characterized by a weaker compatibility prior to integration.

The similarity coefficients presented in tab. 9 and 10 characterise the matching quality in a synthetic way. That is, over all replications and additionally, they do not take into account differences of distributions across domains. Compatibility of the distributions observed for the whole sample, of course, do not translate automatically to all domains for which estimation of economic activity was conducted in the next stage. The discrepancy in the compliance applies to both individual samples and domains. Typically, in the conformity assessment distribution of matching variables is taken into account. In case of a simulation study, there was also the possibility to evaluate distribution of the matched variable.

Comparability of the distributions for the variable in task „education” showed that the distributions were preserved. Table 11 provides the comparison of education distribution by domains in population with direct estimates upon one exemplary sample after matching variable education. The Bhattacharyya coefficient is generally close to one, on average greater than 0.99. Only for domain 42, it takes value lower than 0.95 (in red colour). For this specific domain also the other two similarity coefficients take exceptionally low values. But their more detailed analysis indicates that the education distribution is well maintained only for three domains (number: 1, 6 and 41). The results presented refer to the situation when originally sampling weights were applied. In case of weights calibrated for domains, the distributions were identical.

Table 11. Education distribution by regions in population and direct estimates upon exemplary sample with matched variable

NTS3	Proportion of population of the following education level								$BC(p_f;p_d)$	W_{p1}	W_{p2}
	Exemplary sample*				Population						
	Elementary	Vocational	Secondary	University	Elementary	Vocational	Secondary	University			
1	0,47	0,27	0,20	0,06	0,45	0,28	0,21	0,06	**0,9997**	**0,976**	**0,954**
2	0,55	0,16	0,24	0,05	0,43	0,29	0,22	0,06	**0,9872**	**0,867**	**0,765**
3	0,54	0,19	0,18	0,08	0,47	0,30	0,18	0,04	**0,9900**	**0,894**	**0,808**
4	0,25	0,16	0,41	0,18	0,29	0,19	0,34	0,19	**0,9967**	**0,923**	**0,857**
5	0,49	0,29	0,16	0,05	0,42	0,31	0,20	0,06	**0,9970**	**0,929**	**0,867**
6	0,50	0,28	0,16	0,06	0,49	0,26	0,19	0,06	**0,9994**	**0,971**	**0,944**
38	0,51	0,26	0,21	0,03	0,48	0,29	0,19	0,05	**0,9980**	**0,952**	**0,908**
39	0,46	0,33	0,16	0,06	0,42	0,33	0,19	0,06	**0,9988**	**0,961**	**0,925**
40	0,46	0,34	0,13	0,07	0,43	0,30	0,20	0,06	**0,9944**	**0,924**	**0,858**
41	0,52	0,25	0,17	0,05	0,51	0,25	0,19	0,05	**0,9998**	**0,984**	**0,969**
42	0,54	0,20	0,20	0,07	0,24	0,24	0,34	0,18	**0,9467**	**0,705**	**0,545**
All domains	**0,49**	**0,27**	**0,18**	**0,06**	**0,44**	**0,28**	**0,21**	**0,07**	**0,9990**	**0,956**	**0,916**

* The first sample was compared

Source: Own calculations

Comparability of the distributions for the variable in task „education” showed that the distributions were preserved. Table 11 provides the comparison of education distribution by domains in population with direct estimates upon one exemplary sample after matching variable education. The Bhattacharyya coefficient is generally close to one, on average greater than 0.99. Only for domain 42, it takes value lower than 0.95 (in red colour). For this specific domain also the other two similarity coefficients take exceptionally low values. But their more detailed analysis indicates that the education distribution is well maintained only for three domains (number: 1, 6 and 41). The results presented refer to the situation when originally sampling weights were applied. In case of weights calibrated for domains, the distributions were identical.

Domain Specific Evaluation of Estimation Precision

Assessing the quality of the estimates from domain specific perspective, one can take into account both: single sample and average values for each domain upon 100 replications. The results obtained for estimators used in the study for different research approaches: with imputed education and calibrated weights, are rather extensive. Therefore, due to the limited scope, this article describes only selected results. The exemplary estimates obtained for domain 1 in each of 100 replicates are shown in fig 6.

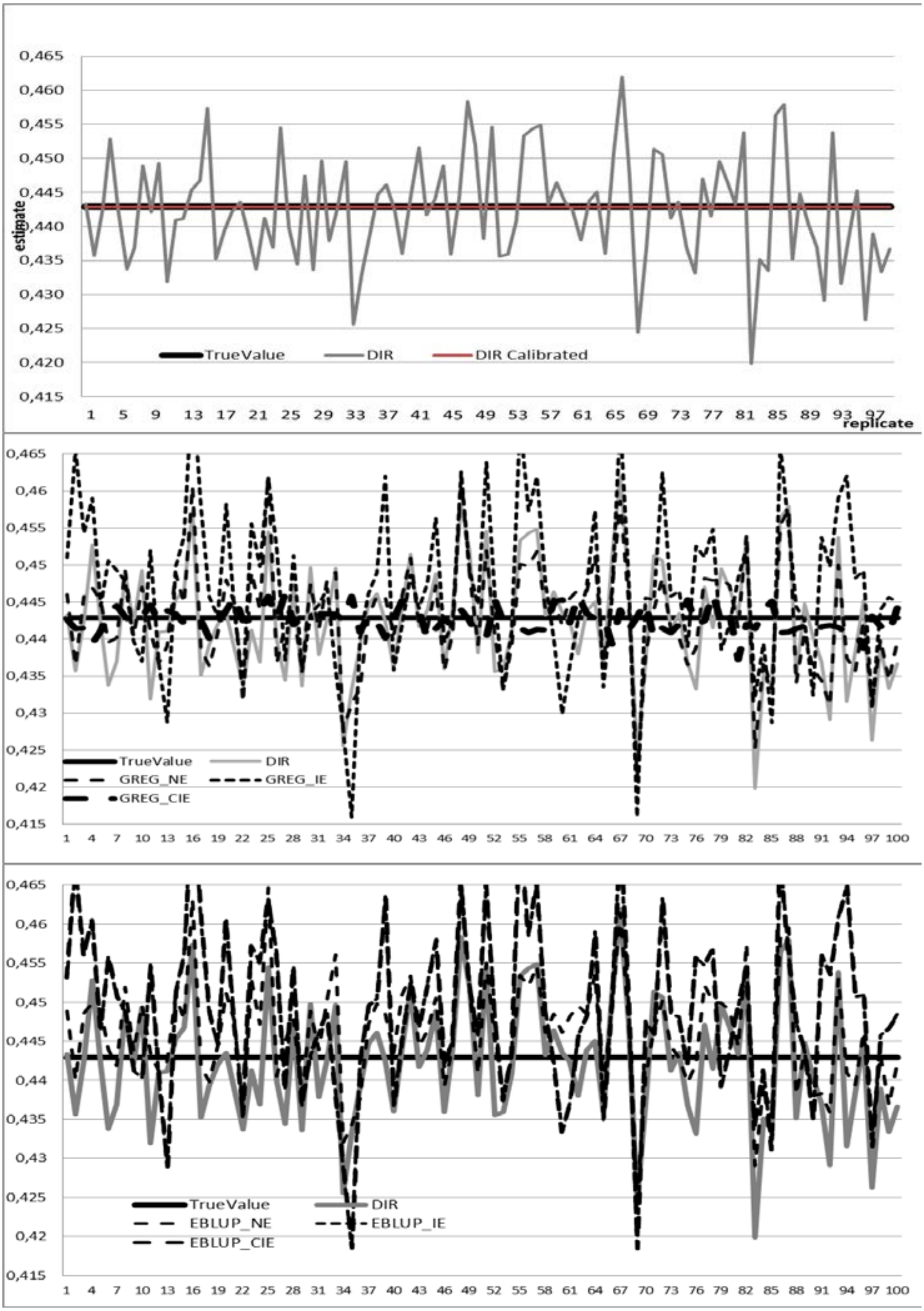


Figure 6: *Estimates of the percentage of economically active, different estimators and research approaches, Domain 1*

Source: Own calculations.

And fig. 7 represents expected values of the one selected estimator (EBLUP) for different approaches by domains.

First, it could be noticed that calibrated weights applied to direct estimator gave the 'true' value in each replicate. As concerns the GREG estimator, the one with imputed education and calibrated weights resulted in estimated close to the 'true value' in all replicates. The variation of the estimates was also small. Combining GREG with synthetics estimator resulted in a considerable increase in EBLUP estimates variation, even in comparison with direct estimator.

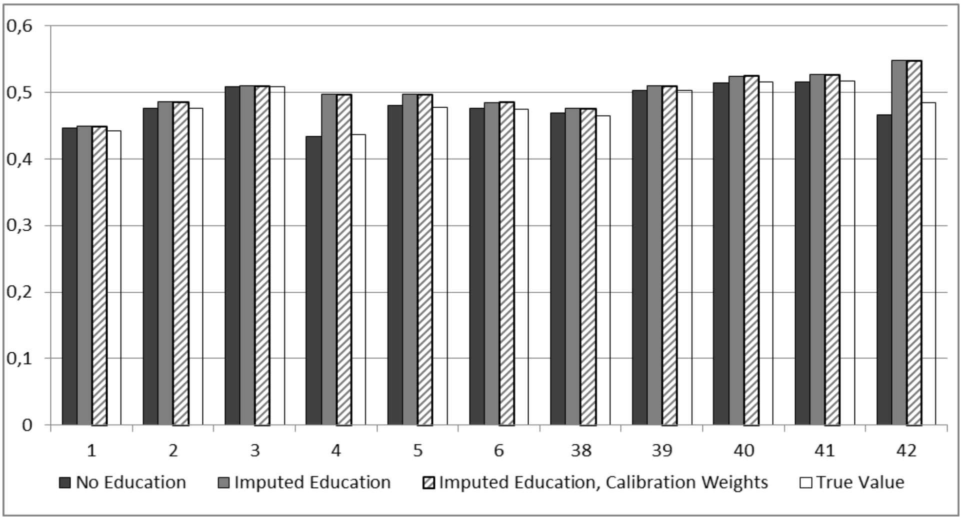


Figure 7: *Expected value of the EBLUP estimator for different approaches by domains,*

Source: Own calculations.

It is worth to noticed that thanks to the simulation approach, the results discussed could be analysed with reference to the 'true' value, which usually is unknown. Another reference values might constitute the estimates obtained model including education or not (fig. 7). No matter which reference value would be chosen, the estimates taking into account the imputed education are on average clearly overestimated in two domains (4 and 42). These results confirm need for careful evaluation of integration process and convergence of the distribution of all variables, especially those exploit as auxiliary.

Synthetic Evaluation of estimation precision over all domains

Assessing the estimation precision over all domains average values of mean and relative estimation errors (MSE and REE) obtained for different research approaches were analysed.

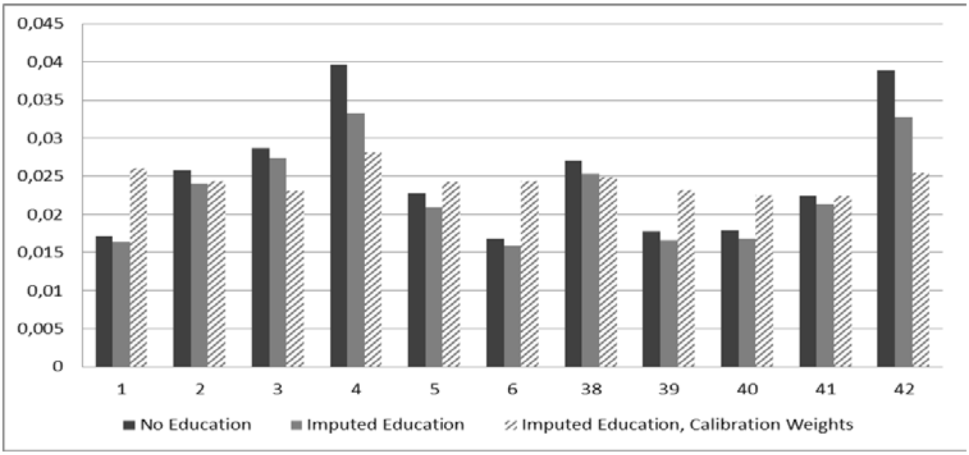


Figure 8: *REE(GREG) for different research approaches by domains*
Source: Own calculations.

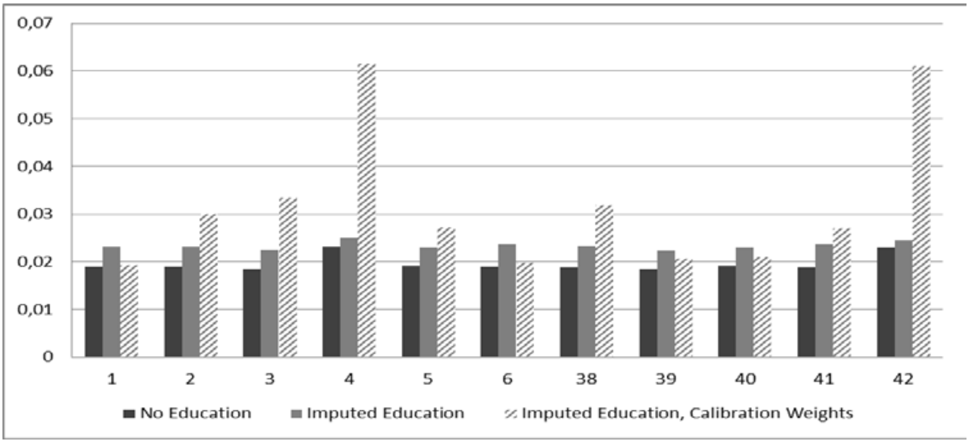


Figure 9. *REE(SYNTH) for different research approaches by domains*
Source: Own calculations.

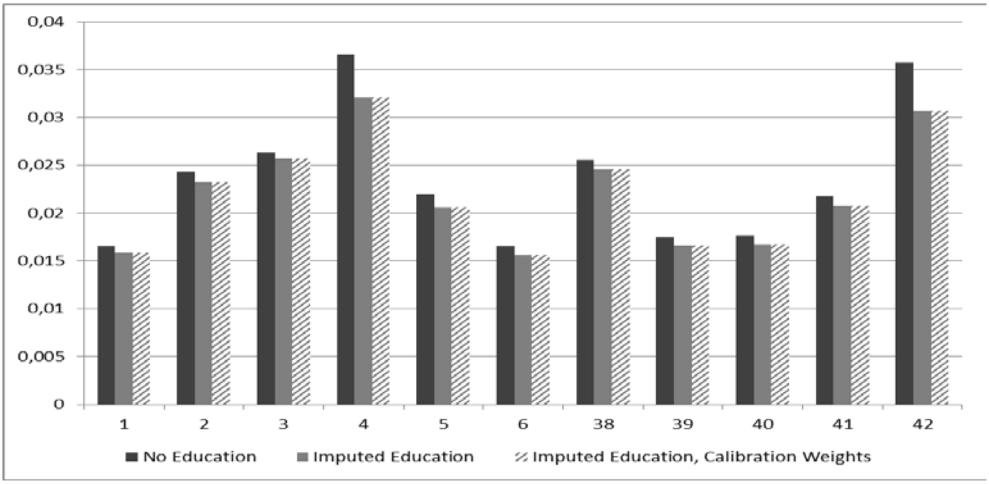


Figure 10: *REE(EBLUP) for different research approaches by domains*

Source: Own calculations.

As it comes from presentation of relative estimation error for GREG and EBLUP estimators across all domains: estimates including imputed education improve precision obtained (red and yellow bars on fig. 8 and 10). Of course, this statement should not be generalised, as in case of SYNTH estimator, the presented results indicate just an opposed opinion (for each domain, fig. 9).

As the main issue in the study was to evaluate the estimates for linked data, the results obtained for samples with real education, were considered for reference purposes (presented in grey in tables 11 and 12). However results obtained for samples with imputed education included in the model (with original or calibrated weights) might also be compared to the ones with no education, as this reflects more realistic situation.

Table 11. MSE for different estimators and research approaches

Research approach	Type of estimator							
	DIR	GREG	SYNTH	EBLUP	DIR	GREG	SYNTH	EBLUP
	Average of MSE over all domains				Weighted average of MSE over all domains			
Education	0,0136	0,0115	0,0082	**0,0108**	0,0117	0,0099	0,0081	0,0094
No Education	0,0136	0,0120	0,0094	0,0113	0,0117	0,0103	0,0093	0,0099
Imputed Education	0,0136	0,0115	0,0117	**0,0111**	0,0117	0,0098	0,0116	**0,0096**
Imputed Education, Calibration Weights	0,0154	0,0131	0,0117	**0,0111**	0,0125	0,0106	0,0116	**0,0096**

Source: Own calculations.

Table 12. REE for different estimators and research approaches

Research approach	Type of estimator							
	DIR	GREG	SYNTH	EBLUP	DIR	GREG	SYNTH	EBLUP
	Average of REE over all domains				Weighted average of REE over all domains			
Education	0,0282	0,0239	0,0171	0,0223	0,0242	0,0205	0,0169	0,0196
No Education	0,0282	0,0248	0,0195	0,0235	0,0242	0,0213	0,0191	0,0205
Imputed Education	0,0282	0,0229	0,0234	**0,0221**	0,0242	0,0199	0,0232	**0,0194**
Imputed Education, Calibration Weights	0,0318	0,0273	0,0234	**0,0221**	0,0259	0,0220	0,0232	**0,0194**

Source: Own calculations.

Similarly as in the simulation study for business statistics, weighting the measures of estimation precision with domain size, indicates on average higher quality assessment. It could be also noticed that estimators for small domains perform typically for linked data equally as for real data. The precision depends on the relation of matched variable and the estimated one. In presented study including imputed education into the model slightly improved estimates of the percentage of economically active population.

5. Conclusions

Data Integration is used to combine information from distinct sources of data which are jointly unobserved and refer to the same target population. Fusing distinct data sources to be available in one set enables joint observation of variables from both files. The integration process is based on finding similar records and the similarity is calculated on the basis of common variables in both datasets. Similarity of the idea concerning small domain estimation and data integration techniques could be specified as follows[1]:

1. Auxiliary information. Both techniques refer to external data sources:
- SDE in order to obtain auxiliary variable that can help to improve estimation precision for domains
- DI to provide more comprehensive data sets which allow for reducing the respondents burden and bias resulting.

Joint application of both methods might result in increasing both: estimation precision and the scope of information available, especially in the context of small domains. But estimates on linked data require good matching quality:

- method for data integration
- direct measure of consistency of the distribution of matched variable is needed

[1] This specification is of course, should not be considered as full and final.

- earlier constrains help to avoid improper values
- micro integration processing
- calibration might be considered as a method for adjusting sample design to estimates for unplanned domains.

2. Correlation and regression. The two data sources are combined upon in-depth correlation analysis:
- in SDE by model-based estimation for domains
- in DI this correlation is crucial in the matching process for a) common matching variables and for b) 'imputed' - jointly unobserved variable '*Z*'.

Taking the above into account, in both groups of methods, variable harmonisation is important. This involves not only definition of the variables, grouping and classification issues, but also designation of statistical units and resulting aggregation level for the analysis. Thus, the danger of so called 'ecological fallacy' or 'ecological error' appears.

When studying the relationship between variables that are specified for different territorial units, or at different levels of aggregation, the concept of ecological error should be understood as taking the relationship observed at a higher level of aggregation, as also valid at a lower level. In practice, estimates for small areas frequently used regression estimators, assuming tacitly that the true values of the parameters (β) in the regression equation at the level of individual units are the same as for the parameters obtained from the mean values for the spatial units (Heady and Hennel, 2002, p. 5). But empirical results show significant differences. Typically, the correlations obtained at the aggregate level are much stronger than the ones obtained for the individual units. This discrepancy in statistics is called ecological or environmental effect illusion (ecological fallacy). The possibility of recognizing a variety of statistical units brings methodological problem, namely how to estimate the relation for a number of levels simultaneously. Application of the mixed models might be considered as one of the solutions suggested to solve the problem and avoid the 'ecological effect'.

It should be stressed that the success of any model-based method depends on distributions of estimated variables and covariates, correlation analysis – choice of good predictors of the study variables and model diagnostic.

3. Sampling design. Often the two data sets are obtained from independent sample surveys of complex designs, this raises a number of methodological problems:
- in SDE with providing the sampling schema that would be optimal in estimation for domains and in assessing precision of the estimates. According to Rao J.N.K. (2003) most important design issues for small domain estimation are the following: number of strata, construction of strata, optimal allocation of a sample, selection probabilities. This list can be enlarged by adding the problem of defining the optimisation criteria, possibilities in obtaining strongly correlated

auxiliary information, choice of estimators taking into account their efficiency under specific sampling designs.

- in DI the sampling design cannot be ignored and different weights assigned to each sample unit must be considered in order to preserve the population structure and variable distribution. In literature Rubin's file concatenation (1986) or Renssen's calibration (1998) is proposed. Alternatively Wu (2004) suggests empirical likelihood method.

4. Stratification. In both methods stratification has a significant meaning. In SDE where data are drawn from population with no respect to domains for which finally estimation is conducted, post-stratification could be considered as a method of optimization the sampling schema. By introducing stratification in DI we optimize the integration process by reducing the computing time.

5. „Theory & Practice". For both groups of methods it is often observed that situations observed in practice do not correspond to the theoretical solutions. On the basis of the study conducted the following of them could be mentioned:

- High differentiation in correlation across domains between variables estimated on the basis of DG-1 statistical reporting and auxiliary variables from administrative databases, including PIT and CIT
- The non-homogenous distributions of estimated variables and covariate data may imply the need for robust estimation (modified GREG, Winsor and local regression). This solution, however, is connected with the highly complicated and time-consuming estimation techniques
- Administrative problems connected with access to auxiliary data, which limit their usefulness in short-term statistics.

6. Estimates on linked data.

According to Rao (2005), small area estimation is a striking example of the interplay between theory and practice. But he stresses that, despite significant achievements, many issues require further theoretical solutions, as well as empirical verification. Among these issues Rao points primarily on: a) benchmarking model-based estimators to agree with reliable direct estimators at large area levels, b) developing and validating suitable linking models and addressing issues such as errors in variables, incorrect specification of the model and omitted variables, c) development of methods that satisfy multiple goals: good area-specific estimates, good rank properties and good histogram for small areas.

Similarly, Data Integration is becoming a major issue in most countries, with a view to using information available from different sources efficiently so as to produce statistics on a given subject while reducing costs and response burden and maintaining quality. However, the use of DI methods requires not only further theoretical solutions, but also many practical tests. Typically, DI methods seem to be understandable and easy to use, but in practice significant complications occur.

Similarity of both methods should be understood also as a set of common problems requiring further research and analysis that could enable their wider use in official statistics.

REFERENCES

BACHER, J. (2002) Statistisches Matching - Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS, ZA-Informationen, 51. Jg.

BALIN, M., D'ORAZIO, M., DI ZIO, M., SCANU, M., TORELLI, N. (2009) Statistical Matching of Two Surveys with a Common Subset, Working Paper n. 124, Universita Degli Studi di Trieste, Dipartimento di Scienze Economiche e Statistiche.

BRACHA, Cz. (1994) Metodologiczne aspekty badania małych obszarów [Methodological Aspects of Small Area Studies], „Studia i Materiały. Z Prac Zakładu Badań Statystyczno-Ekonomicznych" nr 43, GUS, Warszawa (in Polish).

CHAMBERS, R., SAEI A. (2003) Linear Mixed Model with Spatial Correlated Area Effect in Small Area Estimation.

CHAMBERS, R., SAEI A., 2004, Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects, Southampton Statistical Sciences Research Institute.

CHAMBERS, R.L, FALVEY, H., HEDLIN, D., KOKIC P. (2001) Does the Model Matter for GREG Estimation? A Business Survey Example, in: Journal of Official Statistics, Vol.17, No.4, 527-544.

CHAMBERS, R.L. (1996) Robust case-weighting for multipurpose establishment Surveys in: Journal of Official Statistics, Vol.12, No.1, 3-32.

CHOUDHRY, G.H., RAO, J.N.K. (1993) Evaluation of Small Area Estimators. An Empirical Study, in: Small Area Statistics and Survey Designs, eds G. Kalton, J. Kordos, R. Platek, vol. I: Invited Papers, Central Statistical Office, Warsaw.

D'ORAZIO, M., DI ZIO, M., SCANU, M. (2006) Statistical Matching. Theory and Practice, John Wiley & Sons, Ltd.

DEHNEL, G. (2010) Rozwój mikroprzedsiębiorczości w Polsce w świetle estymacji dla małych domen [Development of micro-business in the light of estimation for small domains], Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan (in Polish).

DEHNEL, G. (2011) Use of Administrative Data for Business Statistics, Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

DEVILLE, J–C. SÄRNDAL, C–E. (1992) Calibration Estimators in Survey Sampling, in Journal of the American Statistical Association, Vol. 87, 376–382.

DI ZIO, M. (2007) What is statistical matching, Course on Methods for Integration of Surveys and Administrative Data, Budapest, Hungary.

Eurarea Project Reference Volume All Parts (2004) The EURAREA Consortium http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/downloads/index.html.

GHOSH, M., RAO J.N.K. (1994) Small Area Estimation: An Appraisal, „Statistical Science" vol. 9, no. 1.

GOŁATA, E. (2009) Opracowanie dla wybranych metod integracji danych reguł, procedur integracji danych z różnych źródeł, [Development of selected methods for data integration rules, procedures, data integration from various sources]. GUS Internal materials, Poznań, Poland (in Polish).

GOLATA, E. (2011) A study into the use of methods developed by small area statistics in: Use of Administrative Data for Business Statistics (pp.84-111), G. Dehnel (ed.), Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

HEADY, P., HENNEL S. (2002) Small Area Estimation and the Ecological Effect – Modifying Standard Theory for Practical Situations, Office for National Statistics, London, IST 2000-26290 EURAREA, Enhancing Small Area Estimation Techniques to Meet European Needs.

HERZOG, T. N., SCHEUREN, F. J., WINKLER, W.E. (2007) Data Quality and Record Linkage Techniques, Springer New York.

KADANE, J.B. (2001) Some Statistical Problems in Merging Data Files, Journal of Official Statistics, No. 17, 423-433.

LEHTONEN, R., VEIJANEN, A. (1998) Logistics Generalized Regression Estimators, Survey Methodology, vol. 24.

MŁODAK, A., KUBACKI, J. (2010), A typology of Polish farms using some fuzzy classification method, Statistics in Transition – new series, vol. 11, No. 3, pp. 615 – 638.

MORIARITY, C., SCHEUREN, F. (2001) Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure in: Journal of Official Statistics, No. 17, 407-422.

Multivariate analysis of systematic errors in the Census 2002, and statistical analysis of the variables of NC 2002 supporting the use of small area estimates. J. Paradysz (ed.), Report for Central Statistical Office, November 2008, Centre for Regional Statistics, University of Economics in Poznan (in Polish)

PARADYSZ, J. (2010), Konieczność estymacji pośredniej na użytek spisów powszechnych, [Necessity of indirect estimation in national census] in: Pomiar i informacja w gospodarce [*Measurement and Information in the Economy*] Gołata (ed.) published by Poznan University of Economics (in Polish).

PFEFFERMANN, D. (1999) Small Area Estimation – Big Developments, in: Small Area Estimation, International Association of Survey Statisticians Satellite Conference Proceedings, Riga 20-21 August 1999, Latvia.

PIETRZAK-RYNARZEWSKA, B., JOZEFOWSKI, T. (2010) *Ocena możliwości wykorzystania rejestru PESEL w spisie ludności,* [*Assessment of the possibilities of using population register in the census*] in: Pomiar i informacja w gospodarce [*Measurement and Information in the Economy*], Gołata (ed.) published by Poznan University of Economics (in Polish).

RAESSLER S. (2002) Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, Springer, New York, USA.

RAO, J.N.K. (1999) Some Recent Advances in Model-Based Small Area Estimation in: Survey Methodology, vol. 25, Statistics Canada.

RAO, J.N.K. (2003) Small Area estimation, Wiley-Interscience.

RAO, J.N.K. (2005) Interplay Between Sample Survey Theory and Practice: An Appraisal, Survey Methodology, Vol. 31, No. 2, 117-138.

RENSSEN, R. H. (1998) Use of Statistical Matching Techniques in Calibration Estimation in: Survey Methodology, Vol. 24, No. 2, 171 – 183, Statistics Canada.

ROSZKA, W. (2011) *An attempt to apply statistical data integration using data from sample surveys* in: Economics, Management and Tourism, South-West University "Neofit Rilsky" Faculty of Economics and Tourism Department, Duni Royal Resort, Bulgaria.

RUBIN, D. B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, in: Journal of Business and Economic Statistics, Vol. 4, No. 1, 87 – 94, stable URL: http://www.jstor.org/stable/1391390.

SäRNDAL, C.E., SWENSSON B., WRETMAN J. (1992) Model Assisted Survey Sampling, Springer Verlag, New York.

SÄRNDAL, C. E. (2007) The Calibration Approach in Survey Theory and Practice in: Survey Methodology. Vol. 33, No. 2, 99–119.

SÄRNDAL, C–E., LUNDSTRÖM S. (2005) Estimation in Surveys with Nonresponse, John Wiley & Sons, Ltd.

SCANU, M. (2010) Introduction to statistical matching in: ESSNet on Data Integration. Draft Report of WP1. State of the art on statistical methodologies for data integration, ESSNet.

SCHEUREN, F. (1989) A Comment on "The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration", Survey of Current Business, 40-41.

SKINNER, C. (1991) The Use of Estimation Techniques to Produce Small Area Estimates, A report prepared for OPCS, University of Southampton.

SZYMKOWIAK, M. (2011) Assessing the feasibility of using information from administrative databases for calibration in short-term and annual business statistics in: Use of Administrative Data for Business Statistics (2011) Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

Use of Administrative Data for Business Statistics (2011) G. Dehnel (ed.), Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

VAN DER PUTTEN, P., KOK, J. N., GUPTA, A, (2002) Data Fusion through Statistical Matching, Center for eBusiness, MIT, USA.

VEIJANEN, A., DJERF, K., SŐSTRA, K., LEHTONEN, R., NISSINEN, K. (2004) EBLUPGREG.sas, program for small area estimation borrowing srength over time and space using unit level model, Statistics Finland, University of Jyväskylä.

WALLGREN, A., WALGREN, B. (2007) Registered based Statistics Administrative Data for Statistical Purposes, John Wiley & Sons Ltd.

WINKLER, W.E. (1990) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: Section on Survey Research Methods, 354-359, American Statistical Association.

WINKLER, W.E. (1994) Advanced Methods For Record Linkage, Bureau of the Census, Washington DC 20233-9100.

WINKLER, W.E. (1995) Matching and Record Linkage, in: Business Survey Methods, B. Cox ed. 355-384, J. Wiley, New York.

WINKLER, W.E. (1999) The State of Record Linkage and Current Research Problems, RR99-04, U.S. Bureau of the Census, http://www.census.gov/srd/www/byyear.html.

WINKLER, W.E. (2001) Quality of Very Large Databases, RR2001/04, U.S. Bureau of the Census.

WU, CH. (2005) Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling in: Survey Methodology, Vol. 31, No. 2, 239 – 243.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 143—158

CUSTOMERS RESEARCH AND EQUIVALENCE MEASUREMENT IN FACTOR ANALYSIS

Piotr Tarka[1]

ABSTRACT

Factors Analysis is often tied to specific properties of population and its culture characteristics. If measurement is applied from population to another, then extracted factors may hard to be equally compared on the reflective basic level, unless all conditions of invariance measurement are met. Hence, implementation of customers research and any inter-cultural studies require a multi-cultural model describing statistical differences in both cultures with invariance as underlying assumption. In the article we implement a model for analysis of customers' personal values pertaining to hedonic consumption aspects in two culturally opposite populations. We conducted survey in two countries and the following cities: Poland (Poznan) and The Netherlands (Rotterdam and Tilburg) with randomly prepared samples with youth representatives on both sides. This model permitted us for testing invariance measurement under cross-group constraints and thus examining structural equivalence of latent variables - values.

1. Introduction

One of the main problems in most of socio-economic research is the measurement of equivalence pertaining to samples drawn from the different populations. Equivalence as a word relates to one of the categories of *quality assessment* in studies when scores obtained from e.g., two populations are set for comparison. Equivalence of a measurement is related to the assessment of the extent to which measurements are made in the tested groups using the same units and measures, distribution scores relating to the same characteristics of respondents according to various conditions and context of made observation (e.g., based on socio-economic factors, or frame of time). The measurement is therefore characterized by invariance level. In the absence of invariance in

[1] Poznan University of Economics, Department of Marketing Research, al. Niepodległości 10, Poland. E-mail: piotr.tarka@ue.poznan.pl.

measurement, any differences between individuals and populations cannot be reasonably interpreted as comparisons in any multi-group studies. This is particularly important in studies when we use units of measurement that are relative and conventional, associated with the respondents adopted own independent system reference (Sagan, 2005; Tarka, 2010).

2. Underlying assumptions of invariance measurement

Table 1 shows the main types of invariance research occurring at all stages of the research process. The issue of measurement invariance is crucial for studies that are aimed to investigate group differences. Cross-cultural methodologists have emphasized that group comparisons assume invariance of the elements of the measurement structure (i.e., factor loadings and measurement errors) and of response biases (Billiet, 2002; Little et al., 2006). And group comparisons within a single culture also require measurement invariance to insure that potential differences (e.g., in the means or regression coefficients) can be interpreted reliably (Vandenberg and Lance, 2000).

Sub-groups within populations are often heterogeneous with regard to the parameter values of a model. Nonetheless, most within-society research continues implicitly to assume homogeneity of the population (Muthén, 1989). This is especially happening in the field research pertaining to convenience samples of social, educational, or occupational sub-groups. These groups often differ from one another or from the overall population with regard to measurement or structural parameters. In the worst case, researchers measure different constructs in the groups. Hence, within-society studies should assess possible lack of measurement invariance, when possible, to uncover potential population heterogeneity.

Researchers usually assume invariance of the structure of their measures as they compare them across the groups. The validity of this assumption is critical for any conclusions about group related differences (Vandenberg and Lance, 2000). If it is not true, one cannot even claim that the construct is the same in the different groups (Little et al., 2006). Thus, legitimate comparison of means or structural relations across groups requires invariance of the measurement structures underlying the indicators (Ployhardt and Oswald, 2004; Thompson and Green, 2006).

Table 1. General categories for invariance measurements in different populations

Categories of intercultural equivalence test	**Types of equivalence in the category**	**Description**
Invariance of the research problem	Conceptual Functional	The identity of the constructs examined The similarity function of concepts and actions, predictive validity
Translation invariance	Lexical Idiomatic Grammar Pragmatic	The importance of vocabulary terms The importance of mobile and customary terms The adequacy of grammatical structures The importance of colloquial words in everyday life and action
Measurement invariance	Global Structural Metric Scalar Measurement errors	The similarity of the covariance matrix The adequacy of measurement models Comparability of measurement units Similarity measurement scale Homogeneity of the impact of specific factors
Sample invariance	Sampling units Representativeness	Comparability of sampling units Compliance operational units, the dimensions of socio-demographic stratification
Data collection invariance	Communication with the respondent Context Style and attitude response	The similarity of behaviour patterns, the definition of private and public spheres Commonality of questions of cultural context, the areas of social taboos and permissions Consistency and similarity of responses to the posed questions and themes non-response

Source: Sagan, 2005.

3. Process of invariance measurement

Table 1 shows the main types of invariance research occurring at all stages of the research process. The issue of measurement evaluation based on invariance begins with series of conducted tests where one checks the hypotheses related to dispersions among the groups. These tests should be carried out in a sequence, because the bad model fit, makes another baseless testing measuring the level of cross-cultural equivalence (Meredith, 1993; Sagan 2005). As a result we obtain

configural invariance of the whole factorial structure and metric invariance of the factor loadings which are critical for the interpretation of constructs and are requisites for all other measurement. Configural invariance implies the same number of factors in each group and the same pattern of fixed and free parameters. It is a prerequisite for the other tests. It is the very basic form of invariance and it assesses whether we find the same patterns of loading between indicators and factors in both groups. The parameter restrictions only refer to the patterns of "loading" and "non-loading". Configural invariance is assumed if the same items measure the same factors in both groups. If configural invariance is not supported empirically, there are fundamental distinctions in the measurement structure, which means that the manifest variables measure different latent variable.

The **metric invariance** is more stringent in comparison to the configural invariance, as additional restrictions are adopted. Metric invariance means that, in addition to the conditions of configural invariance for all groups, the factor loadings are equivalent. If the model of metric invariance is maintainable, the manifest variables measure the latent variables equally well. If the model fit of the metric invariance model does not decrease significantly, metric invariance of all items can be assumed. Given a metric invariance, the contents of the factors are assumed to be equivalent. Likewise, the relations of the variables with other variables may be compared across the groups (Bollen, 1989).

The test of metric invariance is conducted by comparing the fit of the metric and configural invariance models to the data with chi-square statistics. Further 'modern' indications for invariance are differences in the indices such as *Comparative Fit Index* (CFI), *Root Mean Square Error of Approximation (RMSEA),* and *Standardized Root Mean Residual* (SRMR). Metric invariance implies equal factor loadings across groups. For instance, the parameter λ_{ij} - representing factor loading must be the same in groups e.g., A and B. And this is tested by imposing equality constraints on the Λ - matrices that contain the factor loadings (i.e., $\Lambda^A = \Lambda^B = ...\Lambda^G$, where superscripts refer to groups A to G). Equal factor loadings indicate that the groups calibrate their measures in the same way. Hence, the values on the manifest scale have the same meaning across groups.

Metric invariance concerns a construct comparability that metric invariance is a stricter condition of construct comparability. According to the common factor perspective, the factor loadings indicate the strength of the causal effect of the latent variable ξ_j on its indicators and can be interpreted as validity coefficients (Bollen, 1989). Significantly different factor loadings imply a difference in the validity coefficients. This raises concerns about whether the constructs are the same across groups. Hence, configural invariance, by providing evidence that the construct is related to the same set of indicators, is a prerequisite for inferring that the construct has the *similar* meaning. However, metric invariance is necessary to infer that the construct has the *same* meaning, because it provides evidence about the equality of validity coefficients.

Additionally, **scalar invariance** refers to invariance of the item intercepts in the regression equations that link the indicators x_i^g to their latent variable ξ_j^g. Item intercepts can be interpreted as the systematic biases in responses of a group to an item. As a result, the manifest mean can be systematically higher or lower (upward or downward biased) than one would expect due to the groups' latent mean and the factor loading. Scalar invariance is present if the degree of up- or downward bias of the manifest variable is equal across groups. It is absent if one of the groups differs significantly in one or more of the item intercepts. To test for scalar invariance, one constrains the tau-vectors to be equal across groups $\tau^A = \tau^B = ...\tau^G$.

Then, we follow **invariance of factor variance / covariance,** which appears when groups have the same variances in their respective latent variables. This is tested by constraining the diagonal of the phi-matrices $\phi_{jj}^A = \phi_{jj}^B = ...\phi_{jj}^G$, to be equal. And invariance of the factor covariances refers to equality of the associations among the latent variables across groups. It is tested by constraining the sub-diagonal elements of the phi-matrices $\phi_{jk}^A = \phi_{jk}^B = ...\phi_{jk}^G$, to be equal. Covariances among constructs have implications for the constructs' meaning or validity (Cronbach and Meehl, 1955). Hence, unequal covariances raise concerns about the equality of construct meanings (Cole and Maxwell, 1985).

As far as the analyses of **invariance of the latent means** are concerned, they are conducted in order to test for differences between groups (or points of time) in their latent means. In contrast, traditional approaches to the analysis of mean differences use composite *manifest* scores and employ *t* tests, ANOVA, or MANOVA. The validity of testing group differences in manifest scores depends on whether the assumptions that underlie such comparisons are correct, specifically, that both the factor loadings and the item intercepts are equal (i.e., metric and scalar invariance) (Tarka, 2011). The relationship between a latent and observed mean or an expected observed value can be written as follows:

$$E\left(x_i^g\right) = \tau_i^g + \lambda_i^g \kappa_i^g. \tag{1}$$

where:

$E\left(x_i^g\right)$ - expected value of the *i*-th manifest indicator in group *g*,

κ_i^g - is the mean of factor f in group *g* to be considered in the tests related with latent means comparison of the particular groups.

It shows that a manifest mean depends not only on its latent mean but also on the factor loading and the item intercept. Thus, a manifest mean difference can be caused either by a latent mean difference or a difference in the loadings, intercepts, or both. Therefore, a test of latent mean difference requires the equality of both the factor loadings and item intercepts. The equality of the latent means is tested by constraining the kappa matrices $\kappa^A = \kappa^B = ...\kappa^G$, to be equal across groups.

Finally, a **measurement of errors invariance** concerns the hypothesis that the measurement error in the manifest indicators $\Theta^{A} = \Theta^{B} = ...\Theta^{G}$, is the same across groups.

4. Factor analysis model for two populations

In factor analysis model we consider a set of *m* populations $\Pi_1, \Pi_2, ..., \Pi_m$. They may be different nations, or culturally different groups, groups of individuals selected on the basis of some known or unknown selection variable, groups receiving different treatments, etc. In fact, they may be any set of exclusive groups of individuals that are clearly defined. And it is assumed that a battery of tests has been administered to a sample of individuals from each population. The battery of tests need not be the same for each group, nor need the number of tests be the same. However, since we shall be concerned with characteristics of the tests that are invariant over populations, it is necessary that some of the tests in each battery are the same or at least content-wise equivalent. A general factor analysis model in each population will be as follows (Jöreskog, 1971):

$$x_g = \mu_g + \Lambda_g f_g + e_g, \tag{2}$$

where:

x_g is random vector with mean vector μ_g (and variance-covariance matrix of population Σ_g) in group g. As a result x_g is explained by k_g common factors f_g and unique factors e_g. Furthermore, we assume that $\varepsilon\left(f_g\right) = 0$ and $\varepsilon\left(e_g\right) = 0$ and so the same with Λ_g a factor pattern. And this implies factor analytic solution as follows (Jöreskog, 1971):

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Psi_g, \tag{3}$$

where:

Φ_g - variance – covariance matrix of f_g

Ψ_g is the diagonal variance – covariance matrix of e_g.

In contrast to Jöreskog general model, Lawley and Maxwell proposed separate models for strictly two populations with variance-covariance matrices denoted as Φ_1 and Φ_2. The coefficients of factor loadings, - if invariant under the changes of populations – will cause loading matrix Λ the same for both populations. They also assumed that Ψ diagonal matrix of e, will be the same. The model can be generalized to some extent by allowing populations to have different unique factors (residual variances) on variance-covariance matrices Ψ_1 and Ψ_2, but this option complicates subsequent estimation procedures. The population variance-covariance matrices for the given x_i are thus given as follows (Lawley and Maxwell, 1963):

$$\Sigma_1 = \Lambda\Phi_1\Lambda' + \Psi, \tag{4}$$

$$\Sigma_2 = \Lambda\Phi_2\Lambda' + \Psi. \tag{5}$$

Such being the case, certain loadings are a priori zero, and that number of and positions of these are such as to determine factors uniquely. The factors are arbitrary and for computational convenience, they can be chosen in such a way that the matrix will be (Lawley and Maxwell, 1963):

$$\Phi = \frac{\left(n_1\Phi_1 + n_2\Phi_2\right)}{\left(n_1 + n_2\right)}. \tag{6}$$

and has unit diagonal elements. As a result there are *k* factors, where certain specified elements of the loading matrix Λ are zero and that the population variance-covariance matrices satisfy assumptions of the Eq. (4) and (5).

5. From the model of two populations towards the model of two samples

For general model, we have N_g respondents in the sample from g-th population, $\overline{x}_g$ as the usual sample mean vector and S_g - sample variance-covariance matrix with $n_g = N_g - 1$ degrees of freedom. Thus, we obtain independent measurements for different groups.

We may thus assume that S_1 and S_2 are the separate variance-covariance matrices for e.g., two groups with respectively n_1 and n_2 degrees of freedom, obtained by taking a random sample from each population. Then, general log-likelihood function for S_g sample will be:

$$\log L_g = -\frac{1}{2} n_g \left\{\log_e\left|\Sigma_g\right| + \mathrm{tr}\left(S_g\Sigma_g^{-1}\right)\right\}. \tag{7}$$

So, if the samples are independent, the log-likelihood for all the samples is:

$$\log L_g = \sum_{g=1}^{m} \log L_g. \tag{8}$$

And the log-likelihood function for two separate groups will be (without function with observations) (Lawley and Maxwell, 1963):

$$-\frac{1}{2} n_1 \left\{\log_e\left|\Sigma_1\right| + \mathrm{tr}\left(S_1\Sigma_1^{-1}\right)\right\} - \frac{1}{2} n_2 \left\{\log_e\left|\Sigma_2\right| + \mathrm{tr}\left(S_2\Sigma_2^{-1}\right)\right\}. \tag{9}$$

To estimate unknown parameters we should have maximized it with respect that non-zero elements of Λ, the elements of Ψ, and the elements of Φ_1 and Φ_2 are subject to restriction that Φ has diagonal elements. The resulting equations of estimations may be simplified and solved iteratively. The hypothesis will be tested by means of the criterion (Lawley and Maxwell, 1963):

$$n_1 \log_e \left(\frac{|\hat{\Sigma}_1|}{|S_1|} \right) + n_2 \log_e \left(\frac{|\hat{\Sigma}_2|}{|S_2|} \right). \tag{10}$$

which for large samples is distributed approximately χ^2 with $(p^2 - k^2 - m)$ degrees of freedom, where *m* is the number of non-zero loadings.

If we want to administer the same test/measurement within different populations, we must follow conditions of invariance as previously discussed. In particular we need to consider invariance of:

- Λ in factorial pattern over populations,
- $\psi_1^2 = \psi_2^2$ of with variances of regression.

Then, we identify parameters where Λ in $\Sigma_g = \Lambda \Phi_g \Lambda' + \psi_g^2$, will be replaced by $\Phi_g^* = T \Phi_g T'$, $g = 1, 2, ..., m$, and where T is an arbitrary non-singular matrix of order $k \times k$. Then, each Σ_g remains the same so that the function F is unaltered.

$$F = \frac{1}{2} \sum_{g=1}^{m} n_g \left[\log|\Sigma_g| + \mathrm{tr}\left(S_g \Sigma_g^{-1}\right) - \log|S_2| - p_g \right]. \tag{11}$$

Since the matrix T has k independent elements, this means that at least k^2 independent conditions must be imposed on $\Lambda, \Phi_2, \Phi_2, ..., \Phi_m$ to make them uniquely defined. And the most convenient way of doing this is to let all the Φ_g, be free and to fix one non-zero element and at least $k-1$ zeros in each column of Λ. In an exploratory study one can fix exactly $k-1$ zeros in almost arbitrary positions. Jöreskog (1971) claims that one may choose zero loadings where one thinks there should be "small" loadings in the factor pattern. The resulting solution may be rotated further, if desired, to facilitate better interpretation. In a confirmatory study, on the other hand, the positions of the fixed zeros, which often exceed $k-1$ in each column, are given *a priori* by a hypothesis and the resulting solution cannot be rotated without destroying the fixed zeros.

In order to make observable variables comparable, according to different units of measurement in different samples, one can rescale these variables before beginning the factor analysis. As a result we assume (Jöreskog, 1971):

$$S = \left(\frac{1}{n} \right) \sum_{g=1}^{m} n_g S_g, \tag{12}$$

where: $n = \sum_{g=1}^{m} n_g$, and

$$D = \left(\mathrm{diag}\ \hat{\Phi} \right)^{-\frac{1}{2}}. \tag{13}$$

Then, the variance-covariance for the rescaled variables is:

$$S_g^* = DS_g D. \tag{14}$$

The weighted average of S_g^* is a correlation matrix. The advantage of rescaling is that, when combined with an option of rescaling the factors, factor loadings are of the same order of magnitude as usual when correlation matrices are analyzed and when factors are standardized to unit variances. This makes it easier to choose starting values for the minimization and interpretation of the results. It should be indicated further that it is not permissible to standardize the variables in each group and to analyze the correlation matrices instead of the variance-covariance matrices. This violates the likelihood function (7-8) which is based on the distribution of the observed variances and covariances. Invariance of factor patterns is expected to hold only when the standardization of both tests and factors are relaxed.

6. Example: values system analysis in Polish and Dutch youth

We drew basic ideas and developed our research on Rokeach (1973) and Schwartz (1992) definition of values, describing them as *"desirable, transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity"*. As a result values are driven by different motivations (Schwartz and Sagiv, 1995) (Table 2).

The theory postulates 10 different types of values and two main value dimensions. The 10 types of values are arranged in a circumplex structure around the following dimensions: *self-transcendence* versus *self-enhancement* and *openness to change* versus *conservation.* Figure 1 displays the circular structure of the types of values as well as the two dimensions behind them (Schwartz and Boehnke, 2004; Schwartz, 2005).

The dimension of *self-transcendence/self-enhancement* describes the possible conflict between the acceptance of others as equal entities and the concern for their well-being (types of values: universalism and benevolence) versus the tendency to try to achieve personal success as well as predominance over others (types of values: power and achievement). The second dimension reflects the possible conflict between independent thought and action and preference for an exciting life (types of values: self-direction and stimulation) versus the tendency to seek stability, security, and attachment to customs, traditions, and conventions (types of values: security, conformity, and tradition). Virtually different types of values correlate differently. And the value type related to *hedonism*, forms a link between *openness to change* and *self-enhancement* (Tarka, 2010).

Table 2. The 10 types of values with motivational goals and the higher-order dimensions

Value	Motivation	Dimension
Self-direction	Independent thought and action-choosing, creating, exploring	Openness to change
Stimulation	Excitement, novelty, and challenge in life	Openness to change
Hedonism	**Pleasure and sensuous gratification for oneself**	**Between self-enhancement and openness to change**
Achievement	Personal success through demonstrating competence according to social standards	Self-enhancement
Power	Social status and prestige, control and dominance over people and resources	Self-enhancement
Security	Safety, harmony, and stability of society, of relationships, and of self	Conservation
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms	Conservation
Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide the self	Conservation
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact	Self-transcendence
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature	Self-transcendence

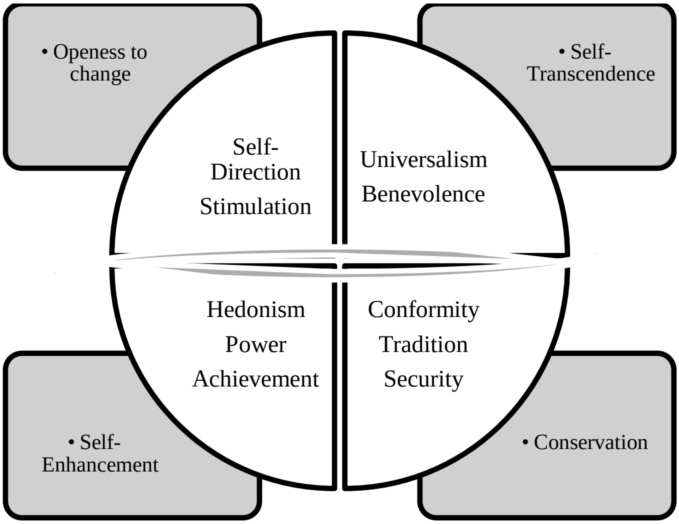
Source: Sagiv and Schwartz, 1995.

Method of data collection

Initially the 19-item question battery was applied in the study to measure value priorities among Polish and Dutch youth representatives in the academic environment. The interviewee was confronted with a five-point Likert scale (where: 1 = totally disagree; 5 = totally agree). This type of scale is a parallel in which each item represents an alternative and equivalent tool for measuring a latent trait. Evaluation of reliability and measurement invariance in scales of this type is made in context of classical theory of the test.

Data collection was based on paper and pencil interviews. In the course of empirical research, printed questionnaires had been handed out to a number of individuals (respondents) at Universities in Poland (in Poznan Universities e.g., Poznan University of Economics, Adam Mickiewicz University of Poznan and Poznan University of Technology) and Universities in the Netherlands in Rotterdam and Tilburg for final evaluation of the prepared sets of items. Sampling frame was derived and prepared according to internal universities database including complete list of participants attending three introductory classes of undergraduate level. Respondents were selected on the rules of simple random sampling. Only a small percentage (less than 5%) of those contacted refused to participate in a study. Thus a collected sample was n = 285. The data was collected between May and June 2009.

Figure 1. Dimensions of value systems



Source: own construction

<u>Data analysis and empirical results</u>

At first the factorial structure was tested and then measurement invariance of the instruments that were operationalized according to the value theory. Sometimes one may also apply the other model for comparison of measurements at different points in time. Such being the case, growth curve models of latent dimension are used (e.g. latent growth curve models). But the latter option was beyond objective of this article.

The assumptions for the assessment of invariance were as follows:

- scale consisted of multiple items,

- items were of reflective form (they reflected latent dimension otherwise latent variable),
- the measurement was performed in two groups at one time.

All analyses were conducted using the computer program LISREL, where a maximum likelihood was applied as the estimation method.

Single group analysis

At first we measured directly the higher-order dimensions of the values by their corresponding items. The two higher dimensions self-transcendence/self-enhancement and openness to change/conservation constituted four factors. The remaining 19 items were attributed to these four factors.

The models required several modifications. At first, items that did not achieve adequate factor loadings were eliminated. The criterion we set for an item to load on a factor was 0.49 and higher. Some loadings were too low for the conservation, self-transcendence, and self-enhancement factors. As the invariance test should be performed on the same measurement model, we eliminated the same items in both samples. Consequently, the final model that we tested for invariance included 15 items.

Multi-group analysis

Next we turned to multiple-group comparison. This model enabled us to test to what extent the value measurements were invariant across the samples. To test it we used the model that included 4 constructs, 15 items, 1 cross-loading and 2 errors. We compared two groups (e.g. Polish and Dutch nationalities). The empirical covariance matrix of the items for each group served as the input. Variance-covariance matrix allowed for comparing outcomes in terms of intergroup value of the original units of measurement. If the variance-covariance matrices are not significantly different in both groups, one can perform further analysis of individual aspects of measurement invariance.

The evaluation of the **structural invariance** of latent variables (also called configural invariance) was conducted in both groups. This reflected a test of the hypothesis of equality variance-covariance matrix based on the degree of goodness of fit of structural independent models made on the basis of data from individual cultures. Good models and their fit to data proved the existence of a configuration invariance and enabled us further comparison between the constructs. The degree of fit was tested using the statistics such as χ^2 index, CFI Bentler, PCLOSE probability of close fit, coefficient RMSEA. Values close to .95 for CFI and below.06 for RMSEA suggest a good fit (Hu and Bentler, 1993).

Next, we assessed **metric invariance,** e.g. the factor loadings of all items that were constrained to be identical across two groups. This assessment was based on a comparison of relative fitting between two structural models. In the first model, corresponding factor loadings were set as equal in all groups (factor loading λ_1 in the first group was equal to the value of the factor loading λ_1 in the second group of respondents). In the second model, factor loadings in both groups were the free parameters. If the fit of the model with defined (fixed) factor loadings was not significantly worse, as compared to model with free – released loadings, then items would measure the latent variables (factors) in a comparable way in both analyzed groups. However, if the degree of data fit in a model, pertaining to fixed factor loadings was significantly worse, then comparison of factor loadings between groups could be only made on partial invariance measurement between the groups.

Table 3. Fit measures for the model assessing configural, metric, and scalar invariance

	Configural invariance	**Metric invariance**	**Scalar invariance**
Chi-square	179,56	192,10	205,60
CFI – comparative fit index	0,935	0,940	0,948
RMSEA – root mean square error of approximation	0,046	0,049	0,040
PCLOSE – probability of close fit	0,495	0,515	0,540
SRMR – standardized root mean square of residuals	0,081	0,082	0,086

Source: own calculation in LISREL.

Next, we turned to the test of ***scalar invariance.*** It allowed us to compare the mean values for the latent variables, especially to detect: 1). inter-group differences in the responses (according to particular statements which determined latent dimensions) and also 2). effects of respondents attitudes and differences in their style of giving responses to these statements.

The global fit measures of the configural invariance model, which are displayed in Table 3, suggest that the model should not be rejected. The results indicate that the metric invariance model is supported by the data. A chi-square difference test between the configural and the metric invariance model revealed that there was no significant difference in the model fit. Also, the fit indices of CFI, RMSEA, and SRMR are further indications for invariance. In case of scalar invariance, we may observe that the constrained intercepts of the items are equal across the samples. As the results we cannot reject the scalar invariance model.

As configural, metric, and scalar invariance has been confirmed, the comparison of latent mean values between the two samples was easy to conduct.

And because one intended to compare the ***latent means*** **in both groups,** therefore we added a vector of manifest means as input. With regard to the parameter matrices, the τ_x -vector and the κ -vector were added. The results are presented in Table 4.

Table 4. Latent mean differences of four constructs (reference group: Polish sample)

	Means for Polish group	**Means for Dutch group**	**Effect sizes (r)**
Openness to change	3,95	3,24	,28∗
Self-enhancement	4,87	4,10	,31∗
Self-transcendence	2,67	2,65	,00
Conservation	3,16	2,53	,12*

Note: Effect sizes of r in the latent means at $*$ *p* $<$ *0,5; Source: own calculation in LISREL.*

Results show significant mean differences for the constructs *openness to change* (estimate = 0,28), *self-enhancement* (estimate = 0,31), and *conservation* (estimate = 0,12). For the construct *self-transcendence* we found no significant mean difference between groups (estimate = 0,0). As a result differences have been found for the latent means of both samples for the constructs ***openness to change, self-enhancement***, the hypothesis that the latent means for value questions were identical in both groups must be rejected. Individuals in the Dutch sample displayed lower levels of ***openness to change*** and ***self enhancement*** (which were also in their own part of hedonic senseless and excessive consumption of the market goods) as compared to Polish sample.

From the above results and application of model it is quite interesting to infer that Polish youth as compared to Dutch youth (being derived from agglomerations such as: Poznan, Rotterdam and Tilburg) exposes more interest towards types of values such as Hedonism in general. Apparently young Poles look now for more pleasure and enjoyable life (also pertaining to products and services consumption) than their foreign colleagues from already developed countries. Events from the past and hard rules of socialism and limitation in access for years to free market goods left their strong impact on young people’s life and behaviour. Being kept too long away from open market sources, citizens of eastern block of Europe, e.g. Poland, seem to recoup their delays and catch up with latest trends arising on the market. In contrast, Dutch youth, being too long exposed to wide markets, virtually grew accustomed to its products and services. As a result this situation affected their life style, lowering also their interests in Hedonism that is senseless and excessive consumption of the market goods. And these facts simply reveal a new perspective for companies business activities that is either to point on new directions associated with entry on new ascending markets.

7. Conclusions

Discussed in article a model of factor analysis model was strongly based on the examination of measurement invariance and specifically, factor invariance. Researcher when using such a type of model avails of the opportunity to detect invariance for tested items and simultaneously generate reliable and valid constructs. If these assumptions are not satisfactory then making further inferences becomes pointless. In a consequence the model requires certain parameters (e.g., factor loadings) to be constrained in the process of identification, which means they need to be invariant across groups, and act as referent variables. If this invariance assumption for some reason would be violated, then location of the parameters that actually differ across groups would become difficult. In case of the conducted analysis and implemented model, it simply turned out to be a satisfactory solution regarding the researched problem and final calculated scores.

REFERENCES

BILLIET, J. (2002), *Cross-cultural equivalence with structural equation modeling,* [in:] Mohler, P.P. (Ed.) *Cross-cultural survey methods,* New Jersey: John Wiley & Sons Inc., pp. 247 - 264.

BOLLEN, K.A., (1989), *Structural equations with latent variables,* New York: Wiley.

COLE, D.A., MAXWELL, S.E. (1985), *Multitrait-multimethod comparisons across populations: a confirmatory factor analytic approach.* Multivariate Behav. Res. 20, pp. 389 – 417.

CRONBACH, L.J., MEEHL, P.E. (1955), *Construct validity in psychological tests,* Psychol. Bull. 52, pp. 281 – 302.

HU, L.-T., BENTLER, P.M., (1999), *Cut-off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives.* Structural Equation Model. 6, pp. 1–55.

JÖRESKOG, K.G. (1971), *Simultaneous factor analysis in several populations,* Psychometrika, 36, pp. 409 – 426.

LAWLEY, D.N., MAXWELL, A.E. (1963), *Factor analysis as a statistical method.* London: Butterworths.

LITTLE, T.D., SLEGERS, D.W., CARD, N.A. (2006), *A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models,* Structural Equation Model, 13(1), pp. 59 – 72.

MEREDITH, W. (1993), *Measurement invariance, factor analysis and factorial invariance,* Psychometrika, Vol. 58, pp. 525 - 543.

MUTHÉN, B., (1989) *Latent variable modeling in heterogeneous populations,* Psychometrika, 54(4), pp. 557 – 585.

PLOYHARDT, R.E., OSWALD, F.L. (2004), *Applications of mean and covariance structure analysis: integrating correlational and experimental approaches,* Organ. Res. Methods, 7(1), pp. 27 – 65.

ROKEACH, M., (1973), *The Nature of Human Values,* New York: Wiley.

SAGAN, A. (2005), *Analysis of equivalence off measurement scales in inter-cultural studies,* Scientific Papers, Vol. 659, pp. 59 – 73.

SCHWARTZ, S.H. (1992), *Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries,* [in:] Zanna, M. (Ed.), *Advances in experimental social psychology,* vol. 25, Orlando: Academic Press, pp. 1 – 65.

SCHWARTZ, S., SAGIV, L. (1995), *Identifying culture-specifics in the content and structure of values,* J. Cross-Cult. Psychol. 26(1), pp. 92 – 116.

SCHWARTZ, S.H., BOEHNKE, K. (2004), *Evaluating the structure of human values with confirmatory factor analysis,* J. Res. Pers. 38(3), pp. 230 – 255.

SCHWARTZ, S.H. (2005), *Basic human values: their content and structure across countries,* [in:] Tamayo, A., Porto, J.B. (Eds.), *Values and Behavior in Organizations,*. Vozes, Petrópolis, pp. 21 – 55.

TARKA, P. (2010), *Measurement scales for customers hedonic values – comparison of reliability techniques* – Scientific papers “Econometrics” UE Wrocław, 29, pp. 23 – 38.

TARKA, P. (2010), *Latent variable models - issues on measurement and finding exact constructs in customers values* – Polish Statistical Review (Przegląd Statystyczny), 4, pp. 142 – 167.

TARKA, P. (2011), *Statistical analysis of youth’s value systems in Poland and Netherlands – an approach to LOV and RVS scale* [in:] Józef Pociecha (Ed.) *„Methods of data analysis”,* Academic papers UE Kraków, 2011, pp. 86 – 94.

THOMPSON, M.S., GREEN, S.B. (2006), *Evaluating between-group differences in latent means.* [in:] Hancock, G.R., Mueller, R.O. (Eds.) *Structural Equation Modeling: A Second Course,* Greenwich: Information Age, pp. 119 - 169.

VANDENBERG, R.J., LANCE, C.E., (2000), *A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research,* “Organ. Res. Methods” 3(1), pp. 4 – 69.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 159—160

EDITOR'S NOTE ON THE STATISTICAL CONGRESS SECTION

The main objective of every statistical congress can perhaps be said as excelling at professional language for equivalence and comparisons. In doing so, statisticians act in the vein of Adolph Quetele's tradition, who (as the president of the Central Committee of Statistics in Belgium) organized the first such a congress, held in Brussels in 1853 (though, as an international event), and who stressed the need to stabilize the language of statistics "specifically to promote the unification of official statistics that the governments published, providing comparable results".

Similar idea has guided the efforts of the Polish Statistical Association that at the outset of resumption of its activity in the late 1930s established two scientific committees: one for statistical vocabulary, other for preparing guidelines for exploring statistical resources.

And this type of idea – contributing to stabilization of the broadly conceived language of statistics, not only cross-nationally but also across sectors and across disciplines – continues to also guide the mission of our Journal, the international scope of which is emphasized in its sub-title "an international journal".

The upcoming congress to celebrate 100th Anniversary of the Polish Statistical Association deserves highlighting for both reasons: (i) as the scientific meeting of representatives of the community of statisticians, and of users of statistics; and (ii) to underlie the importance of one of the most active country's scientific association, which continues to promote and encourage awareness of the statistical profession and shaping several areas of application of statistics, including research, education and dissemination of statistical information.

Therefore, I have asked leading representatives of this community, both researchers and practitioners, who act as members of the Journal's Editorial Board, to comment on the occasion of this anniversary. Especially, to address some issues they consider of particular interest to them, and to the discipline as a field of dynamic development.

This section contains an array of such 'occasional statements', from historical remarks on the Polish Statistical Association (C. Domański and W. Łagodziński), through some challenges of public statistics (M. Szreder), to origin and development of the Journal (J. Kordos and W. Okrasa).

This part is, however, preceded by two biographical notes, devoted to two key figures in modern statistics in general – Jan Czekanowski and Jerzy Neyman. They both, for somewhat different reasons (e.g., J. Czekanowski was also one of the leaders of the Polish Statistical Association before World War II), are to some extent – although not being formally declared as such – considered spiritual (scientific) patrons of the congress.

This section is completed by the congress' organizational materials: the Announcement of the Congress and the Congress Agenda.

Włodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 161—168

JAN CZEKANOWSKI (1882–1965)

Jan Czekanowski was born into a landowning family, on 6 October 1882 in Głuchów, near Grójec in the Masovia region. His father Wincenty (1836–1926) was the owner of the estates of Głuchów and nearby Kośmin. His mother, Amelia von Guthke, was German. Jan had four elder brothers and sisters: Natalia, Aleksander, Stanisław and Maria. He was initially educated at home, but in autumn 1894 he became a third-year pupil at the well-known “real school” run by Wojciech Górski in Warsaw. In autumn 1898 he moved to the real school in Libava (Liepaja) in Latvia, where he passed his *matura* school-leaving exam in June 1901. On 1 September 1901 he joined the army as a volunteer. Through an oversight, in contravention of an instruction of 1888, he was accepted – in spite of his Catholicism – into the defensive artillery of the then military port of Tsar Alexander III in Libava. Unable as a private to be moved to another unit without a Supreme Order, and unable as a Catholic to remain in the defensive artillery of the Vilnius district to which Libava belonged, on 6 December 1901 he was discharged from the army as unfit due to overstraining of the heart. He went abroad, leaving the Russian Empire without the appropriate documents, and with his heart in his mouth. He crossed the border in a saloon car occupied by a high-ranking imperial officer and his wife, which enabled him to avoid any border checks. After journeying to Italy, in spring 1902 he was accepted into the mathematics and natural science section of the Philosophy Faculty at the Cantonal University of Zurich. There he studied anthropology under the superb anthropologist Rudolf Martin, anatomy under Georg Ruge, and mathematics under Heinrich Burghardt. It was to these subjects that he would devote his long, hard-working life. He understood anthropology in a broad sense, from a humanist standpoint – as encompassing knowledge about man and his functions. Anatomy was a part of that knowledge, along with ethnography, anthropogenesis and typology, genetics, linguistics and statistics. Czekanowski saw the human being as a creature characterized by a large number of connected and correlated features. He understood that to study only a few of these features must necessarily lead to a limited, fragmentary, one-sided picture which would obscure or even falsify the real human being as an object of study.

At the beginning of the 20th century, when Czekanowski was studying in Zurich, a multifaceted analysis of the human being was not yet feasible. The English statistical school (Pearson, Yule, Fisher, Student) was just beginning its activity. Czekanowski quickly appreciated the role that statistics could play for anthropology and the empirical sciences, and became a student and pioneer of the new discipline.

Jan Czekanowski's first academic work was a short monograph on statistics, written in the second semester of his university studies. It demonstrates the use of Pearson's correlation coefficient to evaluate various methods of measuring skull height. It should be noted that his was 1902, and Karl Pearson had introduced the correlation coefficient in 1901. With this initial work, Czekanowski travelled in 1903 to a congress of German anthropologists in Worms, accompanying his professor Rudolf Martin. He applied too late to be included on the list of speakers, but he made such an impression with his thorough knowledge of the latest methods of English mathematical statistics, which were not yet known among German anthropologists, that Felix von Luschan, director of the African and Oceanic departments of the Royal Anthropological Museum in Berlin, offered him a research assistant's position with prospects of being sent to Africa or Oceania. Czekanowski agreed to take up the position, but only after completing his studies in Zurich. The monograph, which had proved so significant for Czekanowski's future destiny, eventually appeared in print in *Archiv für Antropologie* in 1904.

In 1903 Czekanowski wrote a paper on the application of the correlation coefficient to the study of muscular anomalies, for which he collected material while working as deputy assistant in the anatomy laboratory. This attempt to apply modern statistical methods in anatomy was published in 1906 in a memorial volume to the American anthropologist Franz Boas.

Czekanowski's work to popularize biometrics is well known. While still a student in Zurich he wrote a paper on the subject, which appeared in 1904 as an introduction to Rudolf Martin's anthropology textbook *Lehrbuch der Anthropologie*, which is still widely known among anthropologists today, and in 1907 was published in Czekanowski's doctoral dissertation. In it he gives a short description of the statistical methods which had been introduced to anthropology by English biometricians. He completed his studies in July 1906, obtaining the degree of doctor of philosophy (his certificate is dated 1907).

In the following winter semester Czekanowski furthered his education by studying mathematics at Berlin University. As a fresh graduate from Zurich, starting from 1 November 1906 he took up the position of assistant at the Royal Anthropological Museum in Berlin. That post provided the possibility of travelling to Africa under a scholarship from the Prussian government. Thus his youthful dreams of an exotic trip to Africa were to be realized. The young Czekanowski was invited by Prince Adolf Frederick of Mecklenburg to join a scientific expedition to the Nile-Congo region in Central Africa. He spent more than two years (from 1 May 1907 to 7 July 1909) in Sudan, the Congo, Uganda and German East Africa, returning to Berlin via Egypt, Syria and the Balkans. His duties included producing an ethnographic map. The enterprise was a huge one. A total of 2230 porters were employed, and seven stations were prepared along the route equipped with food, drink, medicines, tools, clothing, tents, camp beds, firearms, and even folding bathtubs made of a crumpling impermeable material – in short, everything that the explorers would need. Where possible, expedition members stayed with missionaries, at colonial borderland fortresses or at the courts of African rulers. Czekanowski crossed north-western Tanzania, Rwanda and two extensive borderlands: between Uganda and Zaire, and between Zaire and Sudan. The undertaking took place in quite exceptional conditions; the territory explored, with an area almost twice that of Switzerland, covered a region that had been inaccessible to European colonialists and to Arab, Indian and even African merchants. Both the times and the places visited remained politically unstable and of uncertain future. Spending more than two years on the expedition, Czekanowski collected vast amounts of sometimes unique materials from parts of Africa which at that time were entirely unknown. The materials were relevant to both anthropological and ethnological or ethnographic questions, and even to some extent to sociological topics. He published them over many years, some of them even after World War II, in 1951. The first five volumes were published in Leipzig from 1911 to 1927 in the form of a vast monograph titled *Forschungen im Nil-Kongo Zwischengebiet*.

For the results achieved on the African expedition he was decorated with the orders of the Belgian Crown and the Mecklenburg Griffon, and with a Mecklenburg Memorial Medal.

Czekanowski achieved his most important results in studies of racial classification and population structure. The revolution brought about by Jan Czekanowski in human classification was chiefly based on the introduction of a new taxonomic method for racial analysis. This was introduced in 1909 as Czekanowski's diagraphic method. It was published in Czekanowski's fundamental methodological paper *Zur differentialdiagnose der Neandertalgruppe*, which remained a standard work for his pupils for many years afterwards.

On 2 March 1910 Jan Czekanowski married Elizavyetą (Elizabeth) Sergiyevska, daughter of an Orthodox parish priest in Tula. The couple had met in Zurich, where Elizabeth was studying medicine. They would have two daughters: Zofia Teresa (born 25 September 1927) and Anna Katarzyna (born 25 June 1929).

On 1 October 1910 Jan Czekanowski was appointed curator of the Ethnographic Museum of the Imperial Academy of Sciences in St. Petersburg. He moved there at the start of 1911, and remained in the post until the end of September 1913. While he was at St. Petersburg the well-known zoologist Józef Nussbaum-Hilarowicz made a proposal to Czekanowski that he should complete his "habilitation" degree in anthropology and take a university chair at Lvov (Lwów, Lemberg). After a fairly long period of hesitation and delay, he decided to move to Lvov. By a letter of the Imperial Ministry of Denominations and

Enlightenment in Vienna, dated 11 August and with effect from 1 October 1913, he was appointed assistant professor of anthropology and ethnology in the Philosophy Faculty of Lvov University. He was appointed on his merits, without having gained his habilitation. He began lecturing at the start of the 1913/1914 academic year, and would spend the longest period of his life in Lvov, until 1944. Apart from his lectures he also organized the anthropology and ethnology department and engaged in research on national anthropology.

However that work was suspended as a result of the outbreak of the First World War. As a Russian subject being in the state service of Austria, he was compelled to make a hasty escape from the Russian army. In late August 1914 he travelled to Krynica, and later to Busko. On 30 September 1914 he settled in Luhačovice in Moravia, where he continued writing up the materials collected during his African expedition. With the issuing of a decree recognizing citizens of the Polish Kingdom, he obtained a passport, and on 10 October 1916 he returned to Lvov. At the start of the 1916/1917 academic years he renewed his lectures at the university as a full professor. His work at the university was again interrupted on 1 November 1918, after Lvov had been captured by Ukrainian forces. On 10 December Czekanowski travelled to Paris via Warsaw and Prague. There he worked for the Polish Delegation at the Versailles peace conference, acting as an expert and later as a member of the Delegation Council. From 1 March to 1 May 1919 he served as political secretary on the Polish National Committee, from 1 May to 15 June he worked for the Polish Delegation, and from 15 June to 1 October 1919 he headed the offices of the Delegation, in September replacing the Delegation’s secretary general Stanisław Koziekry. Recalled to Paris, he acted as scientific expert to the Polish Delegation until 15 March 1920.

By a decision of the Chief of State dated 9 April 1920, Jan Czekanowski was appointed full professor of ethnology and anthropology at Jan Kazimierz University in Lvov, with effect from 1 January 1920. He returned to Lvov, and on 15 April 1920 he again began lecturing at the university.

In 1913 the Warsaw Scientific Society had published a book by Jan Czekanowski titled *Zarys metod statystycznych w zastosowaniu do antropologii* (“Outline of statistical methods in application to anthropology”). This was the first statistical textbook written in Polish to describe modern methods of handling empirical data and proper interpretation of results. It was published just two years after the appearance of the world’s first textbook of modern mathematical statistics, *An Introduction to the Theory of Statistics* by George Yule, and it played a great role in making biometrics better known among Polish scholars before World War I and in the interwar period. Besides descriptive statistics, this thoroughly modern and precise textbook covers the topics of reasoning based on the correlation coefficient, multiple regression with worked examples, as well as the diagraphic taxonomical method of Czekanowski. I would encourage any authors undertaking work on a modern textbook of statistics to study Czekanowski’s example from almost a hundred years ago.

It is incontestable that Czekanowski made a huge contribution to statistics. This superb scholar also made a greater contribution than anyone else to the flourishing of Polish anthropology, and caused it to gain worldwide renown. Professor Czekanowski was the founder of the Lvov School of anthropology, which set the tone for all research carried out in Poland over many years. It is therefore also referred to as the Polish School of anthropology, distinguished by a totally original approach to individual intrapopulational taxonomy of humans.

Professor Jan Czekanowski was a member of the Lvov Scientific Society. Active local members of the third section, devoted to mathematics and natural science, also included Stefan Banach and Hugo Steinhaus, while members active elsewhere included Marie Curie (Paris), Wacław Sierpiński (Warsaw) and Stanisław Zaremba (Cracow).

In the years 1934–1936 Czekanowski held the post of rector of Jana Kazimierz University in Lvov.

After the arrival of German forces in Lvov, on 30 June 1941 Jan Czekanowski was deprived of the ability to continue work at his beloved Anthropology Department. Thanks to a Ukrainian doctoral student his name was removed from the list of Lvov professors who were shot by the Germans on 4 July 1941. Because he kept his most important materials and books at home, he was able to carry on intensive scientific work even during the occupation. He obtained formal protection from the *Arbeitsamt* by taking up the administration of the estates of Kośmin near Grójec, based on a notarized power of attorney. This also enabled him to place his family in the village of Głuchów and to travel there and also to Warsaw, where he took part in underground educational activities. The Kośmin estates were owned at that time by his elder brother Stanisław, who had formally received them from his father Wincenty in 1895.

On 8 May 1944, Jan Czekanowski and his family left Lvov and, until 14 September of that year, stayed as guests with Professor Jerzy Fuhrich at Broniszów near Ropczyce. Later, taking advantage of a change in situation caused by the movement of Soviet forces to the Wisłoka line, he moved to the village of Cmolas near Kolbuszowa, and taught at the secondary school in Kolbuszowa until the end of April 1945. (The primary school in Cmolas now bears the name of Jan Czekanowski as its patron.) Thanks to the intervention of the Education Ministry, and having received a truck from the Provincial Offices in Rzeszów, he moved to Lublin, where he lectured in anthropology at the Catholic University of Lublin – he had received an appointment from that institution in November 1944, but was not able to travel there earlier because of lack of means of transport.

By letter of the President of the National Council, Bolesław Bierut, dated 28 February 1946, Czekanowski was appointed full professor of anthropology in the Faculty of Medicine at Poznań University. He took up the duties of professor and the chair of anthropology on 1 March 1946. After that faculty was transformed into the independent College of Medicine, he joined the Faculty of Mathematics and Natural Science, and later, when that faculty was divided, the Faculty of Biology and Earth Sciences.

While holding his chair at Poznań he continued to lecture at the Catholic University of Lublin until 1949, when the Ministry refused to allow him to continue working at two universities.

Described below are a few episodes of interest from the career of Jan Czekanowski:

1. During World War I he produced statistics on nationality and religious denomination in the Polish lands for the use of the future Polish Delegation at the Versailles peace conference, which he attended as an expert and as head of its offices. He presented to President Wilson a concept for the positioning of the eastern border which would have placed the same number of Orthodox Christians on the Polish side of the border as Catholics on the Russian side.

2. In the German Nazi period, Czekanowski challenged as Utopian the idea that in prehistory there existed pure racial types such as Germanic, Slav and Ugro-Finnish. He demonstrated this by making measurements on Polish army conscripts. He showed that the highest contribution of the Nordic element, and thus the greatest closeness to the Nazis' Aryan ideal, was found in young Jews who came from Warsaw.

3. The Karaim national minority was spared the fate of the Jews and Gypsies only because, when questioned by the Germans in 1942, Jan Czekanowski gave authoritative confirmation of their Turkish origins.

In 1960, on grounds of age, Jan Czekanowski went into retirement. However he continued to give a seminar in anthropology for master's degree students specializing in that field.

Professor Jan Czekanowski's scientific work was exceptionally wide-ranging. He had a multifaceted mind, interested in many different issues relating to human life and to human beings themselves. However he was able to achieve his greatest successes in theoretical anthropology, by applying statistical methods to anthropometric materials, in ethnography and ethnology, and in Slavic studies, where he provided strong justification for the theory that the original Slavic homeland was situated between the Vistula and Oder rivers. This view was opposed most strongly by German and to some extent by Czech scholars, and even by some from Poland, who were not convinced by the reasoning and documentation put forward by Czekanowski.

Poland recognized his achievements. He was a full member of the Polish Academy of Sciences. Two universities – Wrocław in 1959 and Poznań in 1962 – awarded him the highest available title, that of doctor *honoris causa*. The Polish government awarded him the honours of Commander's Cross of the Order of Polish Rebirth and the Order of the Standard of Labour, First Class.

He was an honorary member of the Polish Anthropological Society, and also honorary member of the anthropological societies of Brno and Zurich, and corresponding member of the Paris Anthropological Society and the Royal Anthropological Institute of Great Britain and Ireland. In 1923–1924 he chaired the Copernicus Polish Society of Natural Scientists. He was a member of the Polish Statistical Society, serving as its vice-chairman in 1937–1939. He was member, vice-chairman and chairman of the Polish Folk Studies Society and a member of the Polish Orientalist Society. He was a founding member and chairman of the Scientific Council of the Polish Biometric Society, from the Society's founding in 1961 until his death.

Jan Czekanowski died on 20 July 1965 in Szczecin. He is buried on the Avenue of Distinguished Citizens in Warsaw's Powązki cemetery. As a result of efforts by the anthropology community, the name of Jan Czekanowski has also been given to one of Poznań's streets.

The following description of Czekanowski comes from an extensive article devoted to the history of anthropology in Poland (T. Bielicki, T. Krupiński, J. Strzałko, *Historia antropologii w Polsce*. Przegląd Antropologiczny 1987, 53(1–2), pp.3-28):

Czekanowski was a scholar in the old, great, professorial style: a man of wisdom adored by some, admired by many and reviled by a few. This tall, grandly built man, with the penetrative gaze of his pale blue eyes, with an inseparable cigarette stuck to the corner of his mouth, could be seductively courteous and gentle-mannered, but also had a sharp tongue and could be caustic in polemics and discussions. He was a polyglot, who besides his native Polish had perfect mastery of German, French and Russian, and could also converse freely in English, Italian and Czech. Coming from the landowning classes, he was a "man of the world", close to a dozen European princes and princesses, and according to legend even to one crowned head. He was charming in company, and in his old age he liked to delight his listeners with spicy anecdotes, such as those about the parties held in private swimming baths in Zurich at the beginning of the century, or those about the revelries of the Russian cavalry stationed in Kalisz when it was a border town of the Russian Empire. He was an erudite person who was able in his time to speak authoritatively about matters of anthropology, Mendelian genetics, European archaeology, Slav linguistics, Slav and African ethnography, and mathematical statistics.

Sources:

The Archives of Adam Mickiewicz University in Poznań.

Chodzidło T., *Wspomnienie pośmiertne. Śp. prof. Jan Czekanowski (1882–1965)*. Zeszyty Naukowe KUL 1966, 9(3), pp. 91–94.

Ćwirko-Godycki M., *Profesor Jan Czekanowski*. Przegląd Antropologiczny 1965, XXXI(2), pp. 211–233.

Gajek J., *Jan Czekanowski. Sylwetka uczonego*. Nauka Polska 1958, 6(2), pp. 118–127.

Malinowski A., *Życie i działalność Jana Czekanowskiego*, in: J. Piontek et al. (eds.), *Teoria i empiria w polskiej szkole antropologicznej. W 100-lecie urodzin Jana Czekanowskiego*. Poznań 1985, pp. 71–77.

Perkal J., *Jan Czekanowski (1882–1965)*. Listy Biometryczne, 1965, 9–11, pp. III–IV.

Szeląg Z., *Grójeckie we wspomnieniach*. Series IV, Adam Mickiewicz Literary Society, Grójec Branch, 2004.

Wanke A., *Sześćdziesiąt lat pracy naukowej Jana Czekanowskiego*. Materiały I Prace Antropologiczne 1964, 70, pp. 7–27.

Wokroj F., *50-lecie promocji doktorskiej prof. dra Jana Czekanowskiego, 1906–1956*. Życie Szkoły Wyższej 1957, 7/8, pp. 93–96.

Mirosław Krzyśko

Adam Mickiewicz University in Poznań

Faculty of Mathematics and Computer Science

Umultowska 87, 61-614 Poznań

mkrzysko@amu.edu.pl

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 169—178

JERZY NEYMAN (1894–1981)

The Russian period, 1894–1921

Jerzy Spława-Neyman was born on 16 April 1894, into a noble family, in the town of Bendery on the river Dniester. He disliked the prefix Spława, and except for some early works he published under the name Neyman. Out of respect for that decision, we use only the shortened form of his surname here. Klonecki (1995) reports that, according to his sources, the Neyman family came to Poland in the 17th century from German or Dutch lands.[1]

Neyman was the grandson of a participant in the uprising of 1863. For his part in the insurrection his grandfather had been burned alive in his own house, his property confiscated, and his ten (according to J. Neyman) sons sent to Siberia. Only his youngest son Czesław – who would be Jerzy's father – was allowed to settle in Bendery in the European part of Russia. Czesław Neyman graduated in law in Kiev. There he also married Kazimiera Lutosławska. Jerzy Neyman, whose father was a successful man, was initially educated at home. He had a governess, and also attended an unofficial Polish school which operated in private homes. When at the age of ten he went to the secondary school in Simferopol, he knew five languages (French, German, Polish, Russian and Ukrainian) and was ahead of his colleagues with his knowledge in many fields, with the exception of Russian history and geography.

In 1906 Neyman's father died, and his family moved to Kharkov, where they had relatives. After he completed secondary school in 1912 his mother sent Jerzy with a group of students on a rail journey around Europe. In autumn of 1912 Neyman began his studies at Kharkov University. At first he was interested in physics, which was a result of the publication at that time of the theory of relativity and the recent Nobel Prize awarded to Marie Curie. In 1914 he went on

[1] He disliked the prefix Spława, and except for some early works he published under the name Neyman. Out of respect for that decision, we use only the shortened form of his surname here. Klonecki (1995) reports that, according to his sources, the Neyman family came to Poland in the 17th century from German or Dutch lands.

a students' academic expedition to Mongolia. However, because he had no talent for manual laboratory work, he dropped physics that same year and began to study Lebesgue's book *Leçons sur l'intégration et la recherche des fonctions primitives*. This resulted in a paper on the Lebesgue integral (530 pages of tiny dense manuscript in the Russian language), for which Neyman received a gold medal in 1916. During his studies Neyman attended S. Bernstein's lectures in probability and mathematical statistics. In his introduction to *Early Statistical Papers of J. Neyman* (University of California Press, 1967), their author began by giving thanks to Bernstein, from whom he had learnt to concentrate on genuinely difficult problems. In 1917 Neyman completed his studies and became a research assistant in the university's mathematics department, as well as a lecturer at Kharkov polytechnic and assistant to A. Przeborski.

The years 1917–1919 were extremely difficult. The First World War, the Bolshevik revolution and the civil war were not conducive to work and led to a marked deterioration in living conditions. In 1919 Neyman was diagnosed with tuberculosis and sent to the Caucasus. There he met the Russian painter Olga Solodovnikova, whom he married in 1920. Ten days after the wedding Neyman was arrested by the Russians and imprisoned for several weeks. In 1920 Neyman passed his master's degree exam and became a university lecturer. He also worked with Professor M. Yegorov in the field of agricultural experimentation.

The Polish-British period, 1921–1938

Following the Riga Treaty of 1921, under an exchange of families, Neyman went with his mother and grandmother to Poland. Thus he saw that country for the first time at the age of 27. They settled in Bydgoszcz, in the house of Neyman's brother Karol. His wife, who had typhus, remained for the time being in Russia. Neyman made contact with Professor W. Sierpiński, who studied the results from Neyman's aforementioned manuscript, and suggested sending one of them, which turned out to be a new result, to the journal *Fundamenta Mathematicae*. The paper was accepted and appeared in 1923 under the title *Sur un théorème métrique concernant les ensembles fermés*. Sierpiński hoped that, starting from the new academic year, it would be possible to obtain a post for Neyman at a Polish university. The most likely institution was the university in Lvov (Lwów). However Neyman wanted to begin working straight away, and became a senior statistical assistant at the National Scientific Agricultural Institute in Bydgoszcz, which was headed by Professor K. Bassalik. Initially he engaged in intense further studies of statistics and agricultural experimentation. Around the end of 1921 he obtained funds for a journey to Berlin and for the purchase of statistical journals and books there. Neyman spent more than a year in Bydgoszcz, and while there wrote several papers on applications of probability theory to agricultural experimentation.

In December 1922 Neyman began working for the National Meteorological Institute in Warsaw, where he looked after equipment and collected data. Also, in

1923, probably thanks to A. Przeborski, who had also come to Poland from Kharkov, Neyman became his assistant at Warsaw University. At the same time he began giving classes as a lecturer in mathematics and statistics at the Central Agricultural College (SGGW). He then had a total of 25 teaching hours a week at those two institutions. From 1924 he gave additional classes at the Jagiellonian University in Cracow, and from 1927 he also worked for the beet producers K. Buszczyński and Sons. In 1928 he organized a Biometrical Laboratory at the M. Nencki Institute. In order to enable his pupils and colleagues to publish their work, and to popularize his own ideas, he founded the journal Statistica and published it from 1929 to 1938. He also worked with the Institute of Social Affairs, the Central Statistical Office and other institutions.

Based on his papers on agricultural experimentation written in Bydgoszcz, in 1924 Neyman received the title of doctor of mathematics from Warsaw University. His examiners were the professors T. Kotarbiński, S. Mazurkiewicz, A. Przeborski and W. Sierpiński. It should be noted that a part of his doctoral thesis, which was published in Roczniki Nauk Rolniczych (Annals of Agricultural Sciences) in 1923, was translated into English and published in 1990 with extensive commentary in the journal Statistical Science. In 1928 Neyman gained his post-doctoral habilitation degree at Warsaw University.

In 1924, thanks to K. Bassalik and W. Sierpiński, Neyman received a one-year Polish government scholarship for a stay at University College London, with Karl Pearson. Among the results was the publication of versions of three earlier works of Neyman in the journal Biometrika. Next, with the support of Pearson and Sierpiński, Neyman received a scholarship from the Rockefeller Foundation, which he used for a year's stay in Paris, with Borel at the Sorbonne and with Lebesgue at the Collège de France. In 1926 he began working with Egon Pearson, the son of Karl. Their contacts were intensive, and in 1934 Neyman gained the post of lecturer at University College London, which solved his problem of having no permanent employment and no real prospects of obtaining a professorship in Poland, which had made his material situation very difficult. In spite of living and working in London, Neyman maintained contact and cooperation with his Polish team. He worked at University College until 1938. It should be pointed out that Neyman wanted to work in Poland. Reid (1982, p. 127) cites dramatic fragments of Neyman's correspondence in the matter of finding a suitable post for him at any institution in Poland.

In the course of those 18 very difficult years, Neyman managed to achieve an unimaginably great amount. A list of his works, included in the above-mentioned volume of early works of J. Neyman, includes 65 papers from 1923–1938, one textbook giving an introduction to probability theory, and two monographs written in Polish in 1933 and 1934. And these are not all of his publications, as can be seen from the bibliography drawn up by B. Łazowska (1995). Many of the works are extremely substantial, which sometimes even led to problems publishing them.

The list of his works naturally includes publications motivated by current application problems arising in connection with Neyman's work at the institutions mentioned earlier. They included in particular agricultural experimentation, biometrics, sampling methods and problems related to insurance.

As a result of questions asked by E. Pearson, Neyman became interested in the issue of hypothesis testing. In 1928 his first joint work with Pearson appeared, titled On the use and interpretation of certain test criteria for purposes of statistical inference, published in the journal Biometrika in two parts (pages 175–240 and 263–294). The work concerns mainly the likelihood ratio test, and introduces the concept of a set of alternatives, errors of the first and second type, the power function, and a definition of the likelihood ratio statistic. It is then shown that different known tests can be obtained by this general method, and an investigation is made of the asymptotic equivalence of the likelihood ratio test and the chi-squared test. As a result of further discussions with Pearson, Neyman formulated a problem of testing in the language of problem of optimization, and in 1930 proved the basic Neyman–Pearson lemma. This was included in 1932 in a paper of Neyman and Pearson concerning uniformly most powerful and uniformly best tests in a class of similar tests. That work, titled On the problem of the most efficient tests of statistical hypotheses, was accepted by the Royal Society, presented by Karl Pearson at the Society's meeting in November 1932, and published in 1933 in Philosophical Transactions of the Royal Society (pp. 289–337). The paper is of fundamental importance in the theory of the testing of hypotheses given a fixed sample size. As is noted by Le Cam and Lehmann (1974), by introducing tests as solutions to clearly defined optimization problems, Neyman and Pearson provided a model for general decision theory, later developed by A. Wald, and for mathematical statistics in general. In 1992 that work was selected for inclusion in a volume of the most important achievements in fundamentals of statistics in the 20th century (Breakthroughs in Statistics, Vol. I, Springer). A similar distinction went to Neyman's work On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, presented in 1934 at a meeting of the Royal Statistical Society and published in the Journal of the Royal Statistical Society (1934, pp. 558–625), which was included among the greatest 20th-century achievements in statistical methodology (Breakthroughs in Statistics, Vol. II, 1992, Springer). This work was based on a monograph of 1933, written in Polish and resulting from Neyman's work for the Institute of Social Affairs. In 1935, in Annals of Mathematical Statistics (pp. 111–116), Neyman published the paper On the problem of confidence intervals. In the summer of 1936 he continued to work intensely on confidence intervals, and presented his result on the duality of interval estimation and testing. E. Pearson rejected it for Biometrika as being too long and mathematical, but the work appeared under the title Outline of a theory of statistical estimation based on the classical theory of probability in 1937 in Philosophical Transactions of the Royal Society (pp. 333–380). It was presented at a meeting of the Royal Society by Jeffreys. In 1935, at a meeting of the

Industrial and Agricultural Section of the Royal Statistical Society, Neyman presented a joint paper written with K. Iwaszkiewicz and S. Kołodziejczyk on orthogonal designs and randomized blocks. This was printed in 1935 in a supplement to the Journal of the Royal Statistical Society (pp. 107–180). The paper, along with the work of R.A. Fisher, had great importance in the development of experimental planning. In 1937 Neyman published a paper in Skandinavisk Aktuarietidskrift (pp. 149–199) titled 'Smooth' test for goodness of fit, which proved another milestone in the development of statistics. In it he gave an asymptotically optimal solution to the problem of testing the fit of a set of observations to a fully known continuous distribution. In this work Neyman introduced sequences of local alternatives (contiguous distributions), which in the 1960s became a standard tool of asymptotic statistics. The test introduced in that paper remained almost completely forgotten for years, although that has changed radically in recent decades.

In 1935 Neyman and E. Pearson founded a new journal called Statistical Research Memoirs. In 1936 Neyman's son Michael was born. In 1937 Neyman was invited to an international probability congress in Geneva. In addition S. Wilks invited him to give a series of lectures in the United States. On his travels in the States his work aroused much enthusiasm, and the visit itself was a huge success. In November 1937 G. Evans sent Neyman an invitation to set up a statistical centre in Berkeley, California. He was also offered a professorship at Ann Arbor, Michigan. In 1938, a few days after his 44th birthday, Neyman accepted the Berkeley offer. Among other things, this decision meant that he would escape the consequences of the Second World War in Europe. It should be remembered that many of Neyman's Polish colleagues and students died during the war. E. Scott (2006) writes that in 1952 Neyman dedicated to them an extended edition of a volume of his thoughts on statistics, titled Lectures and Conferences on Mathematical Statistics and Probability, listing their names and how each of them died. The first edition of the volume (edited with the assistance of W. Deming) had appeared in 1938 under the title Lectures and Conferences on Mathematical Statistics. The book gained great renown in the United States, and helped to popularize Neyman's ideas and results.

The American period, 1938–1981

On 12 August 1938 Neyman arrived in Berkeley and set to work with great vigour. He worked on setting up the Statistical Laboratory and giving numerous lectures (for example, in 1939–1940 he lectured for 25 hours a week). He began gradually to assemble a team. Elizabeth Scott, an astronomy graduate, became his assistant. He also employed E. Fix, but unfortunately was not able to obtain a post for A. Wald, who had escaped from Nazi persecution in Europe. Neyman gained his first distinctions from the American statistics community: he was invited to give a lecture at a joint conference of the American Statistical Association and the International Statistical Institute, and became a member of the organizing committee of the 10th Mathematical Congress and an editor of the journal *Annals of Mathematical Statistics*. The outbreak of war and aggression towards Poland distressed him deeply. He made efforts to help his fellow Poles. Among other things, through the Kosciuszko Foundation, he arranged a scholarship for A. Zygmund, which enabled the latter to emigrate with his family to the United States and probably saved his life. In 1942 E. Lehmann became Neyman's assistant. However racial problems meant that he was not able to employ D. Blackwell. Neyman not only set up the Laboratory, but also began to work closely with many university faculties at Berkeley (Genetics, Geology, Hygiene, Agriculture), and this activity was valued very highly.

In February 1942 Neyman was engaged to solve optimization problems for the military. The project was carried out at the Berkeley Statistical Laboratory, with varying intensity, until the end of the war. In October 1944, together with a group of American mathematicians, he was sent to England for the purpose of researching the effectiveness of certain bombs. Also in 1944 he gained American citizenship. He was also able to bring P. Hsu to work for a time at Berkeley.

In 1945 Neyman organized a symposium in statistics and probability, at which he presented a paper titled *Contribution to the theory of the chi-square test*, which among other things introduced the class of best asymptotically normal (BAN) estimators, which are much more convenient to use than the classical estimators obtained by the method of maximum likelihood, and are useful in many complex problems. The symposium was a great success. The holding of the symposium was motivated by a desire to celebrate the end of the war and to facilitate a return to theoretical research following several years of work on applications for the American military. In 1946 Neyman was invited by President Truman to join a team of international observers for the Greek elections. That summer he was invited to spend a semester at Columbia University, where A. Wald worked, and where Neyman was offered a professorship and numerous privileges. These offers proved effective as a means of applying pressure to gain significant advantages for his Laboratory at Berkeley. In particular, posts were found there for M. Loève and C. Stein. In 1947 Neyman was elected vice-president of the American Statistical Association, and in 1948 he became president of the International Statistical Institute. Recognition for his achievements can also be seen in the fact that most of the papers appearing in *Annals of Mathematical Statistics* at that time related to problems which had been set and considered in earlier works by Neyman. In 1948 Neyman and Pearson renewed publication of *Statistical Research Memoirs*; the series continues to be published today under the new title *University of California Publications in Statistics*. After ten years of Neyman's activities Berkeley had become one of the two strongest centres for statistics in the United States, the other being Columbia University.

In 1949 Neyman took his first sabbatical to Europe. First he visited London, where he gave lectures and had discussions with Pearson. Next he lectured in Paris. While there he met L. Le Cam, and recruited him to the Laboratory. He also received a great distinction while in Paris: he became the first non-French author to be asked to work on a volume in the Borel Series. After Paris he visited Warsaw and many other Polish cities. He also met with his brother Karol.

In 1950, after Neyman's return from Europe, a second Berkeley Symposium took place. Neyman was constantly fighting for the Laboratory's funding and position. The situation was so difficult that Neyman gave up his work on his contribution to the Borel Series. Problems were multiplied by the death of A. Wald in an air crash, and the consequent attempts by Columbia University and other institutions to take over part of Neyman's group. In 1951, in response to questions put by the astronomer C. Shane, Neyman and Elizabeth Scott began a long period of intense collaboration on the dynamics of galaxies. This led to a series of around twenty papers, which are regarded as being among Neyman's most important works on applications. For several months in the academic year 1952/1953 Neyman worked in Bangkok, helping P. Sukhatme to organize a centre for training in sampling methods. In 1953 Neyman employed D. Blackwell and H. Scheffé. Also in 1953 he separated from his wife Olga.

In 1954 a decision was taken to set up a Department of Statistics at Berkeley. Neyman prepared the third Berkeley Symposium, where alongside work on probability and statistics there were papers presented in the fields of astronomy, physics, biology and health issues, econometrics, industrial mathematics and psychometrics. This trend was continued at subsequent Symposia. In the same year Neyman, A. Tarski and three other American mathematicians were invited to the Mathematical Congress in Amsterdam to give lectures on the future of mathematics. In 1955 the Department of Statistics began its work, under Neyman's direction. A year later Neyman resigned from that position, while retaining the lifetime post of head of the Statistical Laboratory.

In 1958 he took another sabbatical. He travelled widely, including to Poland. He also wrote his fundamental work on C(α) tests, which appeared in a volume dedicated to H. Cramér. Until the end of his life Neyman remained bitter that this work had not gained due recognition. Reid (1986, pp. 251–252) quotes a diplomatic statement of Neyman on that subject. The construction of C(α) tests, initiated by a modest publication by Neyman in 1954 in *Trabajos de Estadistica* (pp. 161–168), was key to the development of adaptive methods and asymptotically efficient semiparametric statistics. Unfortunately most works on these subjects make no mention of the originator of the significant idea behind them. E. Scott (2006) notes that during Neyman's lifetime his fundamental results quickly came into practice and found a place in basic textbooks, becoming "classical knowledge" in a sense, and for many it was no longer clear who their originator was. Neyman reconciled himself to the situation.

In 1960, although he had reached retirement age, Neyman continued to work intensively and obtained significant funds for further projects. He was awarded an honorary doctorate by the University of Chicago, became an honorary member of the Royal Statistical Society, and together with Elizabeth Scott received a prize from the American Association for the Advancement of Science. In 1960 the fourth Berkeley Symposium took place. In the following year Neyman spent much time travelling; he visited Leningrad, went to Moscow for a meeting with Bernstein, and also reached Kiev and Kharkov. A direct result of that visit was the arrangement of the translation into English of E. Dynkin's book on Markov processes.

In 1963 Neyman travelled in the southern states of the USA. Moved by the problems of race, he organized a collection of funds for scholarships, and wrote a letter to H. Cramér in the matter of a Nobel Peace Prize for Martin Luther King.

In 1964 Neyman celebrated his 70th birthday. In recognition of that occasion he was given an entry in the Great Book of the National Academy of Sciences, and received an honorary doctorate from Stockholm University. In 1965 a volume of papers dedicated to Neyman was published, edited by F. David. In 1966 he became the first non-Briton to be awarded the gold medal of the Royal Statistical Society, and the University of Berkeley published three volumes of work by Neyman and Pearson. Also in 1966 he became an overseas member of the Polish Academy of Sciences. We should also note that the fifth Berkeley Symposium took place in 1965.

In 1968 Neyman and Le Cam organized protests against the war in Vietnam. In spite of this, in 1969 Neyman became one of twelve Americans to receive the country's highest scientific award, the Medal of Science, "for laying the foundations of modern statistics and devising tests and procedures that have become essential parts of the knowledge of every statistician."

The year 1970 saw the holding of the sixth Berkeley Symposium, with an extensive programme related to biology and environmental pollution. The symposium was supplemented by three conferences held in spring 1971. It should be remembered that proceedings were printed for each of the six symposia, and Neyman was the editor or co-editor of each one of these ever more voluminous works. Also in 1971 Neyman and A. Zygmund began work on a collection of essays on various revolutionary changes in science, which they referred to as "Copernican". The volume, prepared for the 500th anniversary of the birth of Copernicus, and titled *The Heritage of Copernicus: Theories "More Pleasing to the Mind"*, was published in 1974 on the occasion of Neyman's 80th birthday.

In 1974 a meeting *To Honour Jerzy Neyman* took place in Warsaw, and a collection of the papers presented there was published in 1977. Neyman received honorary doctorates from Warsaw University and the Indian Statistical Institute. Volumes of *Annals of Statistics* and *International Statistical Review* were dedicated to him. There was also founded a "Jerzy Neyman Lectureship in Mathematical Statistics". In 1979 Neyman became an overseas member of the Royal Statistical Society.

Neyman's American period produced several works of great importance for the development of asymptotic statistical methods, such as BAN estimators and C(α) tests. However the main topic of interest for Neyman in that period was the building and verification of probabilistic models for a number of natural phenomena. The first paper in that series was published in *Annals of Mathematical Statistics* in 1939 (pp. 35–57) with the title *On a new class of* *`contagious' distributions, applicable in entomology and bacteriology*, and concerned the modelling and analysis of clusters. Subsequent work concerned matters of the formation of clusters with regard to modelling of the spread of epidemics and modelling of the distribution of galaxies in the universe. For more than twenty years Neyman worked on problems of weather modification. He was also interested, among other things, in carcinogenesis, the dynamics of population growth, and analysis of competing risks. Analysing his papers written in Poland and during the American period, we find that more than a half of Neyman's approximately 200 publications relate to matters of applications. More details concerning the entirety of Neyman's work can be found in reports by Klonecki and Zonn (1973), Le Cam and Lehmann (1974), Le Cam (1995) and Scott (2006).

J. Neyman died at Berkeley on 5 August 1981. He had remained active until the very end of his life. In June 1981 he had attended a conference on cancer, organized jointly with Le Cam. Even the day before his death he was working in hospital on a book on the subject of weather modification.

Finally we recall the view expressed by Elizbeth Scott (2006), who knew Neyman well – she wrote that Neyman always spoke of Poland with tenderness, and that he was proud of its heritage, although sometimes he could be critical of the actions of the Polish authorities.

Sources:

Kendall D.G., Bartlett M.S., Page T.L. *Jerzy Neyman 1894–1981*. Biographical Memoirs of Fellows of the Royal Society 1982, 28 , pp. 379–412.

Klonecki W. *Jerzy Neyman (1894–1981).* Probability and Mathematical Statistics 1995, 15, pp. 7–14.

Klonecki W., Zonn W. *Jerzy Spława-Neyman*. Wiadomości Matematyczne 1973, XVI, pp. 55–70.

Le Cam L., Lehmann E.L. *J. Neyman. On the occasion of his 80th birthday*. Annals of Statistics 1974, 2, pp. vii–xiii.

Le Cam L. *Neyman and stochastic models*. Probability and Mathematical Statistics 1995, 15, pp. 37–45.

Lehmann E.L. *Jerzy Neyman 1894-1981*. In: *Biographical Memoir*. National Academy of Sciences. Washington, D.C. 1994, pp. 395–420.

Łazowska B. *Bibliografia prac prof. dr Jerzego Neymana (1894-1981).* Zestawienia Bibliograficzne 22. Centralna Biblioteka Statystyczna im. Stefana Szulca. Warsaw 1995.

Reid C. *Neyman – from life*. Springer. New York 1982.

Scott E.L. *Neyman Jerzy*. In: *Encyclopedia of Statistical Sciences* 8. Wiley-Interscience. New York 2006, pp. 5479–5487.

Teresa Ledwina

Institute of Mathematics of the Polish Academy of Sciences

Branch in Wrocław

Department of Mathematical Statistics

Kopernika 18, 51-617 Wrocław

ledwina@impan.pan.wroc.pl

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 179—186

100 YEARS OF THE POLISH STATISTICAL ASSOCIATION

Czesław Domański

1. The origins of statistics in Poland

The beginning of statistics in Poland is connected with the following facts:

- first estimates and population censuses;
- first publications in the field of statistics ;
- initiation of statistical literature;
- formation of statistical administration;
- first lectures in statistics.

The first ever estimations of population of Poland were supplied by a number of authors: Józef Wybicki estimated the number of population in 1777 at the level of 5 391 364 people; Aleksander Busching in 1772 gave the number of 8.5 million, Stanisław Staszic in the year 1785 provided the estimated number of 6 million; Fryderyk Moszyński in 1788 r. produced the number of 76 3544 620 people.

2. Statistical Institutions

The beginnings of statistical activities on the Polish territories coincide with the proceedings of the so-called Four Year Parliament Session, i.e. the years 1788-1792. The Parliament adopted a resolution on carrying out in 1789 the first national population census combined with smoke registration. The census results were to help the Parliament to pass a law on a new tax, which was supposed to provide money towards expenses on a permanent, one hundred thousand army. The author of the statistical tables of the 1789 census and a statistical method of the military tax calculation was the deputy Fryderyk Józef earl Moszyński (1737-1817).

In 1864 an organizational unit of the Warsaw Municipal Council called the Statistical Section was established. Until the year 1876 its primary objective was to prepare materials which would appear in an annual publication entitled: „Obzor goroda Warszawy”. Since 1877, after the Section had increased the range of its statistical activities, it started to act as a statistical office of the city of Warsaw. Since its foundation the Statistical Section was headed for over 30 years by an economist and a statistician Professor Witold Załęski. His scientific output related to the development of the statistical thought includes the following works: „An Outline of the Theory of Statistics” (1884), „Remarks on the Theory of Statistics” (1888), „The Kingdom of Poland – a statistical approach” (1900-1901), „On Comparative Statistics of the Polish Kingdom” (1908). Załęski’s handbook „An Outline of Theory of Statistics”, encompasses five chapters: *Statistics as a method and science, The history of statistics, The history of administrative statistics, Congresses of statistics, Statistics organization*. The book can be considered the first Polish handbook of statistics.

In 1866 in National Department a project was conceived on establishing a statistical office in Galicia. The project was presented by Mieczysław Marasse (1840-1880) the author of, among others, a dissertation entitled „On Conception and Aim of Statistics” (Kraków, 1866), which was the first Polish publication devoted to the theory of statistics. The author defines there the tasks of statistics, which is subdivided into general and detailed, describes methods of statistics and presents three ways of statistical data compiling: tabular, graphic and descriptive.

A few years later in 1873 the National Statistical Office for Galicia was set up in Lvov and its activity continued until 1918. The founding father and the long-term head of the office was Tadeusz Pilat (1844-1923), a Professor of statistics and administration at the University of Lvov; also he was a co-founder (1885) and the first Pole among 100 members of the International Statistical Institute. Pilat was the first statistician who used estimations in statistical analysis and statistical inference.

The Cracow Municipal Statistical Office was set up in 1884. Its founder was Józef Kleczyński (1841-1900), a Professor of statistics and administrative law at the Jagiellonian University who had worked in the National Statistical Office in Lvov in the years 1875-1880.

J. Kleczyński published in „Polish Review” a lengthy article entitled „International Statistical Institute”, one of the earliest publications devoted to this institution. In 1891 he became the second Polish citizen to become a member of ISI. Kleczynski, who exerted a significant influence on the development of Polish statistics, published the following papers: „On calculating population number between censuses” (1879), and „ Municipal Statistical Offices”.

Another scholar who made a great contribution to the development of Polish statistical thought was a philosopher and an economist Augustyn Cieszkowski (1814-1894). He participated in the Second International Statistical Congress in 1855 r. in Paris as the only Polish representative and the speaker of one of the sections.

During the Congress a statistics of „foresight and future protection” was introduced. This was meant to assist people in making savings and insuring them against consequences of unfortunate events in future. The Congress awarded, among others, the following institutions: Savings Banks, Provident Societies, Pension Funds and Insurance Funds.

The first statistician who clearly defined the tasks of statistics was Ignacy Franciszek Stawiarski (1776-1835). He perceived statistics as the science which „includes all the wishes, demands and expectations of politicians and political economists, and using all available ways and methods provides a detailed description of the country's physical and moral powers. Moreover, using comparisons, calculations and probabilities draws conclusions aimed at improvement in the country's general well-being". „Statistics of Poland", following the example of ”*Statistique generale de la France”* (1806), was to be published in three volumes. Although the book was not completed, the tasks of statisticians described almost 200 years ago, remain basically the same.

Before the World War I a bold initiative of a group of Polish scientists was launched almost simultaneously in Cracow and Warsaw. The initiative was aimed at compiling a statistical publication which would be thematically, methodologically and organizationally independent of the partitioning countries: Russia, Prussia and Austria. The publication, which would encompass the whole territory of the partitioned Poland, was to be published in Cracow. For that purpose the first professional association of statisticians – Polish Statistical Association was established in 1912. The President of the Association became Juliusz Leo (1861-1918) – Professor of finance at the Jagiellonian University, and at the same time the Mayor of Cracow.

In 1915 Polish Statistical Association published „Statistics of Poland" edited by Professor Adam Krzyżanowski (1872-1963) and Professor Kazimierz Władysław Kumaniecki (1880-1941). The publication was the first comprehensive study which presented socio-economic development of the Polish territories from the beginning of the 19th century until the outbreak of the World War I. Kumaniecki, who was one of the founders of the Association and its Secretary, wrote "Studies in Migration Statistics" (1912), „Probability in Statistics" (1910), and many other works.

In the years 1915-1916 Professor of geography and cartography of Lvov University - Eugeniusz Romer (1871-1954), carried out works on compiling „Geographic and Statistical Atlas of Poland“. The atlas, published in Vienna in 1916 in three languages: Polish, French and German, contained 32 tables and 69 maps related to geography, history, demography, industry, agriculture, education and administrative and political entities. Moreover, abundant statistical material collected by Romer during his work on the atlas enabled him to edit, in co-operation with Ignacy Weinfeld, another important statistical publication entitled "Polish Yearbook. Statistical Tables" (Cracow 1917). The book, which came out in Polish, German and French, was comparable in size to Krzyżanowki and Kumaniecki's „Statistics of Poland" as it presented in numerical approach the economic situation and social life in three sectors of the partitioned Poland from the turn of the centuries until World War I.

Statistical yearbooks, „Geographic and Statistical Atlas of Poland" as well as other statistical and historical studies published in wartime proved to be extremely useful for the delegation representing Poland during peace negotiations in Paris (1919) and in Riga (1921).

3. Association of Polish Economists and Statisticians.

The Association of Polish Economists and Statisticians was founded in Warsaw in 1917. The activities of the Association were divided into five sections: economic theory, finance, statistics, economic policy and social policy.

The initial meeting of the Statistics Section was held on 14 January, 1918. The entire board of the Section consisted of the following members: the chairman – Professor Ludwik Krzywicki, the deputy – Professor Edward Grabowski, and the secretary – Stefan Szulc, MSc.

In 1921 the Association Council acknowledged the quarterly "The Economist" to be the official organ of the Association of Polish Economists and Statisticians. Therefore, "The Economist" can be assumed the first Polish statistical periodical with an economic angle.

In August 1929 Warsaw was hosting the 18th Session of the International Statistical Institute. The fact that Polish scholars were entrusted with the task of organizing the session was a sign of recognition for the role that Polish statistics played on the international arena.

The activities of thematic sections of the Association were put into a halt in the years 1929-1932. However, on May 29, 1933 the following sections were reactivated: theory of economics, economic policy, statistics and agriculture economics. The Statistics Section consisted of 14 members: Jan Derengowski, Michał Kalecki, Ignacy Krautler, Ludwik Landau, Zygmunt Limanowski, Stefan Moszczeński, Jerzy Spława-Neyman, Jan Piekałkiewicz, Franciszek Piltz, Edward Strzelecki, Edward Szturm de Sztrem, Stefan Szulc and Jan Wiśniewski.

4. Further activity of Polish Statistical Association.

The Statistics Section working within the framework of Association of Polish Economists and Statisticians was dissolved in December 1937 due to re-establishing of the Polish Statistical Association. The members of the Statistics Section of SPES became the founding members of the PSS.

The resolution on establishing of the Polish Statistical Association was adopted during a meeting of the Statistics Section of SPES held on December 16, 1936. The activity that PSS was engaged in at that time was very intense and fruitful. Polish Statistical Association published two periodicals: „Statistical Review" and „Statistics in Business". Its activities were carried out by four regional branches: Silesia-Dąbrowa – most active, Poznan, Vilnus and Lvov. Moreover, there were four sections: Mathematical Statistics, Statistics in Business, Economic Statistics and Population Statistics. In 1938 three volumes of "Statistical Review" came out together with five volumes of „Statistics in Business" and some other publications. The members of the board of the PSS included the following eminent scholars: Professor Stefan Szulc, Professor Edward Szturm de Sztrem, Professor Ludwik Krzywicki, Professor Edward Grabowski, assistant Professor Jan Wiśniewski, Professor. Jan Czekanowski and Jan Derengowski, MSc.

Within the framework of PSS there were two Scientific Commissions working: for the statistical terminology and for devising a guidebook of statistical sources. The latter one worked under the supervision of Wacław Skrzywan and succeeded in preparing a detailed plan and contents of the guidebook. Unfortunately, the effect of the commission's work was destroyed as a result of the outbreak of war.

In 1939 two volumes of „Statistical Review" came out in print and a number of materials were gathered for subsequent volumes. Also, two volumes of „Statistics in Business" were published. In Volume II of the „Statistical Review" a list of PSS members was included (as of June 15, 1939) and it comprised, 291 names of full members and 30 names of supporting members.

Numerous members of PSS who were not killed as a result of war activities, were deprived of the possibility of performing their normal scientific and professional duties. They got engaged in clandestine education activities and carried on some research getting ready for the after-war period. Many studies which were written by PSS members at that time were unique and had a lasting value e. g. „The Chronicles of War and German Occupation Years" by Ludwik Landau. Other members, who found employment in Statistical Office of the Governor General in Cracow, managed to save research materials of the Central Statistical Office.

After the war Professor Stefan Szulc, the head of the Central Statistical Office, undertook the task of re-activating the Polish Statistical Association, which took place in 1947. He was elected the Chairman of the Association. Despite increasingly unfavourable conditions resulting from political situation, PSS managed to continue its activities up to the end of 1950. In 1949 Volume III of the „Statistical Review" was published. In the years to follow (1953-1981) there was a long break in the existence of a separate organization of statisticians after PSS had been dissolved. Between 1953 and 1980 Polish statisticians conducted their activities in the Statistics Section of the Polish Economic Society.

The highlight of the period was the year 1975 when Polish statisticians were given a task of organizing 39th Session of International Statistical Institute.(ISS).

5. Reactivation of the Polish Statistical Association.

Another initiative aimed at reactivating the organization of Polish statisticians was taken in 1980 by a group of workers of the Central Statistical Office. At the initial stage a team responsible for carrying out the task was selected. The team was led by Doctor Jan Kordos and Professor Leszek Zienkowski, and the team members were: Lucjan Adamczuk, Kazimierz Latuch and others. The founders' meeting was held on 16 of April 1981 and was attended by 40 statisticians from Białystok, Lublin, Łodz, Olsztyn, Poznan, Rzeszow Szczecin, Torun, Warsaw and Wrocław. The resolution on establishing Polish Statistical Association was passed unanimously.

Polish Statistical Association strengthened its position on the international arena and since 1994 it has been affiliated with the International Statistical Institute. It has also achieved a prominent position at home and it is seen as a very active organization whose activities influence both scientific and social environment. Every year the Association is the organizer of two or three scientific conferences attended by members of international statistical community. PSS has a long tradition of organizing seminars and conferences devoted to discussing problems important for both statisticians and local communities. The papers presented at conferences are subsequently published in periodicals issued by the Polish Statistical Association and the Central Statistical Office: „Statistics in Transition", „Statistical News", „Statistical Quarterly" or in special monothematic volumes. At present the main activities of the Association are often supported by assistant bodies: Historical Section, Classification and Data Analysis Section, Mathematical Statistics and Bureau of Statistical Re6search and Analysis.

In 1990 Taxonomy Section, which later changed its name into Classification and Data Analysis Section of PSS, originated and since then it has organized its annual conferences. The most important results of scientific research presented during the conferences were published in 17 volumes entitled "Classification and Data Analysis – Theory and Application".

Classification and Data Analysis Section organized in 2002 in Cracow was one of those conferences, i.e. „The Eighth Conference of the International Federation of Classification Societies" - (IFCS).

In 2006 Polish Statistical Association was the co-organizer of 26th European Meeting of Statisticians, which took place in Torun. Since 1980 PSS has been organizing jointly with The University of Lodz the international conference „Multivariate Statistical Analysis – MSA". The conference proceedings have been published in 20 volumes of "Acta Universitatis Lodziensis". PSS also cooperates with the Economic University of Katowice on organizing international conference „Survey Sampling in Economic and Social Research".

Bureau of Statistical Research and Analysis plays an important role in the activity of Polish Statistical Association. It is an independent research unit which carries out statistical research commissioned by various scientific institutions and institutions of higher education. The profits generated by the Bureau go towards organizational and programme activities of the Polish Statistical Association. The financial means earned by the Bureau enabled the Association to undertake new tasks, improve the quality of audio-visual equipment, produce training films, organize scientific conferences, finance publications, etc.

6. Congress of Polish Statistics.

To celebrate the one hundredth anniversary of the Polish Statistical Association the Congress of Polish Statistics will be held on 18-20 April 2012 in Poznan, combining this event with the celebration of Polish Statistics Day in 2012.

The preliminary programme of the Congress comprises a number of thematic sessions, including the anniversary (historical) session, as well as others devoted to the methodology of statistical research, regional statistics, population statistics, socio-economic statistics, the problems of statistical data and the statistics of health, sport and tourism. The Congress will also host two panel discussions on: fundamental problems of statistics in the modern world, and the future of statistics.

7. Summary.

Let us now turn our attention to the most important achievements of two eminent members of the Polish Statistical Association, who played an active part in the scientific life of the Association - Jan Czekanowski (1882-1965) and Jerzy Neyman (1894-1981).

Researches aimed at finding a method in multi-feature analysis (1909) resulted in developing a diagraphic method (Czekanowski metod). The method enables to group and order a set of multi-feature individuals (each element is defined by a number of features). It includes two steps: 1) a matrix is introduced to the set of individuals, i.e. the distance between each two individuals of the set is calculated, which gives the distance table called Czekanowski table; 2) the numbers in Czekanowski table are replaced by correspondingly blackened fields, and columns and rows are moved in such a way as to get a diagram which would be possibly most blackened at the main diagonal. The obtained result is called Czekanowski diagram (J. Perkal, 1965).

This simple method of multi-feature analysis was used for over half a century to solve problems in anthropology, ethnography, psychology medicine, linguistics, musicology, botany and other fields. The method offers a number of variations (mean differences, features hierarchy) for distance calculation. Thanks to this method Czekanowski became a renowned biometrician. In 1904 he published in introduction to biometrics in anthropology handbook by R. Martin and in 1907 it came out in print as his doctoral dissertation. In 1913 Czekanowski wrote „An outline of statistical methods applied in anthropology". It was the first Polish handbook on modern methods of compiling numerical data and interpretation of findings. It is worth noting here that G. U. Yule published his handbook of mathematical statistics entitled "An introduction to the theory of statistics" just two years earlier. In the years 1913-1941 Czekanowski was a Professor and the head of Chair of Anthropology at the University of Lvov.

Jerzy Neyman arrived in England in 1938 and after some time he settled in Berkeley, USA. At that time statistics was not in the centre of interest of Berkeley University, however Neyman was determined to set up a Department of Statistics, which he later developed and changed the name into Laboratory of Statistics. He organized famous Berkeley Symposiums and showed great interest in astronomy and advanced statistics in less developed countries. This contributed to intensive development in scientific contacts at international level and in the seat of ISI in Holland. The International Association of Statistics in Physical Sciences was founded (I.A.S.P.S.). In 1975 I.A.S.P.S. was dissolved and replaced by Bernoulli Association which constituted a section of ISI. Bernoulli Association absorbed as its sub-section two other major groups of statisticians: The European Meeting of Statisticians (in the USA based on the Institute of Mathematical Statistics), and the international group cooperating on the organization of Conference of Stochastic Processes. All the above mentioned activities aimed at development of statistical movement on an international scale were inspired by Neyman.

Jerzy Neyman became a Professor of Statistics at Berkeley University and his disciples can be met all over the world. His *magnum opus* was the Neyman-Pearson theory of testing statistical hypotheses.

BIBLIOGRAPHY

DOMAŃSKI CZ. (2011), Setna rocznica powstania Polskiego Towarzystwa Statystycznego, Wiadomości Statystyczne nr 9, s. 1-10.

KRZYŚKO M. (2010), Jan Czekanowski – antropolog i statystyk, Kwartalnik Statystyczny nr 4, s. 31-37.

PERKAL J. (1965), Jan Czekanowski, Listy Biometryczne nr 9-11, s. 1-2;.OKTAWA W. (2002), Probabiliści, statystycy, ekonometrycy i biometrycy, Lubelskie Towarzystwo Naukowe, Lublin.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 187—190

THE POLISH STATISTICAL ASSOCIATION (PTS) – RE-ESTABLISHING

Władysław Wiesław Łagodziński[1]

For the hundred-year history of existence and activity of the Polish Statistical Association the year 1981 was an exceptional one since the Association for the first time in its history became part of the process of creating the basis of official statistics for a democratic society. We were set up both by the public demand and how painful need to cleanse science and statistical practice of the previous 25-35 years of availability, dependency, dependencies, bureaucracy, but also of conformism and opportunism to authorities.

Using the privilege of "an eye-witness" I wish to recall how the Association was established for the third time in its history[2].

Before re-establishing

Polish statistics started its official activity immediately after the Second World War, as the statisticians from Warsaw began to operate two days after the liberation of the city (19 January 1945), and CSO started its activity on March 12, 1945, when the front was still in the country, and there were two months left to the end of the war. The Association took action two years later, and on March 5, 1947 the Polish Statistical Association was re-entered into the register of associations and unions. Stefan Szulc, the President of the CSO, was elected the Chairman, and Dr. Kazimierz Romaniuk his deputy[3]. Short was the joy of the activity of both CSO and independent PTS. Already in 1949, Prof. S. Szulc retired, and Dr. K. Romaniuk moved to the Planning Committee, Central Statistical Office withdrew its grant and resigned its position of a supporting member. PTS lost its reasons to exist. Polish statistics were dominated by

[1] Vice-President of the Main Council of PTS. Chairman of the Council of Warsaw Department of PTS.

[2] See "Polish Statistical Association 1912-2012". Joint publication edited by Kazimierz Kruszka, Warsaw 2012. Polish Statistical Association General Council.; especially Part I, Chapters 4,5,6 and 7.

[3] See Lucjan Adamczuk and Kazimierz Latuch, "Reaktywowanie Polskiego Towarzystwa Statystycznego" In: Wiadomości Statystyczne, 1981, No.8.

economists and planners. The final blows for PTS were inflicted by I Congress of Polish Science (1951), and I Scientific Conference of Statistics Departments of Higher School of Economics (1952). The specialists in political economics of socialism and planners put then the finishing touch and consented to the creation of the Section of Statistics in the Warsaw Department of PTS. The final blow was the formal liquidation of PTS on April 4, 1955. And this is how PTS ceased to exist for 26 years.

In retrospect, it is difficult for me to find a good term for what actions we took in 1981. Was it an "establishing of the new PTS", "reanimation" or finally "re-establishing"? For me, it will always be a re-establishing, as in the minds and activity of Polish statisticians an extremely strong survival gene was always present, the gene of the public and science service. This was in the period of the Partitions of Poland, before and after World War I, before, during and after World War II, as well as in the years 1955-1981. Statisticians are reluctant to talk about this, because statistics is essentially the science of the country and it is intended to serve the state in accordance with the principles of science, truth and ethics of the profession. The year 1981 again set before us the questions of where we are in this country and in this science, what this statistics should look like and how we are to serve their country. The next decade showed that these questions were repeatedly returning and are still valid.

First stage of re-establishing

The period 1980-1981 in the Central Statistical Office, and consequently in the whole Polish statistics was very lively in political and union sense. The NSZZ "Solidarność", the most powerful trade union of all central government institutions (more than 1500 members) was active in the CSO. But before the association "Solidarność" was established, we had tried to set up an Independent Self-Governing Trade Union of Polish Statisticians, and only the great emphasis of the crew caused the creation of the circle "Solidarność". Later the idea of reactivating the PTS developed. Many members of the founding group of the "Solidarność" was also among the founders of the PTS. But while the people of “Solidarność" were a group of very radical views and even more radical actions, the group of founding members of the PTS was more moderate in general and represented a full cross-section of socio-professional personnel of statistics, that is official posts, views, political options, interest groups, academic intentions and perhaps also ideological ones.

Founding members - who were they?

The list of founding members who completed the Founding Declaration of Polish Statistical Association included 42 persons, but not all were present at this meeting. Currently, only some of them are living, although the author, despite strenuous efforts, was not able to contact all living founding members. Two groups prevailed: researchers and statisticians practitioners. Among the first ones were men of great stature such as professors Leszek Zienkowski, Wiesław Sadowski, Jerzy Holzer, Zdzisław Hellwig, Mikołaj Latuch, Zbigniew Pawłowski, Stanisław Wierzchosławski and Jan Kordos. Among the founding members were men who worked in the CSO in the thirties (Stanisław Róg, Maria Czarnowska, Józef Wojtyniak)[1]. The majority were heads of departments of statistical organizational units, lecturers from universities and middle-ranking management staff of statistical offices. To date, the following persons are actively participating in the activities of PTS authorities: W. Łagodziński as Vice President of the General Council, K. Kruszka, J. Berger and T. Jurek as members of the Main Council of PTS.

The founding meeting

On 16.04.1981, at the headquarters of one of the museums in Warsaw the founding meeting was held. The above mentioned founding members signed the following declaration:

"After reviewing the Programme Declaration and the Statute of the Polish Statistical Association presented at the Founding Meeting held on 16 April 1981 in Warsaw I confirm participation in the re-establishing of activity of the Polish Statistical Association. At the same time I apply for membership in the Polish Statistical Association as an ordinary member. I agree to comply with provisions of the Statute and the Rules, the Resolutions of General Meetings of Members (Deputies), to make regular payments of membership fees and to implement the goals and tasks of the Polish Statistical Association". Each declaration was confirmed by a signature. The declaration was a careful compromise for those times. What is important is that we did not obtain then any declaration of support from the CSO. This significantly influenced the next 10 years.

The Assembly passed the statute, elected provisional government, adopted a programme declaration, approved the establishment fund and adopted a resolution on the activity and the organizational rules of the Association. We also decided that we are a heir to the traditions and achievements of the Association. The description of the history of the Association and photocopies of original documents can be found in the mentioned publication, edited by Kazimierz Kruszka.

[1] Full list of founding members in “PTS 1912-2012, Op. Cit.

What is worth remembering today when referring to the 1981 convention?

When deciding on the re-establishment of the PTS, we knew that:

1) most academic and professional circles in the years 1950-1980 lost the ability to determine directions of its development and articulation of interests. This led to the superficial actions, loss of prestige and public confidence,
2) statistics deprived of broader social support and subjected to increasing pressure from the authorities lost systematically the status of social service transforming itself in a government agency,
3) statistical information became more and more widely the subject of manipulation. There were attempts at the central level to use it to bolster the ideology of success and prosperity of the Poles by using censorship and selection of statistical data far more than was reasonable, by running psychological mechanisms of self-censorship, blurring of personal responsibility for the content and quality of information and studies (one was also deprived of merit and job satisfaction for the accuracy of research and development, as well as high quality of publications W. L),
4) due to negative phenomena causing statistical information to fail to meet the reality, a crisis of public confidence in statistics started to grow.

All this meant a dramatic crisis in the social situation and the position of statistics, thus “...overcoming this crisis, restoring confidence in the statistics, re-establishing its status of a nationwide social service, raising its renown and prestige as a science...” were the most urgent goals of the PTS in 1981.

After 21 years, while reporting the events from the history of the PTS RE-ESTABLISHING in 1981, I cannot resist the impression that many of the observations and findings made back then retained its validity and its relevance to the present day.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 191—196

NEW ECONOMY – NEW CHALLENGES FOR STATISTICS

Mirosław Szreder[1]

1. Introduction

Congress of Polish Statistics organized in April 2012 seems to be a good opportunity for analyzing and evaluating processes to which statistics in Poland was exposed over the last 22 years, during the period of building democracy and the market economy. Fundamental changes in politics and economics commenced in 1989 have strongly influenced not only Polish public statistics, its institutions and activities, but also have become a new challenge for the science of applied statistics, and for the practice of statistical surveys. This paper involves a condensed description of new tasks which were taken up by Polish statisticians in new circumstances of the market economy, and gives an outline of some new challenges which they will face in the near future. Special attention is paid to statistical surveys.

2. Origins and directions of changes in statistical surveys

Considering sources of the evolution in designing and conducting statistical surveys in Poland over the last 22 years one should take into account two important processes. The first one covers deep changes that affected political, social and economic lives caused by transition from the centrally planned economy to the market economy. The other process, acting independently of the former one, was the general development of the theory of statistics and applied statistics, which in this period was especially fast, and was strongly supported by the progress in IT. An analysis of particular effects of changes which took place in the practice of statistical surveys in Poland, enables one very rarely to identify the single reason. More frequently, a combination of the two described processes, and some other factors, account for those changes.

[1] University of Gdańsk.

The main results of the evolution in statistical surveys in Poland after 1989 can be classified into three groups which involve:

1. significant increase in the areas of economic and social activities where quantitative (statistical) research gained the predominant position among all research works applied in those fields;
2. stronger harmonization of developments of the theory of statistics with practical needs of statistical surveys;
3. increasing efforts of statisticians aimed at providing high quality results of conducted surveys to large parts of the society, especially in topics which attract attention of many people.

In my opinion, the above three groups of the effects of Poland's transformation in statistics have had some impact on the economy and the social life, as a kind of feedback, and additionally they have influenced statistical education within the society.

3. Statistical surveys in new areas of economic and social activities

A system of democracy and principles of the market economy required essential changes in the rules of political and economic lives. The transformation also caused new institutions which were established and new fields of social and economic activities to need more advanced statistical descriptions and analyses.

Soon after the beginning of transformation processes in 1989, many economists and market analysts realized that there was a growing demand for marketing surveys and also for reliable market forecasts both at macro- and micro levels. It was connected, among others, with the needs for preparing business plans for various purposes, including enterprise restructuring and new investments. A development of market research and marketing surveys attracted a genuine interest in statistical surveys, and in statistics as a whole, among increasing number of market specialists. Consumers' needs and preferences constituted a new field of application of statistical surveys which virtually did not exist in the centrally planned economy. Initially, simple statistical techniques were employed in this kind of research, however in response to more complex and more dynamic market phenomena which gradually started to appear, more advanced methods were used. Marketing research created a strong stimulus for including more statistics in university curriculums for students studying economics and management. This, in turn, was accompanied by increasing amount of research output focused on quantitative measurements of market processes.

Another area of economic activity, which succeeded to attract attention of many statisticians was financial market with all its components, including stock exchange, banking system, and rapidly developing processes on this market. That was a new reality in Poland's economic life after 45 years of another economic system which did not need an exchangeable currency or financial markets.

Particularly interesting for statisticians become problems connected with stock exchange, investment funds, currency markets, and insurance. The amount of new issues which required professional analyses was so large that many scientists specializing in mathematics and physics joined the group of economists and statisticians who dealt with the new challenges. Majority of them gained new qualifications and remained in this circle for long, as Poland become a part of the international financial market, and had to cope with all new processes which developed there in the following years. Nowadays, it is difficult to imagine analyses of financial phenomena without applying adequate statistical techniques and tools. And although the applied methods and techniques are sometimes called financial rather than statistical, it is obvious that statistics is the science which works out and develops methods of discovering regularities or patterns in mass phenomena, including the ones which are hidden in financial series. It is worth noting that many advanced methods of economic modelling or statistical inference finds their initial applications in the field of finance. Recent turbulences on the world's financial markets enhanced the demand for proper standards and high quality of statistical analyses concerning processes developing on the domestic financial market. This is a challenge for Polish statisticians for today, and no doubt, also for tomorrow.

Out of many areas of social life, where statistical measurements and research play a crucial role, the most spectacular increase of interest in statistics has occurred in opinion polls. In democracy, the voice of public opinion is important not only for the society but for those who govern the country as well. It is essential that both these groups obtain reliable and accurate measurements of the state of public opinion, mainly expressed by opinion polls. "*Opinion polls enable people to count themselves, in order to find out how many (or how few) of them are there, and the awareness of the number is a starting point for building public opinion*" – says Professor A. Sułek, a Polish sociologist (see Sułek [2011], p. 331). In first several years after 1989, the field of opinion polls was left in Poland entirely to sociologists and pollsters. It did not manage to attract much interest of statisticians. This attitude has changed gradually, when the consecutive national elections brought campaigns with a large number of poor quality polls and disappointing election forecasts. Widespread criticism of the methodology used in opinion polls referred also to statisticians, who eventually took up the challenge. Common efforts made by sociologists and statisticians to improve the quality of opinion polls resulted in more accurate and more precise polls in following years. Higher standards in designing and performing polls have been adapted by majority of pollsters operating in Poland. There are some measurable effects of improved methodology used in opinion polls, which involve decreasing amounts of errors in election surveys and forecasts. For instance, the total error in exit poll conducted during the 2011 parliamentary election was five times smaller than the corresponding error in exit poll performed during the EU referendum in 2003. It seems very likely that in the following years ahead opinion polls will remain an interesting area for statisticians whose competences will be engaged in solving newly aroused methodological or practical problems, such as increasing proportion of non-response.

4. Adequate response of the theory of statistics to the needs of practice

A newly established political and economic system in Poland after 1989 caused an increase in the number of statistical surveys performed beyond the system of public statistics. Needs of economic practice have become an important factor which determines directions of the development of applied statistics, and defines specific issues which ought to be solved in various kinds of statistical surveys. This refers particularly to problems that appeared in quantitative market surveys, in analyses of financial phenomena, and in studies of the activities of small and medium-sized enterprises. Many statistical surveys related to social or economic issues have been carried out for local authorities.

As a response to practical needs one could observe an increasing interest of Polish statisticians in small domain inference (including small area estimation). This is one of the issues which have attracted research interests of statisticians representing both official and commercial statistics. A number of valuable scientific achievements in this field have their roots in original works carried out in Poland's Central Statistical Office (GUS). Therefore, some initial applications of new techniques representing small domain statistics can be found in surveys designed by institutions of public statistics. There also exist several academic institutions in Poland which have been successful in developing small domain methods and techniques. Taking into account increasing information needs expressed by many institutions and enterprises, including local authorities, one should expect further development of this branch of statistics in the future.

One of the crucial challenges for the practice of statistical surveys remains the problem of nonresponse which increasingly strongly affects the results of many surveys. This is the problem which refers to all kinds of statistical surveys, in every country. Efforts of statisticians focused on careful designing of a survey is frequently wasted due to large proportion of respondents who refuse to cooperate, as a result the obtained observations may create the sample structure which is significantly different from the designed one. Furthermore, a large proportion of nonresponses undermine rationality of applying classical inference based on this kind of data. It seems that many professionals who deal with designing and conducting surveys, and also statisticians who try to find efficient techniques which would compensate for the lack of observations realize that "*Today, nonresponse is a normal (but undesirable) feature of the survey undertaking*" (C.E. Särndal, S. Lundström [2005], s. IX). If this is the case, one can expect that the problem of dealing with nonresponses will remain one of important challenges for statisticians in the future. There has been a great deal of research output obtained in this field, including interesting proposals of new imputation and calibration techniques presented by Polish statisticians.

Eventually, however, additional non-sample information seems to be a decisive factor for the efficiency of all those techniques. In my opinion, further studies in this area will concentrate on searching for methods and techniques which would be combine the increasing amount of information, prior and sample one, about various populations that are investigated.

I think that the market economy in Poland, unlike the centrally planned economy, has inspired researchers to respond with new statistical ideas and new solutions to the needs of practice. Moreover, during the last 22 years there has been a tendency for widening the area of social and economic lives, in which statistical surveys have been successfully used on a regular basis. This process is likely to be continued, as the economic and social realities generate new information needs.

5. Statistical education in the society

One of the important challenges for statistics is to work out such measurement and description techniques which would, on the one hand, be adequate to increasingly more complex phenomena, and on the other, be easily and properly understood within the society. Reliable and good quality statistical data are sought not only by enterprises or administration units. In democratic societies, statistics should support new people’s initiatives with data and methods of data analysis. It relates, among others, to activities of non-government organizations in such areas as: environment protection, labour markets, vocational education, health-care and poverty. Promotion of high quality data and results of statistical analyses representing these areas can essentially help people or civic organizations reach the goals in their voluntary activities. Without a clear and credible description of the particular problem which they want to solve, their efforts are likely to be less productive. Quantitative descriptions are preferred at first stages of dealing with a problem, because numbers are able to define the problem more precisely and unambiguously. Polish statistics has been more accessible than in the past, and much has been done in order to provide the required statistical information to various organizations and groups of people whose intention is to do something good for a certain community. However, more can be done in this field, especially by official statistics.

Statistical education which would enable citizens to use and interpret quantitative facts properly is another important challenge. Unprepared people confronted with increasing amount of statistical information in mass media and elsewhere can feel lost or confused, if their individual observations do not confirm the data. This, as a consequence, can create suspicion that statistics generates inadequate pictures of the reality. Lack of confidence, which in such cases can be explained by lack of statistical education, may interfere communication within the society. In the long run, possible lack of trust in statistical data, or in methods of gathering and analyzing them, would be a serious problem in communication between democratic institutions and members of the society. Therefore, the problem of statistical education is now, and likely is going to be in the future, a challenge not only for statisticians but also for political and government bodies.

6. Conclusions

New challenges for statistics in Poland after 1989 have been created by processes connected with building democracy and the market economy, and also by world-wide tendencies like globalization and fast development of IT, which accelerated the need for higher standards of statistical research and surveys. In the last 22 years statistical research and surveys have come up in many new fields of social and economic activities. Statistical description and quantitative explanation of many processes in those fields have become more popular than qualitative ones, which were preferred in the past. The new circumstances constantly create new challenges for the applied statistics and for the theory of statistics. Special care should be paid to statistical education in the society which helps people understand quantitative description of the environment to which they belong.

REFERENCES

SÄRNDAL C.E., S. LUNDSTRÖM, *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd., 2005.

SUŁEK A., *Obrazy z życia socjologii w Polsce*, Wyd. Oficyna Naukowa, Warszawa 2011.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 197—200

" STATISTICS IN TRANSITION" AND "STATISTICS IN TRANSITION - NEW SERIES " - FIRST FIFTEEN YEARS

Jan Kordos

How did we in Poland start preparing a statistical journal in English? I remember that we discussed this issue, after the reactivation of the Polish Statistical Association (PTS) in April 1981, among members of the PTS and some colleagues in GUS. I also discussed this with Prof. Ryszard Zasępa[1] and next with Prof. Wiesław Sadowski[2], who was at that time the President of the Central Statistical Office of Poland (GUS) and the former Editor-in-Chief of "*Przegląd Statystyczny*" (Statistical Review). He advised us to establish a statistical journal in English as the journal of the Polish Statistical Association, and "*Przegląd Statystyczny*" would remain as it was, the journal of the Committee of Statistics and Econometrics and the Polish Academy of Sciences. Professor Sadowski supported that idea fully, and promised some assistance from GUS. We accepted the approach of Professor Sadowski, and discussed it with some colleagues from the PTS.

We realized that there were several problems to be solved, and one of them was a financial support of such an undertaking. In December 1985, when I was elected the President of the Polish Statistical Association[3], we started activities in different fields, and among others, we discussed how to get a permanent financial support for the Association's activities. We realized that the contribution of GUS was important but not sufficient. In 1986 the Main Council of PTS decided to establish the Bureau of Statistical Research and Analysis (BSRA) to get from that activity financial sources for different projects undertaken or supported by the Association. We prepared appropriate documents and sent them to the Ministry of Internal Affairs for approval. It was not an easy task, and Dr. Lucjan Adamczuk, a member of the Main Council of PTS, and I, had to convince the appropriate

[1] An international expert in sampling; Professor at Warsaw School of Economics; an author of the first Polish handbook in sampling; a chairman of the Mathematical Commission of GUS; cooperated with FAO and other international organization; cooperated with Prof. Jerzy Neyman.

[2] The President of the Central Statistical Office of Poland in 1980–1989; Editor-in-Chief of "*Przegląd Statystyczny*" (Statistical Review) in 1970s; Chairman of the Committee of Statistics and Econometrics of the Polish Academy of Sciences; Rector of Warsaw School of Economics (1965-1978), author of several books on mathematical statistics (translated also in English, Italian and Slovakian), theory of decision taking (translated in English and German).

[3] Prof. Jan Kordos was the President of the Polish Statistical Association in 1995-1994.

officials at the Ministry of Internal Affairs that the Bureau would be involved in statistical activity only. We should remember that in was before 1989, when the changes of the system begun. Finally, in March 1987 we got approval of the appropriate officials, and started different projects approved by the Main Council of PTS.

The possibility of issuing a statistical journal in English by the Polish Statistical Association was repeatedly discussed among members of the Association. We got financial support from activities of the Bureau of Statistical Research and Analysis of PTS. Finally, in April 1991, the Main Council of PTS decided to set up a Working Group chaired by Prof. Aleksander Zeliaś (University of Economics, Crakow), for preparing detailed proposals to start a new journal. Initially, the working title of the journal *The Polish Statistician* was adopted.

Decision of the Main Council of PTS

There were different ideas for the journal. Finally, we decided together with the statisticians from other countries, that it should be an international statistical journal published by our Association but edited by statisticians from different countries. The final name of the journal was established in October 1992 during the international conference on small area statistics[1].

I discussed this issue with a number of international statisticians, with Prof. Graham Kalton (USA), and Dr. Richard Platek (Canada), as my main advisors. We came to the conclusion that the most appropriate name would be *Statistics in Transition,* which essentially corresponded to the journal's initial aim to serve mainly countries undergoing transition. Besides, statistics is always *"in transition"* in a sense.

The creation of "Editorial Board"

The Working Group chaired by Prof. Aleksander Zeliaś suggested me as an Editor-in-Chief of the Journal. First, I had some objections to that suggestion since I was already involved in different activities. Finally, after some discussion, I accepted it, and the Main Council of PTS appointed me to these duties, and we started, overcoming some difficulties, to implement this task. The role of *Co-editors* was accepted by Prof. Tomasz Panek and Prof. Adam Szulc of Warsaw School of Economics.

It was necessary to select appropriate colleagues from Poland and abroad, i.e. *Associate Editors* (*AE*). First, we chose Polish AE, which was not easy, and then the *AE* from abroad. The latter came from Belgium, Canada, Holland, India, Italy, Japan, Mexico, the United Kingdom, and the United States, as well as from countries in transition, namely Bulgaria, Czech Republic, Estonia, Latvia,

[1] Small Area Statistics and Survey Designs, International Scientific Conference, Warsaw, Poland, 30th September – 3rd October 1992.

Lithuania, Romania, Russian Federation, Slovakia and Ukraine. I was able to establish cooperation with them due to my long-term (more than eight years) work abroad, participation in various international conferences and acquaintance with many eminent statisticians.

I must admit that I had no experience in running such a serious journal, but I received a considerable assistance in the early years of my work from my colleagues, here I would like to mention: Dr. Richard Platek from Canada, formerly Statistics Canada, Dr. M.P. Singh from Canada, the Editor-in-Chief of *"Survey Methodology"*, and as well from Dr. Lars Lyberg, Sweden, the Editor-in-Chief of "Official Statistics".

Over the course of the years natural changes followed and several of our close colleagues passed away. I shall mention here, above all, our *Associate Editors*: Prof. Ryszard Zasępa (Poland), Prof. Aleksander Ryszard Wójcik (Mexico), Dr. M.P. Singh (Canada), Prof. Ken Takeuchi (Japan), Prof. A. Revenko (Ukraine) and Prof. Aleksander Zeliaś (Poland).

The first issue of the journal

I would like to mention here only the first issue of the journal published in June 1993 which was devoted to various aspects of Polish statistics. That year we celebrated two hundred years of Polish statistics, and 75 years of establishment of the Central Statistical Office of Poland (GUS). The first paper, prepared by Dr. J. Oleński, President of GUS, dealt with different problems connected with transition of Polish statistics to the requirements of market economy. Prof. T. Walczak described seventy five years of official statistics in Poland on background of two hundred years of the Polish statistics. Prof. W. Wojciechowski prepared a note on the Committee of Statistics and Econometrics of the Polish Academy of Sciences. Prof. W. Welfe developed the concept of the computerized system of macroeconomic short and medium-term forecast and economic policy simulation for the period of transition towards the marked economy. Prof. Z. Hellwig considered estimation of linear regression parameters under scarcity of data. Prof. R.. Zasępa described sampling methods used in different household surveys in Poland, including application of sampling methods in population censuses in Poland as a means of speeding up tabulation in 1950 and 1960, and to broaden the scope of the censuses. Dr. J. Wywiał described the estimation of a mean for three sampling designs in a fixed population. There were also three notes on the following topics:

- The Polish Statistical Journal, by Prof. Cz. Domański,
- A short note on Small Area Statistics and Survey Designs, International Scientific Conference, held in 30th September – 3rd October 1992, by Mr. W. Łagodziński,
- The Research Bulletin of the Research Centre for Economic and Statistical Studies which was launched 1992, by Dr. A. Szulc.

I described the activities of the Polish Statistical Association founded in 1912.

Over the15 years of my work as the Editor-in-Chief of *SIT* and *SIT-ns*, we published 46 issues, i.e. over three issues per year, in eight volumes, comprising the total of about 8200 pages, 505 articles, 55 reports, 26 book reviews, and eight obituaries. We have published five special issues devoted to statistics of small areas, a number of articles on sample surveys in various countries, the methodology of household surveys, data quality and the methodology of the population census 2002. We also raised several statistical aspects observed in countries in transition and presented a lot of papers discussed at international conferences, some of them organized by the PTS. We promoted Polish statistics in the international arena, and enabled statisticians from other countries to present their achievements. In 2006 we received a proposal for cooperation with Springer Publishing Company, and I regret to say we failed to take advantage of this offer.

The first version of the journal, known as *Statistics in Transition (SIT)*, was published in the years 1993-2006 twice a year, with additional *"special issues"* published often as well. The journal provided a forum for exchanging views and experience in various fields of statistics, at the beginning regarding mainly the countries in transition from central planning to the market economy.

Gradually the scope of our interests was extended to include wider areas of application of statistical methods, and preparation for a new series of the journal began. *Statistics in Transition - new series (SIT-ns)*, was launched in 2007 as an electronic version of the Polish Statistical Association journal, published by the Central Office Statistics.

We agreed that the journal should be published three times a year (April, August, December). The new edition is to some extent a continuation of the previous release, regarding the numbering of volumes and the logo. However, *SIT-ns* is now taking a broader policy of presenting the application of statistical methods, educational statistics and its development.

After finishing my work as the Editor-in-Chief at the end of 2007, my duties are now limited to those of the *Founder/Former Editor*. I still continue to follow carefully the development of the journal and provide assistance as far as I can.

In 2008 Prof. Włodzimierz Okrasa was appointed a new Editor-in-Chief of *Statistics in Transition - new series*, Dr. Marek Cierpiał-Wolan was approved a new Scientific Secretary, and Dr. Roman Popiński became the Secretary. The issues published since 2000 are available on-line at the address: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

The year 2012 is not only the hundredth anniversary of the Polish Statistical Association, but also the twentieth year of the journal being issued. I hope that the journal will continue to be published in the future, serving Polish statistics and giving statisticians from other countries the chance to present their contributions to the development of statistics.

Jan Kordos

Editor-in-Chief in 1993-2007

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 201—202

STATISTICS IN TRANSITION NEW SERIES – TODAY

As a kind of post scriptum to the above note by Professor J. Kordos (Founder Editor), it seems worth mentioning that since the end of 2007 the Journal has been edited by a new Editor-in-Chief, with the help of also then newly appointed staff of the Editorial Office and with the support of the Editorial Board. Some innovations that have been introduced over that period seems to be worthwhile mentioning.

Responding to growing interest in public statistics, a new section devoted to *Current Issues in Public Statistics* was established in order to report on key issues related to the functioning of official statistics internationally, on subject matters that gain increasing importance in a country's data system and are appealing to community of producers and users of statistics, as well as to other parties concerned about the overall quality of public data, including policy makers and practitioners. It is noteworthy that it was commenced by addressing one of the issues of the most quintessential and vital problem in public statistics that concerns data access while protecting individual entities' data file against disclosure for non-statistical purposes, following the letter of the Polish Statistical Association on reported accidents of breaching the principle of statistical confidentially.

In 2010 Professor Vijaj Verma of the University of Siena, who visited Warsaw on the occasion of international conference on EU-SILC, submitted a proposal to introduce a Journal's new section on *Comparative Surveys*. The idea, which also was supported by Professor Kordos, was to stress the fact that growing demand for a more systematic presentation of huge works being done in the area of multi-population surveys might be seen as a new hallmark for the contents of the term 'statistics in transition'. Professor Verma kindly accepted to serve as a co-editor of this section. The papers included in this section are supposed to discuss issues relating to the meaning, generation and use of comparable data, cross countries, regions, sub- or super-populations or time. We encourage our collaborators and potential authors to submit such papers.

Currently, the Editorial Office of *Statistics in Transition new series* includes:

Włodzimierz Okrasa, Editor-in-Chief – Professor and chair of methodology at the Institute of Sociology of the University of Cardinal Stefan Wyszynski in Warsaw (UKSW), Advisor to the President of the Central Statistical Office of Poland/GUS. He served as a Head of Unit at the European Science

Foundation/ESF (Strasbourg (2000-2003) after his work for the World Bank in Washington DC (1993-2000), and for the Social Science Research Council N.Y., (1991-1993). He was an ASA/NSF (American Statistical Association and National Science Foundation) Senior Research Fellow at the US Bureau of Labor Statistics (1990-91), following teaching in American universities (Univ. of Maryland and Univ. of Mississippi), and researching at the Institute of Economic Sciences of the Polish Academy of Sciences. Research Scholar at the London School of Economics (The British Academy), at the University of Oxford, and others; author of more than 100 publications (most of them in English) in, among others, *European Economic Review, Review of Income and Wealth, Research on Economic Inequality*; *World Bank Working Papers.*

Marek Cierpiał-Wolan, Scientific Secretary – Ph.D. of Warsaw School of Economics, lecturer in econometrics and macroeconomics at the University of Rzeszów and the Rzeszów School of Business, Director of Statistical Office in Rzeszów; author of more than 50 publications in domestic and international journals.

Roman Popiński, Editorial Secretary – Ph.D., Political Science in Polish Institute of International Affairs, Lead Specialist of CSO/P Information Division (since April 2012, part-time)**.**

Beata Witek, Editorial Secretary, graduated in sociology (UKSW), joined the SiTns Office in April, 2012.

STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 203—218

Information on the Congress of Polish Statistics to Celebrate the 100th Anniversary of The Polish Statistical Association

In 2012 we will be celebrating **the 100th anniversary of the Polish Statistical Association (PSA)**, an organization created to integrate specialists involved in public statistical services as well as representatives of the academic community, local and economic government and agencies of state administration interested in the theory and implementation of statistical research. The Association contributes to the development of theoretical, methodological and practical aspects of statistical research and tries to promote statistical knowledge in society. It maintains cooperation with statistical associations in other countries and such organizations as Bernoulli - Society for Mathematical Statistics and Probability, International Society for Quality of Life Research, International Society for Quality-of-Life Studies or International Federation of Classification Societies. Polish Statistical Association is an affiliated member of International Statistics Institute.

To celebrate the 100th anniversary of The Polish Statistical Association we decided to hold the Congress of Polish Statistics on 18 – 20 April 2012 in Poznan and combine this event with the celebration of Polish Statistics Day. The Anniversary of the Polish Statistical Association, as well as the Congress of Polish Statistics are undoubtedly major events. It is our great honour to inform that the **President of the Republic of Poland, Mr Bronisław Komorowski has extended his Honorary Patronage over the Congress**, while numerous prominent scientists and economists have become members of its Honorary and Scientific Committees.

In 2018 we will celebrate 100th Anniversary of recovery of independence. The Central Statistical Office was the first unit of central administration of the independent Polish state. It was established by Regulation of the Regency Council already in July 1918. And it is worth to note that Polish statisticians established their official association even six years earlier, before recovery of the state! And now the Polish Statistical Association can boast 100 years of history and rich tradition.

The Polish Statistical Association was established in Cracow in 1912 and its main objective in the pre-war period was to prepare a statistical description of the traditionally Polish territories from the earliest times (http://www.stat.gov.pl/pts/9_ENG_HTML.htm). It was published in Cracow in 1915 and titled "Statistics of Poland" ("Statystyka Polski"). Juliusz Leo, its first President, was a professor of Jagiellonian University and the then President of the city of Cracow. Members of the PSA were famous Polish statisticians and economists, among others professors: Józef Buzek, Ludwik Landau, Jan Piekałkiewicz, Jerzy Spława-Neyman, Stefan Szulc. Contributors of the PSA are not only eminent Polish statisticians, but also a large number of anonymous people dedicated to serve science and the state.

In the interwar period the president of the Polish Statistical Association was Professor Edward Szturm de Sztrem and the deputy president was Professor Jan Czekanowski. The association was mainly a scientific and research one at that time. The official journal of the PTS was the quarterly "Statistical Review".

After World War II, in 1947 the PTS started its activity under the leadership of Professor Stefan Szulc. But in 1953 the decision on its liquidation was taken and entered in force in 1955. A part of the PSA members has become members of the Polish Economical Association, where the Section of Statistics was created. The last reactivation of PTS took place in April 1981.

In 1993, following the initiative of Professor Jan Kordos (the President of PSA in 1985 – 1994) and under his edition, the journal of the PSA edited in

English was funded. It is entitled „Statistics in Transition”. Currently the Polish Statistical Association has about 750 members organized in 17 regional branches. The PSA publishes jointly with the Central Statistical Office the monthly journal „Statistical News” („Wiadomości Statystyczne”) and the scientific journal of an international character entitled „Statistics in Transition – New series”.

In recognition of exceptional activities of statisticians form Poznan, the General Assembly of Polish Statistical Association, appointed Poznan as the localisation of the Congress. Acknowledged Poznan Universities, city and regional authorities were invited to participate in organization of the Congress.

In order to organize the Congress of Polish Statistics a special resolution was adopted by: Polish Statistical Association, Central Statistical Office, Poznan University of Economics and Statistical Office in Poznan:

Polish Statistical Association

Central Statistical Office

Poznan University of Economics

Statistical Office in Poznan

Congress of Polish Statistics is one of the most significant events in the history of Polish scientific thought and statistical practice. It marks the celebration of the 100th anniversary of the Polish Statistical Association, in recognition of its leading role and legacy.

The organizers of the Congress of Polish Statistics hope that it will provide a special opportunity for the exchange of ideas and experience between representatives of public statistics, research centres as well as other partners involved in investigating and monitoring social, economic and demographic processes. This form of discussion should also help to direct methodological and research work to be undertaken by the Polish statistical community in the years to come.

The Congress of Polish Statistics will have an international scope mainly by focusing on the presentation of the Polish contribution to the world repository of statistical knowledge. The international character of the Congress will also be highlighted by the participation of numerous representatives from foreign institutions and the promotion of its ideas and contributions outside Poland. We also look forward to the participation of leading statisticians, who will enrich the Congress with advances in research and applications. Numerous well-known scientists in the area of statistical research were invited to participate in and contribute to the Congress. The following respected statisticians have announced their participation in the Congress: Raymond Chambers, Malay Ghosh, Lorenzo Fattorini, Jean-Claude Deville, Reinhold Decker, Patrick Groenen, Wojciech Krzanowski, Nico Keilman, Francesco Billari, Achille Lemmi, Anne Valia Goujon and Li-Chun Zhang.

The Congress programme comprises a number of thematic sessions, including an anniversary (historical) session, a methodological session devoted to the methodology of statistical research, regional statistics, population statistics, socio-economic statistics, statistical data and statistics of health, sport and tourism. The Congress will also host discussion panels focusing on fundamental problems of statistics in the modern world and the future of statistics. The task of organizing the various sessions and panels has been undertaken by Polish most distinguished statisticians, recognised both at the national and international level. Please see the a detailed Congress programme attached.

Congress proceedings will take place in the plenary halls (aula) of Adam Mickiewicz University in Poznan and Poznan University of Economics. The Organizing Committee would like to invite everyone to join in the unique celebration of Polish statistics and take part in the Congress of Polish Statistics. Updated information about the conference are published on the Congress website: http://www.stat.gov.pl/pts/kongres2012/english/index.htm.

Chairperson of the Organizing Committee of the Congress:
Assoc. Prof. Elżbieta Gołata, UEP: e-mail: elzbieta.golata@ue.poznan.pl

Secretariat of the Organizing Committee of the Congress:
Statistical Office in Poznan,
ul. J. H. Dąbrowskiego 79, 60-959 Poznań: e-mail: kongres2012@stat.gov.pl
tel.+ 48 61 27 98 325 (Mo-Fr: 8-15),
+48 61 27 98 343 (Mo: 9-12, Tu: 7-10,We: 7-10),
fax. +48 61 27 98 101

The Programme of the Congress of Polish Statistics
Poznań 18 – 20 April 2012

Plenary sessions:

Anniversary session

Plenary session 1
Development of Polish statistical research

1. **Presentation of the achievements of Polish statistics, organizer prof. dr hab. M.Krzyśko**
   - **Walenty Ostasiewicz** (Uniwersytet Ekonomiczny we Wrocławiu), Development of statistical research in Poland
   - **Jan Mielniczuk,** Polish Contributions in the development of mathematical and applied statistics
   - **Tadeusz Caliński,** Development of Polish Statistical Research – Advances in biometrics
   - **Krzysztof Jajuga** (Uniwersytet Ekonomiczny we Wrocławiu), Development of Polish statistical research in economic sciences
   - **Janusz Witkowski, Tadeusz Walczak, Jan Berger** (GUS), Public statistics – historical development and modern challenges

Plenary session 2
Development of Polish statistical research

1. **Eminent Polish statisticians**
   - **Teresa Ledwina** (Instytut Matematyczny PAN), Jerzy Neyman **– invited paper**
   - **Mirosław Krzyśko** (UAM), Jan Czekanowski **- invited paper**
2. **Stanisława Bartosiewicz** (Wyższa Szkoła Bankowa we Wrocławiu), **Stańczyk Elżbieta** (Urząd Statystyczny we Wrocławiu)**,** A short overview of the socio-economic history of Poland in the years 1918–2008
3. **Jan Kordos** (Wyższa Szkoła Menedżerska, Warszawa)**,** Interdependence between the development of theory and practice in survey methodology in Poland

Discussion Panel
The Future of Statistics
Janina Jóźwiak (Szkoła Główna Handlowa w Warszawie), Population statistics

Janusz Witkowski (GUS), Social statistics

Tomasz Panek (Szkoła Główna Handlowa w Warszawie), Social statistics

Jerzy Wilkin (Uniwersytet Warszawski), Economic statistics

Jan Paradysz (Uniwersytet Ekonomiczny w Poznaniu), Regional statistics

Jerzy Andrzej Moczko (Uniwersytet Medyczny w Poznaniu), Health statistics
Mirosław Szreder (Uniwersytet Gdański), Statistical data

Jacek Koronacki (Instytut Podstaw Informatyki PAN, Warszawa), Methodology of statistical research

Parallel Sessions:

History and Development of Polish Statistics

1. **Józef Pociecha** (Uniwersytet Ekonomiczny w Krakowie), Circumstances surrounding the foundation of the Polish Statistical Association and its first President **– invited paper**
2. **Bożena Łazowska** (Centralna Biblioteka Statystyczna), The legacy of professor Władysław Bortkiewicz
3. **Cezary Kuklo** (Uniwersytet w Białymstoku), The family and household in Poland in the pre-industrial age – myths and facts **– invited paper**
4. **Witold Zdaniewicz** (Intytut Statystyki Kościoła Katolickiego, Warszawa), Contribution of the Statistical Institute of the Catholic Church to Polish public statistics

Mathematical statistics 1

1. **Tadeusz Bednarski** (Uniwersytet Wrocławski), A statistical analysis of causes of bias in labour market surveys **– invited paper**
2. **Wioletta Grzenda** (Szkoła Główna Handlowa w Warszawie), Using a semiparametric Bayesian Cox model to study determinants of long-term unemployment among young people
3. **Jan W. Owsiński** (Instytut Badań Systemowych PAN w Warszawie), An optimal partition of empirical distribution (and some problems assiociated with it)

Mathematical statistics 2

1. **Lesław Gajek** (Komisja Nadzoru Finansowego i Politechnika Łódzka), Modelling insolvency risk of investment companies using statistical methods – **invited paper**
2. **Daniel Kosiorowski** (Uniwersytet Ekonomiczny w Krakowie), Measures of position and dispersion in robust analysis of economic data streams
3. **Paweł Kobus** (Szkoła Główna Gospodarstwa Wiejskiego w Warszawie), The use of the Bayesian approach to estimate distributions of variables for selected units in a population using aggregate data

Mathematical statistics 3

1. **Jacek Koronacki** (Instytut Podstaw Informatyki PAN w Warszawie), Analysing multivariate data given a small sample size **– invited paper**
2. **Tomasz Górecki, Mirosław Krzyśko** (Uniwersytet im. Adama Mickiewicza w Poznaniu), Kernel principal component analysis

3. **Bronisław Ceranka, Małgorzata Graczyk** (Uniwersytet Przyrodniczy w Poznaniu) Estimating the total weight of objects in weighing designs

Mathematical statistics 4

1. **Zbigniew Szkutnik** (Akademia Górniczo-Hutnicza w Krakowie), EM algorythm and its modifications **– invited paper**
2. **Jan Mielniczuk** (Instytut Podstaw Informatyki PAN w Warszawie i Politechnika Warszawska), **Małgorzata Wojtyś** (Politechnika Warszawska), Post-model estimation of grade density
3. **Teresa Ledwina, Grzegorz Wyłupek** (Instytut Matematyczny PAN, Oddział Wrocław), A test for two samples given one-sided alternatives

Survey sampling and small area statistics 1 - English language session

1. **Malay Ghosh** (University of Florida), Finite population sampling: model-design synthesis **- invited paper**
2. **Ray Chambers, Gunky Kim** (Wollongong University), Regression analysis using data obtained by probability linking of multiple data sources **- invited paper**
3. **Imbi Traat** (University of Tartu), Domain estimators calibrated on reference survey

Survey sampling and small area statistics 2 - English language session

1. **Jean-Claude Deville**, **Daniel Bonnéry, Guillaume Chauvet** (ENSAE), Neyman type optimality for marginal quota sampling **- invited paper**
2. **Janusz L. Wywiał** (Katowice University of Economics), Estimation of population mean on the basis of a simple sample ordered by an auxiliary variable
3. **Wojciech Zieliński** (Warsaw University of Life Sciences), Statistical properties of a control design of controls provided by Supreme Chamber of Control
4. **Wojciech Gamrot** (Katowice University of Economics), On empirical inclusion probabilities

Survey sampling and small area statistics 3 - English language session

1. **Lorenzo Fattorini** (University of Siena), Design-based inference on ecological diversity **- invited paper**
2. **Tomasz Żądło** (Katowice University of Economics), On prediction of totals for spatially correlated domains
3. **Tomasz Klimanek** (Poznań University of Economics, Poznań Statistical Office), Using indirect estimation with spatial autocorrelation in social surveys in Poland
4. **Tomasz Józefowski** (Poznań Statistical Office), Using a SPREE estimator to estimate the number of unemployed across subregions

Survey sampling and small area statistics 4 - English language session

1. **Li-Chun Zhang** (Statistics Norway), Micro calibration for data integration **- invited paper**

2. **Sara Franceschi** (University of Tuscia), **Lorenzo Fattorini** (Università di Siena), Maffei D. (Università di Firenze) Design-based treatment of unit nonresponse by means of the calibration approach
3. **Marcin Szymkowiak** (Poznań University of Economics), Construction of calibration estimators of total for different distance measures
4. **Jan Kubacki** (Łódź Statistical Office), Estimation of parameters for small areas using the hierarchical Bayes method in the case of known model hyperparameters

Population statistics 1 - English language session

1. **Nico Keilman** (University of Oslo), Challenges for statistics on households and families **- invited paper**
2. **Irena E.Kotowska** (Warsaw School of Economics), Evolving population structures and their relevance for demographic and social change **- invited paper**
3. **Marta Styrc, Anna Matysiak** (Warsaw School of Economics), Socioeconomic status and marital stability

Population statistics 2 - English language session

1. **Francesco Billari** (Bocconi University, Milan), Challenges for new methods in demographic analysis (tentative) **- invited paper**
2. **Ewa Frątczak** (Warsaw School of Economics), Tradition and modernity in demographic analysis. From Graunt and Lexis to longitudinal models with fixed and random effect
3. **Marek Kupiszewski** (CEFMPR), What drives population change? **- invited paper**

Population statistics 3

1. **Elżbieta Gołata** (Economic University Poznań), Population census and truth
2. **Lucyna Nowak** (GUS), Development of statistical research in the field of demography and migration
3. **Wiktoria Wróblewska** (SGH), Changes in maximum life span and longevity – challenges for statistics
4. **Jerzy T.Kowaleski, Anna Majdzińska** (Uniwersytet Łódzki), Population ageing in EU countries – the near future and projections
5. **Jadwiga Borucka Joanna Romaniuk , Ewa Frątczak** (SGH), Duration of first (marital and cohabiting) unions in birth cohorts 1950-1970

Social Statistics 1 - English language session

1. **Achille Lemmi** (Siena University, Dipartimento di Economia Politica) Dimensions of poverty. theory, models and new perspectives **- invited paper**
2. **Anna Szukiełojć-Bieńkuńska** (Główny Urząd Statystyczny, Departament Badań Społecznych i Warunków Życia), Measurement of poverty and social exclusion in Polish public statistics

3. **Stanisław Maciej Kot** (Politechnika Gdańska, Wydział Ekonomii i Zarządzania, Zakład Statystyki), Stochastic equivalence scales for Poland
4. **Krystyna Hanusik, Urszula Łangowska-Szczęśniak** (Uniwersytet Opolski, Wydział Ekonomiczny), Determinants of variation in the level and structure of consumption of Polish households

Social Statistics 2

1. **Walenty Ostasiewicz** (Uniwersytet Ekonomiczny we Wrocławiu), Life quality as a subject of statistical research **– invited paper**
2. **Jolanta Perek-Białas** (Szkoła Główna Handlowa, Instytut Statystyki i Demografii), The situation of older generations in Central and Eastern Europe on the basis of EU-SILC data
3. **Anna Szukiełojć-Bieńkuńska** (Główny Urząd Statystyczny, Departament Badań Społecznych i Warunków Życia), Life quality in GUS surveys
4. **Krzysztof Szwarc** (Uniwersytet Ekonomiczny w Poznaniu, Katedra Statystyki i Demografii), The standard of living and variation in the demographic situation at NUTS 4 level in the Wielkopolska province

Social Statistics 3

1. **Urszula Sztanderska** (Uniwersytet Warszawski, Wydział Nauk Ekonomicznych, Katedra Makroekonomii i Teorii Handlu Zagranicznego), Polish public statistics as a source of inspiration for and an obstacle to the development of labour market research **– invited paper**
2. **Paweł Ulman** (Uniwersytet Ekonomiczny w Krakowie, Katedra Statystyki), The influence of economic activity of the disabled on the economic situation of their households
3. **Dominik Śliwicki** (Urząd Statystyczny w Bydgoszczy), Using a logit model in econometric analysis of gross salary
4. **Aleksandra Matuszewska-Janica** (Szkoła Główna Gospodarstwa Wiejskiego w Warszawie, Katedra Ekonometrii i Statystyki), Statistical analysis of the gender wage gap in Poland as compared to the EU countries by age, company type and occupational group

Social Statistics 4 - English language session

1. **Anne Valia Goujon, Ramon Bauer, Samir K.C., Michaela Potančoková** (Vienna Institute of Demography (VID) of the Austrian Academy of Sciences), The human capital puzzle: why it is so hard to find good data on educational attainment? **- invited paper**
2. **Marcin Stonawski** (Cracow University of Economics), Human capital and population ageing in Poland
3. **Agnieszka Chłoń-Domińczak** (Warsaw School of Economics), The use of indicators in the context of development of evidence-based social policy

4. **Dominik Rozkrut** (Urząd Statystyczny w Szczecinie), Differentiation of innovation strategies

Economic statistics 1

1. **Włodzimierz Siwiński** (Uniwersytet Warszawski, Akademia L. Koźmińskiego), A statistical overview of the 2008 crisis **– invited paper**
2. **Eugeniusz Gatnar** (Narodowy Bank Polski), The role of the National Bank of Poland in the system of Polish public statistics
3. **Barbara Pawełek** (Uniwersytet Ekonomiczny w Krakowie), A study of the usefulness of economic situation test results published by GUS for short-term projections of changes in economic activity in Poland using vector autoregression
4. **Andrzej Czyżewski, Aleksander Grzelak** (Uniwersytet Ekonomiczny w Poznaniu), Possibilities of using input-output statistics for macroeconomic assessment in economy

Economic statistics 2

1. **Barbara Liberda** (Uniwersytet Warszawski), Generational accounting by means of measuring the wealth of consecutive generations **– invited paper**
2. **Krzysztof Malaga** (Uniwersytet Ekonomiczny w Poznaniu), Dilemmas of modern statistics of economic growth
3. **Stanisław Lewiński Vel Iwański, Zofia Wilimowska (**Politechnika Wrocławska we Wrocławiu), Financial structure of Polish companies: statistical modelling
4. **Maria Parlińska, Robert Babiak** (Szkoła Główna Gospodarstwa Wiejskiego w Warszawie), Screening and risks of information asymmetry

Economic statistics 3

1. **Elżbieta Mączyńska** (INE PAN, Prezes PTE), Statystyka jako źródło danych w badaniach naukowych. Dylematy i niedostosowania - **invited paper**
2. **Monika Natkowska, Jacek Kowalewski** (Urząd Statystyczny w Poznaniu), Using a map of short-term statistics in the process of organizing business surveys conducted by public statistics
3. **Grażyna Dehnel** (UE Poznań), Indirect micro-estimation in business statistics
4. **Aneta Ptak-Chmielewska** (Szkoła Główna Handlowa), Conditions of the business sector in Poland. Statistics of business survival

Economic statistics 4

1. **Andrzej P. Wiatrak**, Contemporary problems of agricultural statistics **– invited paper**
2. **Robert Pacuszka** (Główny Urząd Statystyczny), Methods of agricultural production and their classification. An idea for a GUS methodological study

3. **Renata Bielak** (GUS) The role of statistics in the process of planning development policies – priorities and challenges
4. **Grzegorz Kowalewski** (Uniwersytet Ekonomiczny we Wrocławiu), Methodology of economic situation surveys

Regional statistics 1

1. **Jan Paradysz** (Centrum Statystyki Regionalnej, Uniwersytet Ekonomiczny w Poznaniu), Regional statistics: the current state, problems and directions **– invited paper**
2. **Dominika Rogalińska** (Główny Urząd Statystyczny), Challenges of regional statistics in the context of the debate about policy coherence
3. **Danuta Strahl, Małgorzata Markowska** (Uniwersytet Ekonomiczny we Wrocławiu), Possibilities of assessing the level of innovation in NUTS 2 regions in the light of Eurostat information resources
4. **Ewa Kamińska-Gawryluk** (Urząd Statystyczny w Białymstoku), Regional variation in Poland and selected EU countries (UE- 15)

Regional statistics 2

1. **Tadeusz Borys** (Uniwersytet Ekonomiczny we Wrocławiu), **Tomasz Potkański** (Związek Miast Polskich), Monitoring regional and local development
2. **Dorota Doniec** (Urząd Statystyczny w Katowicach), Regional accounts – methodological problems and practice
3. **Bartosz Bartniczak** (Urząd Statystyczny we Wrocławiu), The module of sustainable development indicators in the local databank
4. **Dorota Wyszkowska** (Uniwersytet w Białymstoku, US w Białymstoku), Regional statistics as an instrument of supporting the development of Polish provinces
5. **Jacek Batóg, Barbara Batóg, Magdalena Mojsiewicz** (Uniwersytet Szczeciński), **Katarzyna Wawrzyniak** (Zachodniopomorski Uniwersytet Technologiczny w Szczecinie), A statistical system of monitoring the implementation stage of the city development strategy

Regional statistics 3

1. **Janusz Dygaszewicz** (Główny Urząd Statystyczny), The agricultural census of 2010, the national census of 2011 and the use of GIS in public statistics
2. **Paweł Chlebicki,** (ESRI Polska), ArcGIS as a powerful took for statistical data visualization and analysis
3. **Sylwia Filas-Przybył, Maciej Kaźmierczak Dorota Stachowiak (**Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Miast), The use of administrative data in urban statistics
4. **Robert Buciak, Marek Pieniążek** (Główny Urząd Statystyczny, Uniwersytet Warszawski), Spatial classification of rural areas in Poland

5. **Roma Ryś-Jurek** (Uniwersytet Przyrodniczy w Poznaniu), Regional variation in agricultural production in EU countries (UE-27)

Data analysis and classification 1 - English language session

1. **Reinhold Decker** (Universität Bielefeld), Model-based analysis of online consumer reviews – methods and applications **- invited paper**
2. **Patrick Groenen** (Erasmus University Rotterdam), The support vector machine as a powerful tool for binary classification **- invited paper**
3. **Andrzej Sokołowski** (Cracow University of Economics), **Beata Basiura** (AGH University of Science and Technology in Cracow) – Ward's agglomerative method

Data analysis and classification 2 - English language session

1. **Wojtek Krzanowski** (University of Exeter), Classification: some old principles applied to new problems **- invited paper**
2. **Justyna Wilk** (Uniwersytet Ekonomiczny we Wrocławiu), The symbolic approach in regional analyses
3. **Marcin Pełka** (Wrocław University of Economics), The ensemble approach for clustering interval-valued symbolic data
4. **Justyna Brzezińska** (University of Economics in Katowice), Independence analysis of nominal data with the use of R software

Data analysis and classification 3

1. **Dorota Rozmus** (University of Economics in Katowice), The ensemble approach in taxonomy
2. **Małgorzata Markowska, Bartłomiej Jefmański** (Uniwersytet Ekonomiczny we Wrocławiu), Assessing the dynamics and direction of changes in the development of intelligent specialization of European regions using fuzzy classification
3. **Kamila Migdał-Najman** (Uniwersytet Gdański), The structure of a self-organizing hybrid neural network based on the SOM-GNG algorithm
4. **Krzysztof Najman** (Uniwersytet Gdański), Self-organizing networks in cluster analysis

Data analysis and classification 4

1. **Aleksandra Łuczak, Feliks Wysocki** (UP Poznań), Assessing the level of socio-economic development of NUTS 4 units (*poviats*) of the Wielkopolska province
2. **Katarzyna Kopczewska** (Uniwersytet Warszawski), Regions in space. Does location and accessibility matter for socio-economic development?
3. **Robert Pietrzykowski** (Szkoła Główna Gospodarstwa Wiejskiego w Warszawie), The use of quantile regression in spatial analyses of arable land prices
4. **Marzena Piotrowska-Trybull, Stanisław Sirko** (AON) The influence of a military base on the development of the surrounding NUTS 5 unit (*gmina*)

5. **Joanna Zyprych-Walczak, Alicja Szabelska, Idzi Siatkowski** (Uniwersytet Przyrodniczy w Poznaniu), Methods of gene selection that account for Inter-gene correlation

Statistical data 1

1. **Jan Zawadzki, Maria Szumksta-Zawadzka (**Zachodniopomorski Uniwersytet Technologiczny w Szczecinie), On methods of predicting missing data in time series with periodical (seasonal) fluctuations
2. **Jadwiga Suchecka, Emilia Modranka** (Uniwersytet Łódzki), Spatial statistics – methods and applications
3. **Ewa-Zofia Frątczak, Adam Korczyński** (Szkoła Główna Handlowa w Warszawie), Tradition and modern trends in data imputation methods. An overview of theories – selected examples of applications
4. **Maja Rynko** (GUS), Index and aggregation theory – detailed analysis
5. **Cezary Głowiński** (SAS Institute sp. z o.o.), Using social networks to model the behaviour of customers of telecommunications companies

Statistical data 2

1. **Grażyna Trzpiot, Agnieszka Orwat-Acedańska** (Uniwersytet Ekonomiczny w Katowicach), The classification of mutual funds based on the management style—the quantile regression approach
2. **Tadeusz Kufel** (Uniwersytet Mikołaja Kopernika), **Marcin Błażejowski, Paweł Kufel** (Wyższa Szkoła Bankowa w Toruniu), The creation of databanks as the basis of socio-econometric analyses in GRETL software
3. **Zbigniew Augustyniak**, (Główny Urząd Statystyczny), Online presentation of statistical data as an element of implementing the SISP project
4. **Arleta Olbrot-Brzezińska** (Urząd Statystyczny w Poznaniu), Social functions of statistical information and the responsibility of public statistics
5. **Katarzyna Wawrzyniak** (Katedra Zastosowań Matematyki w Ekonomii Wydział Ekonomiczny Zachodniopomorski Uniwersytet Technologiczny w Szczecinie), Assessing the degree of implementing the main objectives of the National Developmental Strategy by province using correspondence analysis

Poland and Ukraine in the European context

1. **Rusłan Motoryn** (UPUFiHM w Kijowe), **Iryna Motoryna**, **Tetiana Motoryna** (Uniwersytet Kijowski), A comparison of indicators of the use of household savings in Poland and Ukraine
2. **Mieczysław Kowerski** (Wyższa Szkoła Zarządzania i Administracji w Zamościu), Studying economic mood of entrepreneurs and consumers at the regional level based on example of the Lubelskie province
3. **Andrzej Młodak** (Urząd Statystyczny w Poznaniu), Modernization of economic activity statistics on a pan-European scale

4. **Magdalena Gabińska, Piotr Gołos, Alina Warelis** (US w Białymstoku), The value of forest resources in Poland – the theoretical aspect and empirical research
5. **Karol Andrzejczak** (Politechnika Poznańska), Changes and the growth limit of the indicator of the automobile market saturation at the global level and in Poland
6. **Dorota Kwiatkowska-Ciotucha, Urszula Załuska** (UE we Wrocławiu), Assessing the use of ESF funding in different regions of Poland in the current programming cycle

Health statistics

1. **Agnieszka Dyzmann-Sroka, Jerzy Moczko, Andrzej Roszak, Maciej Trojanowski** (Uniwersytet Medyczny im. K. Marcinkowskiego w Poznaniu), Methods of increasing data quality in malignant tumor registers
2. **Urszula Wojciechowska, Joanna Didkowska,** (Centrum Onkologii), Malignant tumors in Poland as a public health problem. Modern tools for measuring the phenomenon
3. **Barbara Kołodziejczak, Magdalena Roszak** (Uniwersytet Medyczny im. K. Marcinkowskiego w Poznaniu), Rapid e-learning in biostatistics
4. **Magdalena Roszak, Barbara Kołodziejczak** (Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu), E-assessment of statistical knowledge of medicine students
5. **Anna Wierzbicka** (Instytut Statystyki i Demografii, Łódź), A taxonomic analysis of public health in Poland in comparison with selected European countries

Statistics of sport and tourism

1. **Jacek Foks** (Ministerstwo Sportu i Turystyki) , Sport statistics in practice **– invited paper**
2. **Barbara Liberda** (Uniwersytet Warszawski), **Łucja Tomaszewicz, Iwona Świerczewska** (Uniwersytet Łódzki), Problems of creating satellite accounts as exemplified by sport satellite account for Poland
3. **Iwona Bąk** (Zachodniopomorski Uniwersytet Technologiczny w Szczecinie), A statistical analysis of tourist activity of senior citizens in Poland
4. **Marcin Błażejowski** (Wyższa Szkoła Bankowa w Toruniu), Applying modelling and prediction algorithms to the tourism market

Classifications and standards

1. **Włodzimierz Okrasa** (Główny Urząd Statystyczny), Statystyka i socjologia: współzależności rozwoju z pespektywy inter-dyscyplinaryzacji badań społecznych. Aspekty metodologiczne i instytucjonalne

2. **Sergiy Gerasymenko** (Uniwersytet Kijowski) **Olena Chupryna**, (Narodowy Uniwersytet Karazina w Chakowie), Optimizing an inventory of indicators for the comparative assessment of socio-economic entities
3. **Grzegorz Krzykowski, Marcin Kalinowski**, Data selection as a key element in statistical surveys in the context of financial markets
4. **Piotr Krasucki (Instytut Spraw Publicznych), Władysław Wiesław Łagodziński**, An integrated system of statistical data on healthcare – the essential minimum
5. **Artur Mikulec, Aleksandra Kupis - Fijałkowska,** (Uniwersytet Łódzki), An empirical analysis of Mojena and Wishart's efficiency criterion in cluster analysis
6. **Mariusz Kraj**, (GUS), A structure and development plan of a statistical education and research centre of the Central Statistical Office in Jachranka

Statistical inference

1. **Mirosław Krzyśko, Łukasz Waszak** (Uniwersytet im. Adama Mickiewicza w Poznaniu) Kernel canonical correlation analysis
2. **Grzegorz Kończak** (UE Katowice), Statistical inference for multivariate data in multiway contingency tables
3. **Joanna Kisielińska** (SGGW Warszawa), Using the bootstrap method to produce precise estimates of mean and variance
4. **Ewa Meller** (US Poznań), Small area estimation for selected variables of the labour market
5. **Łukasz Wawrowski** (US Poznań), Poverty analysis at the NUTS 4 level in the province of Wielkopolska using small area statistics methodology

Economic and social development analysis

1. **Elżbieta Stańczyk** (Urząd Statystyczny we Wrocławiu, Dolnośląski Ośrodek Badań Regionalnych), Occupational qualifications of the unemployed on the basis of research by CSO. Intervoivodship Comparative Analysis
2. **Józef Hozer, Marta Hozer-Koćmiel** (Uniwersytet Szczecinski), *Quantum satis* for companies in Poland in 2012
3. **Janusz Kornecki, Piotr Cmela** (Urząd Statystyczny w Łodzi), Strategic dilemmas of supporting micro business development in Poland
4. **Rafał Nowakowski** (Urząd Statystyczny we Wrocławiu), Local government elections based on the example of capital cities of the Dolnośląski, Małopolski and Wielkopolski regions – 20 years of experience
5. **Jacek Białek** (Uniwersytet Łódzki), A special case of a certain general formula of the price index

Poster session

1. **Beata Bal-Domańska** (Uniwersytet Ekonomiczny we Wrocławiu, WGRiT w Jeleniej Górze) Convergence processes modelling the European regional space
2. **Maciej Beręsewicz** (UE w Poznaniu), Internet data sources in the light of the national census based on the example of the real estate market in Poznan
3. **Anna Błaczkowska** (Wyższa Szkoła Bankowa we Wrocławiu), Alicja Grześkowiak (Uniwersytet Ekonomiczny we Wrocławiu), Demographic and economic conditions of the process of junior high school education in the regions of Dolny Śląsk and Opolszczyzna in the years 2003-2010
4. **Hanna Dudek, Joanna Landmesser** (Katedra Ekonometrii i Statystyki SGGW w Warszawie) Wage satisfaction and relative deprivation
5. **Konrad Furmańczyk, Stanisław Jaworski** () Detecting changes in a series of independent observations
6. **Alicja Ganczarek-Gamrot** (Samodzielny Zakład Demografii i Statystyki Ekonomicznej UE w Katowicach), Multivariate FIGARCH models for risk assessment on the Polish electricity market
7. **Hanna Gruchociak** (UE w Poznaniu), Delimitacja lokalnych rynków pracy w Polsce
8. **Marta Małecka** (Katedra Metod Statystycznych, Uniwersytet Łódzki), Assessing the variance of Value at Risk (VaR) estimators
9. **Magdalena Okupniak** (UE w Poznaniu), Applying intra- and inter-block analysis in economic statistics
10. **Elżbieta Paszko, Anna Maria Jurek** (Katedra Metod Statystycznych, Uniwersytetu Łódzkiego) Application of TQM - Total Quality Management - in the process of improving teaching methods
11. **Agnieszka Pobłocka** (Uniwersytet Gdański), Polski rynek ubezpieczeń w latach 1999-2010
12. **Wojciech Roszka** (UE w Poznaniu), Applying mass imputation methods to integrate databases from various sources
13. **Danuta Rozpędowska-Matraszek** (Instytut Ekonomii Stosowanej PWSZ w Skierniewicach) Assessing the efficiency of healthcare in Poland – an analysis based on groups of similar NUTS 4 units (*poviats*)
14. **Elżbieta Sobczak** (Uniwersytet Ekonomiczny we Wrocławiu) Workforce by the intensity of R&D activity in EU countries – space-structural analysis
15. **Agnieszka Stanimir** (Uniwersytet Ekonomiczny we Wrocławiu), Different techniques of presenting non-metric variables
16. **Marta Styrc** (Instytut Statystyki i Demografii, Szkoła Główna Handlowa), Factors affecting the stability of first marriages in Poland
17. **Marta Styrc** *(Instytut Statystyki i Demografii, Szkoła Główna Handlowa), The educational gradient of divorce in Poland*