STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 193

FROM THE EDITOR

A collection of a dozen papers composing this issue of *Statistics in Transition new series* is conventionally divided into three sections: *Sampling and Estimation Methods, Other Articles,* and *Current Issues in Public Statistics.* The latter is devoted to one of the most timely and challenging for public statistics question of cross-border cooperation and movements (flows of peoples and goods), especially between countries of European Union, like Poland or Slovakia, and their non-EU neighbors, like Ukraine or Belarus. And more generally, to the question of transnational phenomena, starting from an attempt to covering them properly at the regional level.

In addition to the problem of confidentiality and data access, which was of focus of the call for papers in the journal's previous issue (as the main object of the planned *Special Issue* , "Confidentiality and Data Access — Theory and Evidence from a Global Perspective"), this is the second question to also be discussed comprehensively in its future edition. By the way, I am pleased to announce that the deadline for submission of the papers for the confidentiality and data access issue has been shifted to the 30th of January 2009.

Since the intention of the third section *(Current Issues in Public Statistics)* is to go beyond presentation of scientific ideas and discussion of theoretical problems - which constitute the primary goal of the journal - and to address the questions considered special interest to public statistics, I would like to signal in advance the issues related to the census (conduct of which is under preparation in almost all countries) as such an object of interest to the journal from now on. It means that we would give priority to publishing the census-related papers, even before officially announcing a call for papers on census-related problems.

Let me take this opportunity to express my gratitude to peer reviewers for their collaboration and generous contribution to the quality of our journal.

WŁODZIMIERZ OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 1, pp. 195

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/gus/45_2638_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 197—206

AN APPLICATION OF SHRINKAGE TO DEALING WITH EXTREME VALUES

Nicholas T. Longford[1]

ABSTRACT

Trimming and winsorisation are commonly used devices for reducing the influence of exceptionally large observations in positively skewed data. We explore a flexible version winsorisation based on the general idea of shrinkage. It can be interpreted as a compromise between retaining and truncating an outlying observation. We show that the details of winsorisation should be informed not only by the distribution of the data but also by the target of estimation.

Key phrases: log-normal distribution, Pareto distribution, shrinkage, trimming, winsorisation.

1. Introduction

Trimming and winsorisation are two methods for dealing with observations which on the count of their very large values exercise undue influence on statistics (estimates), such as the sample mean and sample variance (García-Escudero *et al.*, 2003; and Balog and Thorburn, 2007). They are commonly applied in studies in which population summaries (expectation, standard deviation, and the like) are sought for a positively skewed variable, such as income, expenditure, house price, company's levels of assets and liabilities, number of events of a particular kind, and the like. Transformations to approximate normality and application of generalised linear models with suitable link functions often make the analysis of such data tractable. These approaches are not useful when the population mean or total are estimated, because the operations of a nonlinear transformation and averaging cannot be interchanged and the results obtained on the transformed scale are difficult to translate to the original scale.

Trimming is defined as discarding a given percentage $100p$ of the largest observations in the sample. As an alternative, observations that exceed an *a priori*

[1] Address for correspondence: N. T. Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain. Email: NTL@SNTL.co.uk.

specified threshold value v may be discarded. Winsorisation is defined as truncating the top $100p$% of the observations to their sample $100p$-percentile, or as truncating to a set threshold value v all the observations that exceed it. Trimming is easy to justify when there are good grounds for believing that the affected observations are in error (erroneous values or erroneous inclusion of the units in the sample), whereas winsorisation is applied when the scale used for the outcomes may be inappropriate, or the affected observation exerts a disproportional influence on the value of the statistic or estimator.

The next section presents our proposal for a smoother version of winsorisation. The following section evaluates its properties and explores how its parameters should be set by simulations. Throughout, we consider only variables with positive values.

2. Shrinkage winsorisation

Instead of replacing each value $x_k > v$ by v in a sample $(x_1, ..., x_n)$, we propose to shrink its value toward v by replacing x_k with

$$\text{xk'} = (1 - \text{b})\ \text{xk} + \text{bv}, \qquad (1)$$

where $b \in [0,1]$ is a constant. We refer to this transformation as *shrinkage winsorisation*. The established form of winsorisation corresponds to $b = 1$ and operating with the original sample to $b = 0$. The log-transformation is frequently applied to positively skewed variables to make the assumptions of normality or symmetry more palatable or merely to reduce the influence of the largest values. Shrinkage winsorisation can be applied on the log-scale, yielding the adjustment

$$x_k' = \exp\{(1 - b)\log(x_k) + b\log(v)\}, \qquad (2)$$

for $b \in [0,1]$. In principle, any function increasing in $(0,+\infty)$ and its inverse can be used in place of log and exp in this transformation, although log and exp with $b \in (0,1)$, in addition to identity-identity in (1), probably offer sufficient flexibility for most situations. To distinguish these two ways of shrinkage winsorisation, we refer to (1) as *linear* and to (2) as *power* shrinkage winsorisation. Further, we refer to the originally defined winsorisation (with $b = 1$) as *full*.

Shrinkage winsorisation is specified by the type (linear or power), the threshold v and the shrinkage coefficient or *extent of shrinkage b*. The threshold v may be set *a priori* or be data-dependent, such as a particular sample percentile, so that a given fraction of the observations is affected. There is little clinical advice in the literature about how to set v in full winsorisation, and the problem of setting the extent b and the threshold v for shrinkage winsorisation is even more difficult. The additional flexibility need not yield success unless b and v are set with care, drawing on (prior) information about the distribution of the data.

The next section explores the properties of linear and power shrinkage winsorisation by simulations. We intend it as an outline for how *b* and *v* might be set in a particular study; a simple prescription for them is beyond what can be responsibly provided. The simulations indicate that setting the values of *b* and *v* should be informed not only by the distribution of the data but also by the target of estimation; winsorisation may be detrimental for estimating the population mean and very effective for estimating the population variance or *vice versa*.

3. Simulations

We consider two classes of distributions for generating positively skewed outcomes: log-normal and Pareto. Their variety can be enhanced by mixing. For example, a mixture of two log-normal distributions is constructed by generating two subsamples (not necessarily of equal sizes) from distinct log-normal distributions. In the analysis, the identity of the generating distribution for an observation is regarded as not known. It may be argued that one of the mixture components is bound to dominate the right-hand tail of the mixture distribution, and so only this component is important for setting the details of (shrinkage) winsorisation.

We study first the performance of shrinkage winsorisation for log-normally distributed samples. Such samples are generated as exp(x_k), where x_k , k =1, ..., n, are a random sample from a normal distribution, $\mathrm{N}(\mu,\sigma^2)$. No generality is lost by setting $\mu = 0$; otherwise exp(μ) is a multiplicative factor in $\{ x_k \}$.

We evaluate the bias and root mean squared error (rMSE) of the estimators of expectation, standard deviation and variance. Apart from the two types of shrinkage winsorisation, with the full range of the extent-coefficient *b*, we also consider the naive estimators of the moments

$$\mathrm{E}\{\exp(X)\} = \exp\left(\mu + ½\sigma^2\right)$$

$$\mathrm{var}\{\exp(\mathrm{X})\} = \exp(2\mu + \sigma 2\ )\ \{\exp(\sigma 2\ ) - 1\}, \qquad (3)$$

which are obtained by replacing μ and σ^2 with their standard estimators. Each simulation comprises 5000 replications for each of the values $b = h/40$, $h = 0, 1, ..., 40$.

Figure 1 summarises the results of a set of simulations for sample size $n = 200$ from the log-normal distribution with the underlying distribution N(0, 0.25) and threshold set at the 95th percentile of the distribution, equal to 2.276. The left-hand panels confirm that the estimators with shrinkage winsorisation have negative biases and they decrease with *b*. In contrast, the rMSEs in the right-hand panels decrease with *b*, reach a minimum, and then they increase.

For estimating the expectation, full winsorisation is nearly efficient; the optimal shrinkage is attained for *b*=0.7 for linear shrinkage and 0.65 for power shrinkage, but the minima differ only slightly. For estimating the standard deviation, the optima are attained for *b*=0.325 and *b*=0.375 for power and linear shrinkage, respectively. The minimum for power shrinkage is slightly lower, but neither attains the efficiency of the naive estimator based on (3). Full winsorisation (*b*=1) is counterproductive for estimating the standard deviation. In contrast, full and no winsorisation are about equally efficient for estimating the variance, but shrinkage winsorisation leads to more efficient estimation than either of these alternatives. Optimum is attained for *b*=0.4 for power shrinkage, and for slightly higher *b* for linear shrinkage. At their optima, power shrinkage is slightly more efficient than linear shrinkage. With the optimal coefficients *b*, shrinkage winsorisation comes close to matching the efficiency of the naive estimator.

Figure 1. The biases and rMSEs of estimators of the expectation, standard deviation and variance of the log-normal distribution. Data simulated from log-N(0, 0.25). Sample size *n* =200.

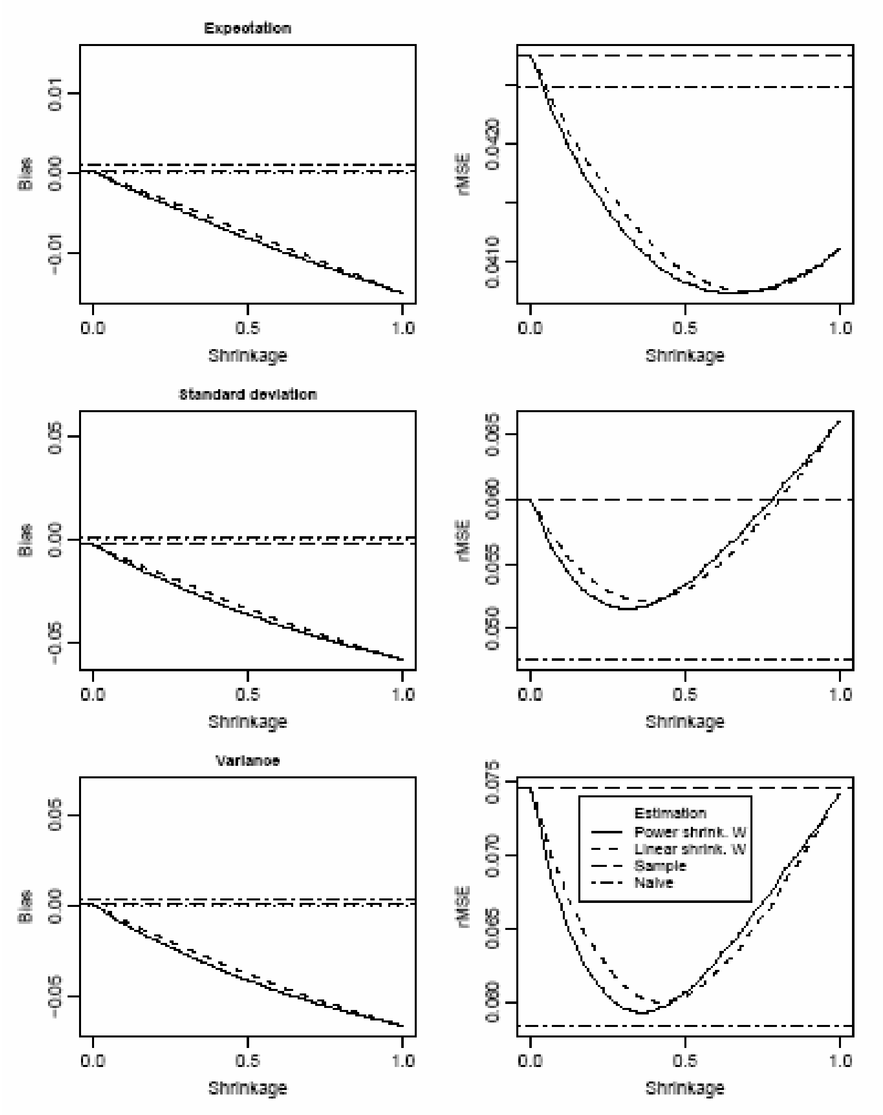


Figure 1 illustrates that shrinkage winsorisation is more efficient than full winsorisation, but also that no winsorisation with a single (fixed) coefficient *b* is efficient for estimating all population summaries. Greater efficiency may be attained by selecting *b* specifically for the target, even if it runs counter to the convention and convenience of working with a single (adjusted) dataset. Power shrinkage has a slightly greater potential than linear shrinkage.

Although the naive estimator is efficient for estimating the standard deviation and variance, it is less robust than an estimator based on shrinkage winsorisation. We illustrate this by applying shrinkage winsorisation to a mixture of two log-normal distributions: log-N(−0.5, 0.25) with probability 0.7 and log-N(1, 1) with probability 0.3. The density of this distribution is plotted in Figure 2 together with the density of log-N(0, 0.25), used in the previous simulation. The mixture distribution has much thicker right-hand tail. It could not be matched closely by a single log-normal distribution.

Figure 2. The densities of the log-normal distribution log-N(0, 0.25) and of the 0.7/0.3 mixture of log-N(−0.5, 0.25) with probability 0.7 and log-N(1, 1) with probability 0.3.

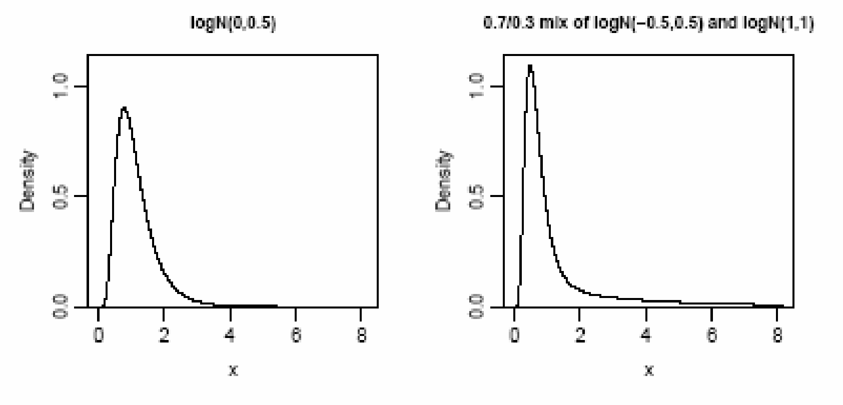
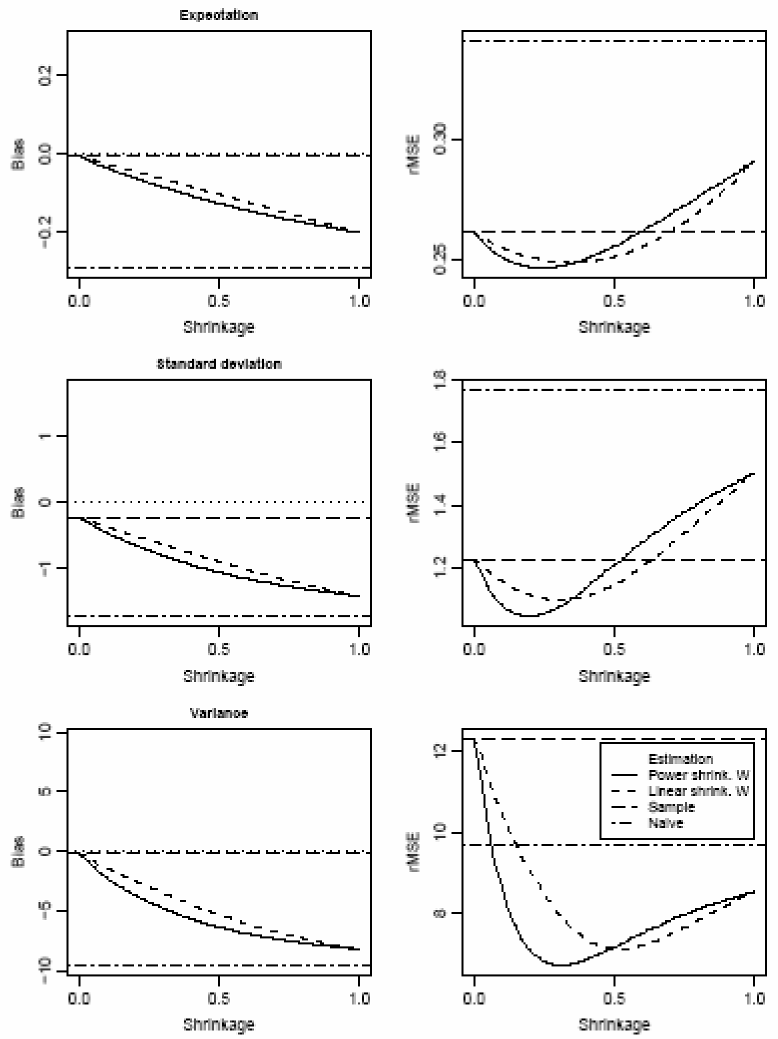


Figure 3 shows that the naive estimators of the mean, standard deviation and variance have substantial negative biases. The shrinkage winsorisation with coefficient around 0.3 is more efficient than both the sample and naive alternatives for all three targets. The naive estimator is very inefficient for the mean and standard deviation, but not for the variance. To see this, we compare the expectations and standard deviations of the mixture on the underlying scale and after exponentiation. On the underlying scale, the expectation is −0.05 and the standard deviation 0.973. The normal distribution with these moments has, after exponentiation, expectation 1.527 and standard deviation 1.918. However, the mixture of two log-normals has expectation 1.826 and standard deviation 3.670.

Figure 3. The biases and rMSEs of estimators of the expectation, standard deviation and variance of a mixture of two log-normal distributions. Data simulated from the 0.7/0.3 mixture of log-N(−0.5, 0.25) and log-N(1, 1). Sample size *n* =200.



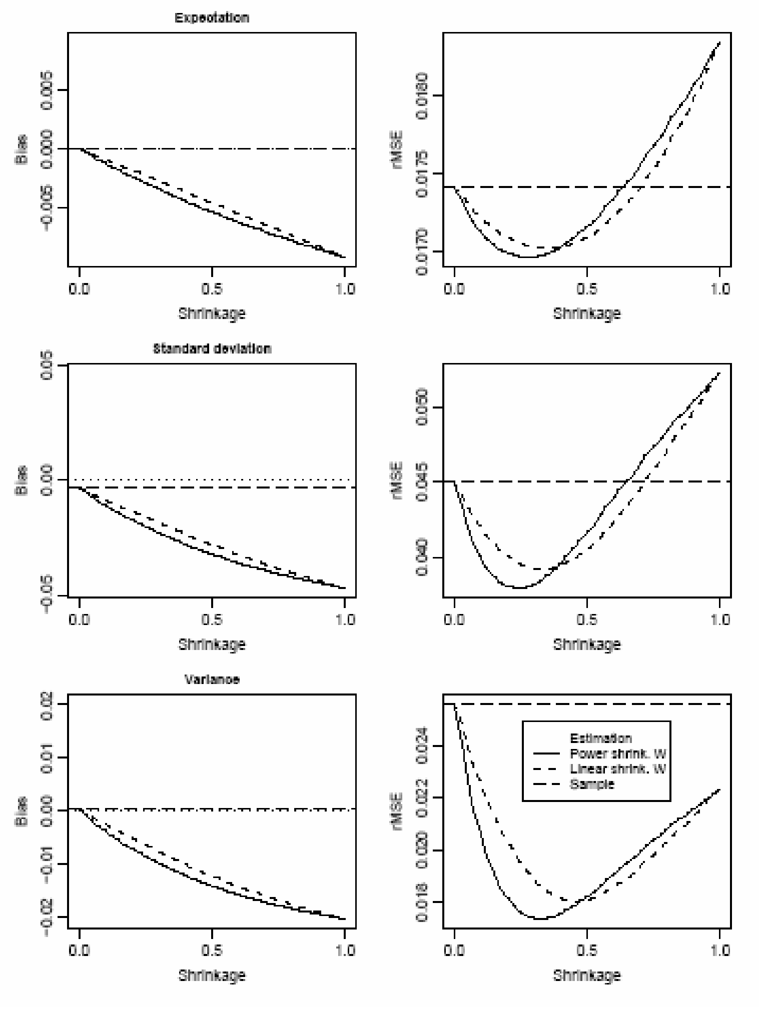
We explore next the performance of shrinkage winsorisation with Pareto distributed data. The class of Pareto distributions is defined by the densities

$$f(x;\, \theta, x_0) \;=\; \theta x^{\theta} \left(x+x_0 \right)^{-(\theta+1)}$$

for $x \in (0,+\infty)$, where the *exponent* $\theta > 0$ and origin $x_0 > 0$ are parameters. The expectation of a Pareto distribution, denoted by $\mathrm{P}(\theta, x_0)$, is $x_0/(\theta-1)$, so long as $\theta >$

1, and the variance is $x_0\theta/\{(\theta-2)(\theta-1)^2\}$, so long as $\theta>2$. Figure 4 summarises the simulations from P(6,1). It confirms that (shrinkage) winsorisation results in a bias in estimating all three moments. For the mean and the standard deviation, estimation after full winsorisation is less efficient than with no winsorisation; the optimal shrinkage coefficient for winsorisation is around 0.25. For estimating the variance, full winsorisation is preferable to no winsorisation, but the optimal winsorisation is with shrinkage coefficient around 0.3. For all three quantities, multiplicative shrinkage winsorisation is better than linear shrinkage.

Figure 4. The biases and rMSEs of estimators of the expectation, standard deviation and variance of the Pareto distribution. Data simulated from P(6, 1). Sample size *n* =200.



The naive estimator used with log-normally distributed data is more efficient than the sample estimator; see Figure 1. When applied to Pareto-distributed data it is (relatively) very inefficient.

4. Discussion

We have generalised winsorisation from a discrete choice, whether to apply it or not, to a continuum of alternatives defined by the extent of shrinkage *b* and applied on the linear or log-scale. The simulation study confirms that this generalisation is useful and shows that the decision whether to apply winsorisation in the discrete-choice setting, and how to set the shrinkage coefficient *b*, together with the setting of the threshold *v*, should in ideal circumstances be informed not only by the distribution of the data but also by the target of estimation; different coefficients *b* are optimal for the expectation, standard deviation, and variance. The actual choice of band *v* is an outstanding problem that defies any straight-forward solution. We have conducted simulations similar to those reported in Section 3 with a wide range of distributions but have failed to discover any easy-to-describe associations that would enable us to generate a simple rule for the choice of the constants *b* and *v*.

The natural counterpart of shrinkage winsorisation for trimming, called *partial trimming*, is defined by reduced weights for the extreme observations. Thus, for a constant $w \in [0,1]$, each observation y that exceeds the threshold v is associated with weight $1-w$. For example, the partially-trimmed estimator of the population mean is

$$\left[\sum_i y_i\{1 - wI(y_i > v)\}\right] \Big/ \left[n - w \sum_i I(y_i > v)\right] , \tag{4}$$

where n is the sample size and I the indicator function; it is equal to unity when its argument is true and to zero otherwise. No trimming corresponds to $w = 0$ and 'full' trimming to $w = 1$. When the observations are associated with sampling weights, the estimator in (4) has an obvious adaptation — the sampling weights are multiplied by the factor $1 - wI(y_i > v)$.

Shrinkage winsorisation is an application of the general idea of shrinkage or *composite estimation*; see Longford (2007). Instead of choosing one of the contending estimators, we consider their convex combinations and choose or estimate the coefficient that yields maximum efficiency. Among other factors, the optimal coefficient depends on the target of estimation. That implies that the commonly adopted strategy of adjusting a dataset by winsorisation and using it for estimating several quantities is suboptimal. Our simulations show that shrinkage winsorisation has a potential, although its full exploitation requires further research.

The simulations were conducted using R (R Development Core Team, 2006) and the code developed (R functions) is available from the author on request.

Acknowledgement

Research for this article was partly supported by the Grant SEJ2006–13537 from the Spanish Ministry of Science and Technology.

REFERENCES

BALOG, M., and THORBURN, D. (2007). Extreme value treatment for samples from skew income distributions. *Statistics in Transition* **7,** 139–153.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., and MATRÁN, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics* 12, 434–449.

LONGFORD, N. T. (2007). *Studying Human Populations. An Advanced Course in Statistics*. Springer-Verlag, New York.

R Development Core Team. (2006). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 207—214

GENERAL CLASS OF ESTIMATORS FOR ESTIMATING THE POPULATION MEAN INTWO-STAGE CLUSTER SAMPLING

Mohanad Al-khasawneh[1]

ABSTRACT

This paper presents a general class of estimators for estimating the finite population mean in two-stage cluster sampling with unequal first-stage units. The mean square error (MSE) and minimum mean square erroe of this class of estimators have been derived.

1. Introduction

Let the set of first-stage units (fsu) of a finite population be denoted by U=(1,2… i, … N) such that the i-th fsu contains M_i second-stage units (ssu) and $M=\sum_{i=1}^{N} M_i$. Let $\overline{Y}_i$, $\overline{X}_i$ and $\overline{Z}_i$ be the means in U_i in respect of the study variable *Y* and the two auxiliary variables *X* and *Z* respectively.

Define,

$$\overline{Y}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i}=\frac{\sum_{i=1}^{N} M_i \overline{Y}_i}{M}=\sum_{i=1}^{N} w_i \overline{Y}_i \,, \overline{X}=\sum_{i=1}^{N} w_i \overline{X}_i \qquad (1.1)$$

and $\overline{Z}=\sum_{i=1}^{N} w_i \overline{Z}_i$, $w_i=\dfrac{M_i}{M}$

where $\overline{Y}_i, \overline{X}_i$ and $\overline{Z}_i$ are the population means per element of the i-th cluster.

[1] Department of Statistics, Yarmouk University Irbid, Jordan.

Let us consider a general class of two-stage sampling to estimate $\bar{Y}$. At stage one, a sample s ($s \subset U$) of n fsu's is drawn from U according to any design with equal probabilities. Then for every $i \in s$, a sample S_i of m_i ssu's is drawn from U_i ($s_i \subset U_i$) with suitable selection probabilities at the second-stage.

Let E_1, V_1, and COV_1 denote the expectation, variance, and covariance over repeated sampling in the first stage. E_2, V_2, and COV_2 denote the expectation, variance, and covariance over repeated sampling in the second-stage. E, V, and COV we denote overall expectation, variance and covariance.

It is assumed that from the second-stage sample s_i, $i \in s$, unbiased estimates $\hat{\bar{Y}}_i$, $\hat{\bar{X}}_i$ and $\hat{\bar{Z}}_i$ respectively for $\bar{Y}_i$, $\bar{X}_i$ and $\bar{Z}_i$ are available such that

$$\hat{\bar{Y}}_i = \sum_{j \in s_i} Y_j / m_i \text{ , } \hat{\bar{X}}_i = \sum_{j \in s_i} X_j / m_i \text{ , } \hat{\bar{Z}}_i = \sum_{j \in s_i} Z_j / m_i$$

$$V_2(\hat{\bar{Y}}_i) = \sigma^2_{i\bar{y}} \text{ , } V_2(\hat{\bar{X}}_i) = \sigma^2_{i\bar{x}} \text{ , } V_2(\hat{\bar{Z}}_i) = \sigma^2_{i\bar{z}} \text{ , } COV_2(\hat{\bar{Y}}_i, \hat{\bar{X}}_i) = \sigma_{i\overline{yx}} \text{ , }$$

$$COV_2(\hat{\bar{Y}}_i, \hat{\bar{Z}}_i) = \sigma_{i\overline{yz}} \text{ and } COV_2(\hat{\bar{Z}}_i, \hat{\bar{X}}_i) = \sigma_{i\overline{xz}}$$

Where for Simple Random Sampling Without Replacement (SRSWOR),

$$\sigma^2_{\bar{y}} = \frac{1-f}{n} S^2_y \text{ , } \sigma^2_{\bar{x}} = \frac{1-f}{n} S^2_x \text{ , } \sigma^2_{\bar{z}} = \frac{1-f}{n} S^2_z$$

$$\sigma_{\overline{yx}} = \frac{1-f}{n} S_{yx} \text{ , } \sigma_{\overline{yz}} = \frac{1-f}{n} S_{yz} \text{ , } \sigma_{\overline{xz}} = \frac{1-f}{n} S_{xz}$$

$$\sigma^2_{i\bar{y}} = \frac{1-f_i}{m_i} S^2_{iy} \text{ , } \sigma^2_{i\bar{x}} = \frac{1-f_i}{m_i} S^2_{ix} \text{ , } \sigma^2_{i\bar{z}} = \frac{1-f_i}{m_i} S^2_{iz}$$

$$\sigma_{i\overline{yx}} = \frac{1-f_i}{m_i} S_{iyx} \text{ , } \sigma_{i\overline{yz}} = \frac{1-f_i}{m_i} S_{iyz} \text{ and } \sigma_{i\overline{xz}} = \frac{1-f_i}{m_i} S_{ixz}$$

Where $f = \dfrac{n}{N}$ and $f_i = \dfrac{m_i}{M_i}$ are the correction factors in the first and the second stage respectively.

Given the first-stage sample *s*, and the second-stage samples s_i (i=1,2,…,*n*)we define the estimators:

$$\hat{\bar{Y}} = \sum_{i=1}^{n} w_i \hat{\bar{Y}}_i \, , \quad \hat{\bar{X}} = \sum_{i=1}^{n} w_i \hat{\bar{X}}_i \text{ and } \hat{\bar{Z}} = \sum_{i=1}^{n} w_i \hat{\bar{Z}}_i$$

then $E(\hat{\bar{Y}}) = \bar{Y}$, $E(\hat{\bar{X}}) = \bar{X}$ and $E(\hat{\bar{Z}}) = \bar{Z}$,

Define,

$$S_y^2 = \frac{\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2}{N-1} \, , \quad S_x^2 = \frac{\sum_{i=1}^{N} (\bar{x}_i - \bar{X})^2}{N-1}$$

$$S_z^2 = \frac{\sum_{i=1}^{N} (\bar{z}_i - \bar{Z})^2}{N-1} \, , \quad S_{yx} = \frac{\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})(\bar{x}_i - \bar{X})}{N-1}$$

$$S_{yz} = \frac{\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})(\bar{z}_i - \bar{Z})}{N-1} \, , \quad S_{xz} = \frac{\sum_{i=1}^{N} (\bar{z}_i - \bar{Z})(\bar{x}_i - \bar{X})}{N-1}$$

$$S_{iy}^2 = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{M_i - 1} \, , \quad S_{ix}^2 = \frac{\sum_{j=1}^{M_i} (x_{ij} - \bar{X}_i)^2}{M_i - 1}$$

$$S_{iz}^2 = \frac{\sum_{j=1}^{M_i} (z_{ij} - \bar{Z}_i)^2}{M_i - 1} \, , \quad S_{iyx} = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)(x_{ij} - \bar{X}_i)}{M_i - 1}$$

$$S_{iyz} = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)(z_{ij} - \bar{Z}_i)}{M_i - 1} \text{ and } S_{ixz} = \frac{\sum_{j=1}^{M_i} (z_{ij} - \bar{Z}_i)(x_{ij} - \bar{X}_i)}{M_i - 1}$$

Then

$$V(\hat{\bar{Y}}) = \sigma_{\bar{y}}^2 + \frac{1}{nN} \sum_{i=1}^{N} w_i^2 \sigma_{\bar{iy}}^2 \, , \quad V(\hat{\bar{X}}) = \sigma_{\bar{x}}^2 + \frac{1}{nN} \sum_{i=1}^{N} w_i^2 \sigma_{\bar{ix}}^2 \, ,$$

$$V(\hat{\bar{Z}}) = \sigma_{\bar{z}}^2 + \frac{1}{nN} \sum_{i=1}^{N} w_i^2 \sigma_{\bar{iz}}^2 \, , \quad COV(\hat{\bar{Y}}, \hat{\bar{X}}) = \sigma_{\bar{yx}} + \frac{1}{nN} \sum_{i=1}^{N} \sigma_{\bar{iyx}}$$

$$COV(\hat{\bar{\bar{Y}}},\hat{\bar{\bar{Z}}}) = \sigma_{\overline{yz}} + \frac{1}{nN}\sum_{i=1}^{N}\sigma_{i\overline{yz}} \text{ and } COV(\hat{\bar{\bar{Z}}},\hat{\bar{\bar{X}}}) = \sigma_{\overline{zx}} + \frac{1}{nN}\sum_{i=1}^{N}\sigma_{i\overline{zx}}$$

2. The Class of Estimators:

For given s_i motivated by Srivastava (1980) let us define a class of estimators for $\bar{Y}_i$ by:

$$\hat{\bar{\bar{Y}}}_{hi} = h_i(\hat{\bar{Y}}_i, \hat{\bar{X}}_i), i \in s \qquad (2.2)$$

Where $h_i(\hat{\bar{Y}}_i, \hat{\bar{X}}_i)$ is a known function of $\hat{\bar{X}}_i$ and $\hat{\bar{Y}}_i$ which may contain $\bar{X}_i$ but independent of $\bar{Y}_i$ such that $h_i(\hat{\bar{Y}}_i, \hat{\bar{X}}_i) = \bar{Y}_i$. For given s let $\hat{\bar{\bar{Y}}}_s = \sum_{i\in s} w_i \hat{\bar{Y}}_i$ and $h(\hat{\bar{\bar{Y}}}_s, \hat{\bar{\bar{Z}}})$ be a function of $\hat{\bar{\bar{Y}}}_s$ and $\hat{\bar{\bar{Z}}}$ which may contain $\bar{\bar{Z}}$ but independent of $\bar{\bar{Y}}$ such that $h(\hat{\bar{\bar{Y}}}_s, \hat{\bar{\bar{Z}}}) = \bar{\bar{Y}}$.

Further, following Srivastava (1980), let us assume that:

i) i) $(\hat{\bar{Y}}_i, \hat{\bar{X}}_i), i \in s$ and $(\hat{\bar{\bar{Y}}}_s, \hat{\bar{\bar{Z}}})$ take values in a bounded, closed convex subspace (R^3) of three-dimensional real space containing the points $(\bar{Y}_i, \bar{X}_i)$ i.e. $(\bar{Y}_i, \bar{X}_i, 0)$ and $(\bar{\bar{Y}}, \bar{\bar{Z}})$ i.e. $(\bar{\bar{Y}}, 0, \bar{\bar{Z}})$.

ii) ii) The functions h_i and h are continuous having first and second order partial derivatives which are also continuous in R^3.

Thus the proposed class of estimators of $\bar{\bar{Y}}$ may be defined by:

$$\hat{\bar{\bar{Y}}}_g = h(\hat{\bar{\bar{Y}}}_s, \hat{\bar{\bar{Z}}}) \qquad (2.3)$$

On expanding $\hat{\bar{\bar{Y}}}_{hi} = h_i(\hat{\bar{Y}}_i, \hat{\bar{X}}_i)$ around the point $(\bar{Y}_i, \bar{X}_i)$ by first order Taylor's series and neglecting the reminder term. It can be observed that to a first order of approximation $E_2(\hat{\bar{\bar{Y}}}_i) = \bar{Y}_i$

$$V_2(\hat{\bar{Y}}_i) = E[h(\bar{y}_i,\bar{x}_i) - h(\mu_{\bar{y}_i},\mu_{\bar{x}_i})]^2$$

$$= E(\bar{y}_i - \mu_{\bar{y}_i})^2[\frac{\partial h(\bar{y}_i,\bar{x}_i)}{\partial \bar{y}_i}\Big|_{\bar{y}_i=\mu_{\bar{y}_i}}]^2$$

$$+ E(\bar{x}_i - \mu_{\bar{x}_i})^2[\frac{\partial h(\bar{y}_i,\bar{x}_i)}{\partial \bar{x}_i}\Big|_{\bar{x}_i=\mu_{\bar{x}_i}}]^2$$

$$+ 2E(\bar{y}_i - \mu_{\bar{y}_i})\ (\bar{x}_i - \mu_{\bar{x}_i})\ \frac{\partial h(\bar{y}_i,\bar{x}_i)}{\partial \bar{y}_i}\Big|_{\bar{y}_i=\mu_{\bar{y}_i}}\ \frac{\partial h(\bar{y}_i,\bar{x}_i)}{\partial \bar{x}_i}\Big|_{\bar{x}_i=\mu_{\bar{x}_i}}$$

$$V_2(\hat{\bar{Y}}_i) = \sigma^2_{i\bar{y}} + \sigma^2_{i\bar{x}}h^2_{2i}(\bar{y}_i,\bar{x}_i) + 2h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{y}\bar{x}} \qquad (2.4)$$

where $h_{2i}(\bar{y}_i,\bar{x}_i)$ is the value of the first order partial derivative of h_i w.r.t. $\bar{x}_i$ at $\bar{x}_i = \mu_{\bar{x}_i}$.

Similarly an expansion of $h(\hat{\bar{Y}}_s,\hat{\bar{Z}})$ about the point $(\bar{y},\bar{z})$ in a first order Taylor's series yields an asymptotic variance of $\bar{Y}_g$ as:

$$V(\hat{\bar{Y}}_g) = V(\hat{\bar{Y}}_s) + h_2^2(\bar{y}\ ,\bar{z})V(\hat{\bar{z}}) + 2h_2(\bar{y}\ ,\bar{z})COV\ (\hat{\bar{y}}_s,\bar{z}) \qquad (2.5)$$

where $h_2(\bar{Y},\bar{Z})$ is the value of the first order partial derivative of h$(\bar{y},\bar{z})$.

We know that,

$$V(\hat{\bar{Y}}_s) = V_1E_2(\hat{\bar{Y}}_s) + E_1V_2(\hat{\bar{Y}}_s)$$

and

$$COV(\hat{\bar{Y}}_s,\hat{\bar{Z}}) = COV_1[E_2(\hat{\bar{Y}}_s),E_2(\hat{\bar{Z}})] + E_1COV_2(\hat{\bar{Y}}_s,\hat{\bar{Z}})$$

After simplification

$$V(\hat{\bar{Y}}_s) = \sigma^2_{\bar{y}} + \sum_{i=1}^{N}[\sigma^2_{i\bar{y}} + 2h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{y}\bar{x}} + h^2_{2i}(\bar{y}_i,\bar{x}_i)\sigma^2_{i\bar{x}}] \qquad (2.6)$$

and

$$COV(\hat{\bar{Y}}_s,\hat{\bar{Z}})=\sigma_{\bar{y}\bar{z}}+\sum_{i=1}^{N}[\sigma_{i\bar{y}\bar{z}}+h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{x}\bar{z}}] \tag{2.7}$$

Finally, from (2.6) and (2.7)

$$V(\hat{\bar{Y}}_g)=\sigma_{\bar{y}}^2+2h_2(\bar{y},\bar{z})\sigma_{\bar{y}\bar{z}}+h_2^2(\bar{y},\bar{z})\sigma_{\bar{z}}^2+2h_2(\bar{y},\bar{z})[$$

$$\sum_{i=1}^{N}(\sigma_{i\bar{y}\bar{z}}+h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{x}\bar{z}})]+h_2^2(\bar{y},\bar{z})\sum_{i=1}^{N}\sigma_{i\bar{z}}^2$$

$$+\sum_{i=1}^{N}[\sigma_{i\bar{y}}^2+2h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{y}\bar{x}}+h_{2i}^2(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{x}}^2] \tag{2.8}$$

The optimum choices of *h*(.), which minimize the variance, be obtained from the following equations

$$\frac{\partial V(\hat{\bar{Y}}_g)}{\partial h_2}=2\sigma_{\bar{y}\bar{z}}+2h_2(\bar{y},\bar{z})\sigma_{\bar{z}}^2+2\sum_{i=1}^{N}[\sigma_{i\bar{y}\bar{z}}+h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{y}\bar{x}}]$$

$$+2h_2(\bar{y},\bar{z})\sum_{i=1}^{N}\sigma_{i\bar{z}}^2=0$$

$$\frac{\partial V(\hat{\bar{Y}}_g)}{\partial h_{2i}}=2N\sigma_{i\bar{x}\bar{z}}h_2(\bar{y},\bar{z})+2N\sigma_{i\bar{y}\bar{x}}+2Nh_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{x}}^2=0$$

$$h_2(\bar{y},\bar{z})=-\frac{\sigma_{\bar{y}\bar{z}}+\sum_{i=1}^{N}(\sigma_{i\bar{y}\bar{z}}+h_{2i}(\bar{y}_i,\bar{x}_i)\sigma_{i\bar{x}\bar{z}})}{\sigma_{\bar{z}}^2+\sum_{i=1}^{N}\sigma_{i\bar{z}}^2} \tag{2.9}$$

and

$$h_{2i}(\bar{y}_i,\bar{x}_i)=-\frac{\sigma_{i\bar{z}\bar{x}}h_2(\bar{y},\bar{z})+\sigma_{i\bar{y}\bar{x}}}{\sigma_{i\bar{x}}^2} \tag{2.10}$$

The results in (2.9) and (2.10) can be written as follows:

$$h_2(\bar{y},\bar{z}) = -\frac{\sigma_{\bar{y}\bar{z}} + \sum_{i=1}^{N} \sigma_{i\bar{z}}^2(\beta_{i\bar{y}\bar{z}} - \beta_{i\bar{y}\bar{x}}\beta_{i\bar{z}\bar{x}})}{\sigma_{\bar{z}}^2 + \sum_{i=1}^{N} \sigma_{i\bar{z}}^2(1-\rho_{i\bar{z}\bar{x}}^2)} = \hat{h}_2 \text{ (Say)} \qquad (2.11)$$

and

$$h_{2i}(\bar{y}_i,\bar{x}_i) = -(\beta_{i\bar{z}\bar{x}}\hat{h}_2(\bar{y},\bar{z}) + \beta_{i\bar{y}\bar{x}}) = \hat{h}_{2i} \text{ (Say)} \qquad (2.12)$$

where

$$\beta_{i\bar{y}\bar{x}} = \sigma_{i\bar{y}\bar{x}} / \sigma_{i\bar{x}}^2 \text{ , } \beta_{i\bar{z}\bar{x}} = \sigma_{i\bar{z}\bar{x}} / \sigma_{i\bar{x}}^2 \text{ ,}$$

$$\beta_{i\bar{y}\bar{z}} = \sigma_{i\bar{y}\bar{z}} / \sigma_{i\bar{z}}^2 \text{ and } \rho_{i\bar{z}\bar{x}} = \sigma_{i\bar{z}\bar{x}} / \sigma_{i\bar{x}}\sigma_{i\bar{z}}$$

Thus, the minimum variance can be determined after estimating the value of $\hat{h}_2$ which can be used to compute $\hat{h}_{2i}$. Thus, approximate variance (mean square error) is

$$V(\hat{\bar{Y}}_g)_{min} = \sigma_{\bar{y}}^2(1-B^2) + \sum_{i=1}^{N} \sigma_{i\bar{y}}^2(1-\rho_{i\bar{y}\bar{x}}^2) \qquad (2.13)$$

where

$$B^2 = -\hat{h}_2^2[\sigma_{\bar{z}}^2 + \sum_{i=1}^{N} \sigma_{i\bar{z}}^2(1-\rho_{i\bar{z}\bar{x}}^2)] / \sigma_{\bar{y}}^2$$

The estimator attaining this bound (may be called the minimum variance bound (MVB) estimator) is a regression-type estimator of the form:

$$\hat{\bar{Y}}_{RG} = \sum_{i \in s}(\hat{\bar{Y}}_i - \hat{h}_{2i}(\hat{\bar{x}}_i - \bar{X}_i)) - \hat{h}_2(\hat{\bar{z}} - \bar{Z}) \qquad (2.14)$$

REFERENCES

COCHRAN, W.C., *Sampling Techniques*, third edition. John Wiley & Sons, 1977.

HOSSAIN, M.I. & AHMED, M.S., A class of predictive estimators in two-stage sampling. *Jour. of Information and Management Sciences. Vol.*12, No.1,P. 49—55, 2001.

ROBINSON, P.M., A class of estimators for the mean of finite population using auxiliary information. *Sankhy. Vol.56, series B*, P.389—399, 1994.

SCOTT A. and SMITH, T. M. P., Estimation in multistage surveys, *Jour. Amer. Stat. Assoc., Vol.* 64, P.830—840, 1969.

SINGH, D. and F. S. CHAUDHARY*, sample survey designs*, 1997.

SRIVASTAVA, S.K., A generalized estimator for the mean of finite population using multi-auxiliary information. *Jour. Amer. Stat. Assoc. Vol.*66, P.404—407, 1971.

SRIVASTAVA, S.K. & JHAJJ, H.S., A class of estimators using auxiliary information for estimating finite population variance. *Sankhy. Vol.*42, P.87—96, 1981.

TRIPATHI, T.P., A class of estimators for population mean using multivariate auxiliary information under general sampling designs. *Alig. Jour. Stat. Vol.*7, P.49—62, 1987.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 215—232

ON THE USE OF GUESS VALUE FOR THE ESTIMATION OF POPULATION MEDIAN ON CURRENT OCCASION IN TWO OCCASIONS ROTATION PATTERNS

G. N. Singh[1], Kumari Priyanka[2]

ABSTRACT

The present work is an attempt to present a sample rotation pattern, in which individuals are sampled for two successive occasions. Here the problem of estimation of a finite population median at current occasion, in two occasions successive sampling (rotation patterns) is undertaken. Some available guess value of the population median at current occasion is extensively utilized in the proposed sampling strategy. The asymptotic properties of the proposed estimator are studied providing the expressions of its bias and mean square error. The proposed estimator is compared with the sample median estimator when there is no matching. Optimum replacement policy is also discussed. Results have been justified by empirical and pictorial means of elaboration with the help of some natural populations.

Key words: Population median, guess value, successive (rotation) sampling, mean square error, optimum replacement policy.

1. Introduction

A variety of practical problems could fall in the arena of applied and environmental sciences in which the various characters opt to change over time with respect to different parameters. Such changes are the inherent behavior of the nature. Some type of changes directly or indirectly affects the quality of living and surrounding of the human beings. Such changes draw the attention of human intelligentsia to know the pattern or the rate of change at different points

[1] Department of Applied Mathematics, Indian School of Mines University, Dhanbad- 826 004, India, E-mail: gnsingh_ism@yahoo.com.

[2] Department of Applied Mathematics, Indian School of Mines University, Dhanbad- 826 004, India, E-mail: priyanka_ismd@yahoo.co.in.

(occasions) of time or to know the amount of change at any given point of time (occasion) or simultaneously to know both the situations. This requires the continuous monitoring of the real life situation in hand. If the situations required to be monitored are concerned with very large group of individuals, it is difficult, time taking and costly affair. For example, an investigator or owner of the industry of cold drinks may be interested in the following type of problems: (a) The average or total sale of cold drink for the current season; (b) The change in average sale of cold drinks for two different seasons; or (c) Simultaneously to know both (a) and (b).

The follow-up of objectives is carried out by means of sampling on successive occasions (over years or seasons or months) according to a specific rule, with partial replacement of units, called successive (rotation) sampling. This rotation (successive) sampling provides a strong tool for generating the reliable estimates at different occasions. The problem of sampling on two successive occasion was first considered by Jessen (1942), and latter this idea was extended by Patterson (1950), Narain (1953), Eckler (1955), Kulldorff (1963), Rao and Graham (1964), Sen (1972, 73), Adhvaryu (1978), Gordon (1983), Singh et al. (1991), Arnab and Okafar (1992), Feng and Zou (1997), Biradar and Singh (2001), Singh and Singh (2001), Singh (2003, 05), Singh and Priyanka (2006a, 06b) and many others. All the above studies were concerned with the estimation of population mean on two or more occasions.

Frequently there are many problems of practical interest that involves variables with extreme values, which strongly influence the value of mean. In such situations the study variables is having highly skewed distributions. For example, the study of environmental issues, the study of social evil such as number of abortions, the study of income, expenditure etc. In such situations, the estimation of mean does not suffice the purpose, so the estimation of median deserves special attention. Sample medians have also long been recognized as simple robust alternatives to the sample means in estimating the location of markedly skewed population from simple random samples. Gross (1980), Sedransk and Meyer (1978) and Smith and Sedransk (1983) have considered the problem of estimation of the median using simple random sampling. Kuk and Mak (1989) are the first researchers to attempt the estimation of the median using auxiliary information. Rao et al. (1990) and Francisco and Fuller (1991) have also considered the problem of estimation of the median as a part of estimation of finite population distribution function. Further works on median estimation were extended by Meeden and Vardeman (1991), Mak and Kuk (1993), Kuk and Mak (1994), Meeden (1995), Rueda et al (1998), Rueda and Arcos (2001), Singh et al. (2001), Singh and Joarder (2002), Rueda and Arcos (2002), Allen et al. (2002), Singh et al. (2003), Singh (2003) and Singh et al. (2004). Singh, Joarder and Tracy (2001) have generalized the work of Kuk and Mak (1989, 94) and Chen and Qin (1993) in double sampling.

We propose to investigate in the present work some theories of successive (rotation) sampling applied to the median estimation. The main objective is to propose a new estimator for estimating the finite population median at current occasion in two occasions successive (rotation) sampling. It has been assumed that some guess value of the population median to be estimated is available at current occasion. This guess value has been extensively utilized in the proposed sampling strategy. The asymptotic properties of the proposed estimator are studied providing the expressions of its bias and mean square error. The proposed estimator is compared with the sample median estimator when there is no matching. Optimum replacement policy is also discussed. Results have been supported with empirical and pictorial means of representation with the help of some natural populations.

2. Notations

Let U = (U_1, U_2, - - -, U_N) be the finite population of N units, which has been sampled over two occasions. The character under study is denoted by x (y) on the first (second) occasions respectively. A simple random sample (without replacement) of n units is taken on the first occasion. A random sub sample of m = n λ units is retained (matched) for use on the second occasion. Now, at the current occasion a simple random sample (without replacement) of u = (n-m) = nμ units is drawn afresh from the remaining (N-n) units of the population so that the sample size on the second occasion is also n. λ and μ (λ+ μ =1) are the fractions of matched and fresh samples respectively at the second (current) occasion. The following notations are considered for the further use:

M_x, M_y : The population median of the variables x and y respectively.

M_g: Guess value of the population median at current occasion.

$\hat{M}_{x(n)}$: The sample median at first occasion.

$\hat{M}_{x(m)}$, $\hat{M}_{y(m)}$: The sample median of the matched sample on the first and the second occasions respectively.

$\hat{M}_{y(u)}$: The sample median of the unmatched sample on the current occasion.

P_{xy} : The proportion of elements in the population such that $x \le M_x$ and $y \le M_y$.

$f_x\left(M_x\right)$, $f_y\left(M_y\right)$: The marginal densities of x and y respectively.

3. Proposed Estimator

To estimate the population median M_y on the second occasion, two independent estimators are suggested. One is based on sample of size $u\,(= n\mu)$ drawn afresh on the second occasion is given by

$$T_{1u} = W\,\hat{M}_{y(u)} + (1 - W)M_g \tag{1}$$

where W is a constant chosen to achieve the minimum variance of the estimator T_{1u}.

Second estimator is a ratio type estimator based on the sample of size $m\,(= n\lambda)$ common with both the occasions and is defined as

$$T_{2m} = \hat{M}_{y(m)}\left[\frac{\hat{M}_{x(n)}}{\hat{M}_{x(m)}}\right] \tag{2}$$

Considering the convex linear combination of the estimators T_{1u} and T_{2m}, we have the final estimator of M_y as

$$\hat{T} = \varphi\, T_{1u} + (1 - \varphi)T_{2m} \tag{3}$$

where φ is an unknown constant to be determined to achieve the minimum mean square error of the estimator $\hat{T}$.

4. Bias and Mean Square Error of $\hat{T}$

Following Kuk and Mak (1989) and Singh et. al (2001), let $p_{x(m)}$ and $p_{y(m)}$ denote the proportion of x and y values in the sample for which $x \le \hat{M}_x$ and $y \le \hat{M}_y$ respectively. P_{xy} is the proportion of elements in the population such that $x \le M_x$ and $y \le M_y$. Further, it is also assumed that as $N \to \infty$ the distribution of the bivariate variable (X, Y) approaches a continuous distribution with marginal densities $f_x(x)$ and $f_y(y)$ for x and y respectively. The proposed estimator defined in equation (3) is biased for $\hat{T}$. So, its bias B (.) and mean square error MSE (.) up to the first order of approximations are derived under the following large sample approximations:

$$\hat{M}_{y(m)} = M_y\left(1 + e_{0(m)}\right),\ \hat{M}_{y(u)} = M_y\left(1 + e_{0(u)}\right),\ \hat{M}_{x(m)} = M_x\left(1 + e_{1(m)}\right),$$

$\hat{M}_{x(n)} = M_x\left(1 + e_{(n)}\right)$ such that E ($e_{0(m)}$) = E ($e_{0(u)}$) = E ($e_{1(m)}$) = E ($e_{1(n)}$) = 0,

$$E\left(e_{0(m)}^2\right) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4M_y^2}, E\left(e_{0(u)}^2\right) = \left(\frac{1}{u} - \frac{1}{N}\right)\frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4M_y^2}$$

$$E\left(e_{1(m)}^2\right) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{\left\{f_x\left(M_x\right)\right\}^{-2}}{4M_x^2},\ E\left(e_{1(n)}^2\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\left\{f_x\left(M_x\right)\right\}^{-2}}{4M_x^2},$$

$$E\left(e_{0(m)}e_{1(m)}\right) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{\left(4P_{xy} - 1\right)\left\{f_y\left(M_y\right)\right\}^{-1}\left\{f_x\left(M_x\right)\right\}^{-1}}{4M_xM_y}$$

$$\text{and } E\left(e_{1(m)}e_{1(n)}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\left\{f_x\left(M_x\right)\right\}^{-2}}{4M_x^2}.$$

Under the above transformations the estimators T_{2m} takes the following form:

$$T_{2m} = M_y\left(1 + e_{0(m)}\right)\left(1 + e_{1(n)}\right)\left(1 + e_{1(m)}\right)^{-1},$$

further we assume that $\left|e_{1(m)}\right| < 1$. Thus, we have the following theorems.

Theorem 4.1: The bias of the estimator $\hat{T}$ is given by

$$B\left(\hat{T}\right) = \varphi B\left(T_{1u}\right) + \left(1 - \varphi\right)B\left(T_{2m}\right) \tag{4}$$

where $B\left(T_{1u}\right) = (1 - W)\,(M_g - M_y)$ (5)

and

$$B\left(T_{2m}\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\left[\frac{\left\{f_x\left(M_x\right)\right\}^{-2}M_y}{4\,M_x^2} - \frac{\left(4P_{xy} - 1\right)\left\{f_x\left(M_x\right)\right\}^{-1}\left\{f_y\left(M_y\right)\right\}^{-1}}{4M_x}\right] \tag{6}$$

Proof: The bias of $\hat{T}$ is given by

$$B\left(\hat{T}\right) = E\left(\hat{T} - M_y\right) = \varphi B\left(T_{1u}\right) + \left(1 - \varphi\right)B\left(T_{2m}\right)$$

$$B\left(T_{1u}\right) = E\left[T_{1u} - M_y\right] = (1 – W)\ (M_g - M_y)$$

and

$$B\left(T_{2m}\right) = E\left[T_{2m} - M_y\right] = E\left[M_y\left(1 + e_{0(m)}\right)\left(1 + e_{1(n)}\right)\left(1 + e_{1(m)}\right)^{-1} - M_y\right]$$

Now expanding the right hand side of above expression binomially, and retaining the terms up to the first order of approximation, we have

$$B\left(T_{2m}\right) = M_y\, E\left[e_{0(m)}e_{1(n)} - e_{0(m)}e_{1(m)} - e_{1(n)}e_{1(m)} + e_{1(m)}^2\right]$$

$$= \left(\frac{1}{m} - \frac{1}{n}\right)\left[\frac{\left\{f_x\left(M_x\right)\right\}^{-2} M_y}{4\, M_x^2} - \frac{\left(4P_{xy} - 1\right)\left\{f_x\left(M_x\right)\right\}^{-1}\left\{f_y\left(M_y\right)\right\}^{-1}}{4M_x}\right]$$

which is same as the expression for bias given in equation (6).

Theorem 4.2: The mean square error of the estimator $\hat{T}$ is given by

$$MSE\left(\hat{T}\right) = \varphi^2 M\left(T_{1u}\right)_{opt} + \left(1 - \varphi\right)^2 M\left(T_{2m}\right) \tag{7}$$

where

$$M\left(T_{1u}\right)_{opt} = \frac{a\alpha^2}{a + u\,\alpha^2} \tag{8}$$

and

$$M\left(T_{2m}\right) = \frac{a}{m} + \left(\frac{1}{m} - \frac{1}{n}\right)\left(b - c\right) \tag{9}$$

where

$$\alpha = \left(M_g - M_y\right),\quad a = \frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4},\quad b = \frac{M_y^2}{M_x^2}\frac{\left\{f_x\left(M_x\right)\right\}^{-2}}{4}$$

and

$$c = \frac{\left(4P_{xy} - 1\right)M_y\left\{f_y\left(M_y\right)\right\}^{-1}\left\{f_x\left(M_x\right)\right\}^{-1}}{M_x}$$

Proof: Since T_{1u} and T_{2m} are based on two independent samples, therefore for large N (i.e. N→ ∞), the covariance type term has been ignored. Hence

$$M\left(\hat{T}\right) = \varphi^2 M\left(T_{1u}\right) + \left(1-\varphi\right) M\left(T_{2m}\right)$$

where

$$M\left(T_{1u}\right) = E\left[T_{1u} - M_y\right]^2 = W^2\ V\left(\hat{M}_{y(u)}\right) + \left(1-W\right)^2 \left(M_g - M_y\right)^2$$

The above expression is a function of W, hence minimizing for W, we have

$$\frac{\partial M\left(T_{1u}\right)}{\partial W} = 0\text{, this gives } W_{opt} = \frac{\alpha^2}{V\left(\hat{M}_{y(u)}\right) + \alpha^2} = W^*\text{ (say)}$$

Hence, the minimum mean square error of T_{1u} is given by

$$M\left(T_{1u}\right)_{opt} = \frac{\alpha^2 V\left(\hat{M}_{y(u)}\right)}{\alpha^2 + V\left(\hat{M}_{y(u)}\right)} = \frac{\alpha^2\left(\frac{1}{u} - \frac{1}{N}\right)\frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4}}{\alpha^2 + \left(\frac{1}{u} - \frac{1}{N}\right)\frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4}} \tag{10}$$

Now, since the population size is sufficiently large (i.e., $N \to \infty$) therefore, the finite population correction (fpc) is ignored. In the light of this assumption the mean square error of T_{1u} takes the following form.

$$M\left(T_{1u}\right)_{opt} = \frac{a\alpha^2}{a + u\alpha^2}\text{, where } a = \frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4}$$

Now, using the transformations considered above, we have the following steps for evaluation of mean square error of T_{2m}.

$$M\left(T_{2m}\right) = E\left[T_{2m} - M_y\right]^2 = E\left[\hat{M}_{y(m)}\frac{\hat{M}_{x(n)}}{\hat{M}_{x(m)}} - M_y\right]^2$$

$$= E\left[M_y\left(1 + e_{0(m)}\right)\left(1 + e_{1(n)}\right)\left(1 + e_{1(m)}\right)^{-1} - M_y\right]^2$$

Expanding the right hand side of the above expression binomially, retaining the terms up to the first order of approximation, we have

$$M\left(T_{2m}\right)=M_y^2E\left[e_{0(m)}+e_{1(n)}-e_{1(m)}\right]^2=\left(\frac{1}{m}-\frac{1}{N}\right)a+\left(\frac{1}{m}-\frac{1}{n}\right)\left(b-c\right) \quad (11)$$

where $b=\frac{M_y^2}{M_x^2}\frac{\left\{f_x\left(M_x\right)\right\}^{-2}}{4}$ and $c=\frac{\left(4P_{xy}-1\right)M_y\left\{f_y\left(M_y\right)\right\}^{-1}\left\{f_x\left(M_x\right)\right\}^{-1}}{M_x}$

For the large population size (i.e. $N\rightarrow\infty$), finite population corrections are ignored, hence the result of equation (11) reduces to equation (9).

5. Minimum Mean Square Error of $\hat{T}$

Since, mean square error of $\hat{T}$ in equation (7) is a function of unknown constant φ, therefore, it is minimized with respect to φ and subsequently the optimum value of φ is obtained as

$$\varphi_{opt}=\frac{M\left(T_{2m}\right)}{M\left(T_{1u}\right)_{opt}+M\left(T_{2m}\right)} \quad (12)$$

Substituting the value of φ_{opt} in equation (7), we get the optimum variance of $\hat{T}$ as

$$M\left(\hat{T}\right)_{opt}=\frac{M\left(T_{1u}\right)_{opt}.M\left(T_{2m}\right)}{M(T_{1u})_{opt}+M(T_{2m})} \quad (13)$$

Further, substituting the values from equations (8) and (9) in equation (13) the simplified value of $M(\hat{T})_{opt}$ is shown in theorem (5.1).

Theorem 5.1: The minimum mean square error of $\hat{T}$ (i.e., $M(\hat{T})_{opt}$) is derived as

$$M\left(\hat{T}\right)_{opt}=\frac{t_4+t_5\mu}{[t_1\mu^2+t_2\mu+t_3]} \quad (14)$$

where $\alpha=\left(M_g-M_y\right)$, $t_1=n\alpha^2\left(b-c\right), t_2=a\left(b-c\right)$, $t_3=\alpha^2na+a^2$, $t_4=\alpha^2a^2$, $t_5=\alpha^2a\left(b-c\right)$ and $\mu\left(=\frac{u}{n}\right)$ is the fraction of fresh sample taken at second (current) occasion.

6. Optimum Replacement Policy

The key design parameter affecting the estimates of change is the overlap between successive samples. Maintaining high overlap between repeats of a survey is operationally convenient, since many sampled units have been located and have some experience of the survey. Hence, to determine the optimum value of μ (fraction of sample to be taken afresh at second occasion) so that M_y may be estimated with maximum precision, we minimize $M\left(\hat{T}\right)_{opt}$ in (14) with respect to μ and hence we get

$$\hat{\mu} = \frac{-t_1 t_4 \pm \sqrt{t_1^2 t_4^2 + t_1 t_5 \left(t_3 t_5 - t_2 t_4\right)}}{t_1 t_5} \tag{15}$$

The real values of $\hat{\mu}$ exist if $t_1^2 t_4^2 + t_1 t_5 \left(t_3 t_5 - t_2 t_4\right) \geq 0$. For any situation which satisfies this condition, two values of $\hat{\mu}$ are possible, hence to choose a value of $\hat{\mu}$, it should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values of $\hat{\mu}$ are inadmissible. Substituting the admissible value of $\hat{\mu}$ (say μ_0) from equation (15) in equation (14), we have

$$M\left(\hat{T}\right)_{opt^*} = \frac{t_4 + t_5 \mu_0}{[t_1 \mu_0^2 + t_2 \mu_0 + t_3]} \tag{16}$$

7. Efficiency Comparison

The percent relative efficiencies of $\hat{T}$ with respect to $\hat{M}_{y(n)}$ when there is no matching have been obtained for different choices of n and $\delta \left(= \frac{M_g}{M_y} \right)$.

Since, $\hat{M}_{y(n)}$ is the sample median, hence the variance of $\hat{M}_{y(n)}$ for large N (i.e., $N \to \infty$) is given by

$$V\left(\hat{M}_{y(n)}\right) = \frac{\left\{f_y\left(M_y\right)\right\}^{-2}}{4\,n} \tag{17}$$

For different choices of n and δ, table 1 shows the optimum value of μ i.e., μ_0 and table 2 shows the percent relative efficiencies E of $\hat{T}$ (under optimal condition) with respect to $\hat{M}_{y(n)}$, where

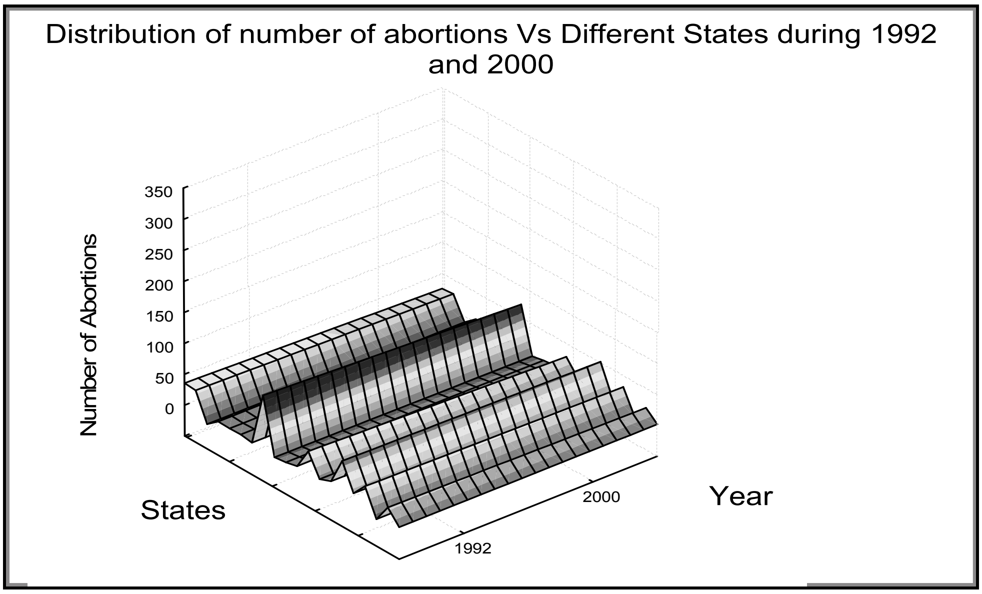
$$E=\frac{V\left(\hat{M}_{y(n)}\right)}{M\left(\hat{T}\right)_{opt^*}}\times 100 \qquad (18)$$

To have the clear-cut idea, the percent relative efficiencies of $\hat{T}$ with respect to $\hat{M}_{y(n)}$ are viewed through suitable graphs, which are shown in figures 3—6.

8. Empirical Study

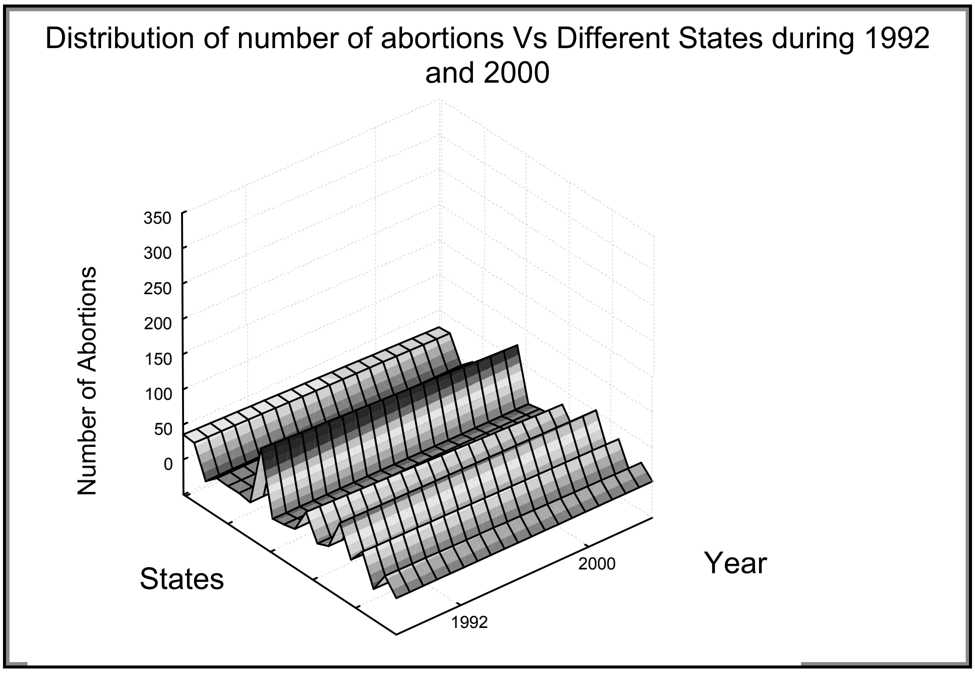
Populations Source: [Free access to the data by the Statistical Abstracts of the United States]; for the purpose of empirical study we have considered two natural populations. In the first case the aim is to analyze the educational status of US in the year 2004. The population consists of N = 50 states. Further, let Y_i (study variable) be the percent of bachelor degree holders or more in the year 2004 and X_i be the percent of bachelor degree holders or more in the year 2000 in the ith state of US. The following graph in figure 1 shows that the distribution of percent of educational attainment in different states is skewed towards right.

Figure 1. Distribution of educational attainment



Similarly, in the second natural population, the aim is to estimate the number of abortions in different states during 1992 and 2000. Here, also the population consists of N = 50 states. Let Y_i (study variable) denote the number of abortions during 2000 and X_i be the number of abortions in the year 1992 in the i^{th} state of US. The following graph in figure 2 shows that the distribution of number of abortions in different states is also skewed towards right.

Figure 2. Distribution of number of abortions



One reason of skewness in both the above cases may be the distribution of population in different states, that is, the states having large populations are expected to have large number of abortion cases or larger percent of bachelor's degree holders or more. Thus, the skewness of the data indicates that the use of median is a good measure of central location than mean in such situations. The sample size at both the occasions is assumed to be the same. Here, it has been assumed that in both the above cases the study variable (dynamic in nature) follows independent normal distributions. Table 1 shows the optimum fraction of the sample to be drawn at current occasion for the two populations. The percent relative efficiency of the proposed estimator for different values of $\delta = \frac{M_g}{M_y}$ is calculated. The above experiments were repeated for different sample sizes n = 10, 12, 15, 20 and 25 for both the populations described above and are tabulated in table 2. On the basis of above description, the values of the different required parameters for both the populations are calculated as follows:

Population I:

$N = 50$, $M_y = 25.5$, $M_x = 24.6$, $\mu_y = 27.174$, $\mu_x = 25.19412$, $\sigma_x = 4.661348$, $\sigma_y = 5.40155$ and $P_{xy} = 0.4510$.

Also, we have $f(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}$ and $f(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2}$

Which implies that $f(M_x) = 0.0849$ and $f(M_y) = 0.0704$

Population II:

$N = 50$, $M_y = 12.0$, $M_x = 14.5$, $\mu_y = 26.34$, $\mu_x = 30.56$, $\sigma_x = 51.2983$, $\sigma_y = 42.77688$, $P_{xy} = 0.46$, $f(M_x) = 0.0074$ and $f(M_y) = 0.0088$

Table 1. Optimum value of the fraction of sample to be drawn afresh at current occasion.

Population	I	II
Optimum value of μ	0.6403	0.6260

Table 2. Percent relative efficiencies of the proposed estimator for different choices of n and δ

	δ / n	0.1	0.3	0.5	0.7	0.9	1.2	1.4	1.6	1.8	2.0
Population I	10	403.66	583.64	1021.0	2608.4	22451.0	5708.9	1523.3	748.16	476.87	351.30
	12	357.73	507.71	872.18	2195.0	18731.0	4778.8	1290.7	644.81	418.73	314.09
	15	311.79	431.78	723.36	1781.6	15010.0	3848.6	1058.2	541.46	360.60	276.89
	20	265.86	355.85	574.53	1386.3	11290.0	2918.5	825.67	438.11	302.47	239.68
	25	238.30	310.30	485.24	1120.2	9057.4	2360.4	686.15	376.10	267.59	217.36
Population II	10	126.17	126.79	128.31	133.83	202.76	144.60	130.06	127.36	126.42	125.98
	12	126.01	126.53	127.79	132.39	189.84	141.36	129.25	127.00	126.22	125.86
	15	125.85	126.26	127.28	130.95	176.91	138.13	128.44	126.64	126.02	125.73
	20	125.69	126.00	126.76	129.52	163.98	134.90	127.63	126.29	125.81	125.60
	25	125.59	125.84	126.45	128.66	156.23	132.96	127.15	126.07	125.69	125.52

Figure 3. PRE of $\hat{T}$ for Population I for n = 10

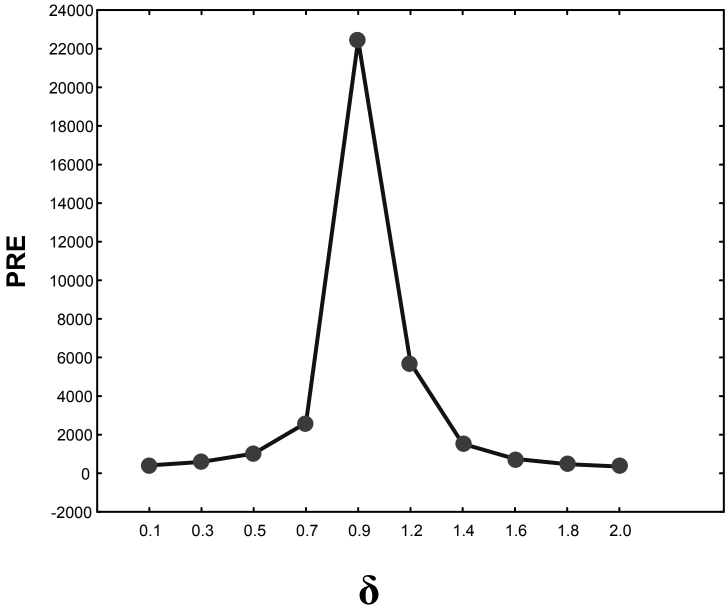


Figure 4. PRE of $\hat{T}$ for Population I for n = 20

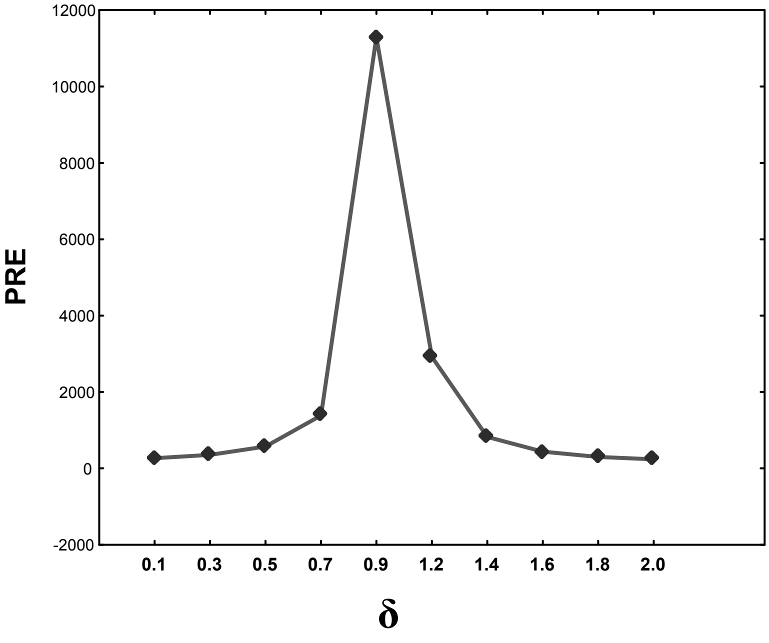


Figure 5. PRE of $\hat{T}$ for Population II for n = 12

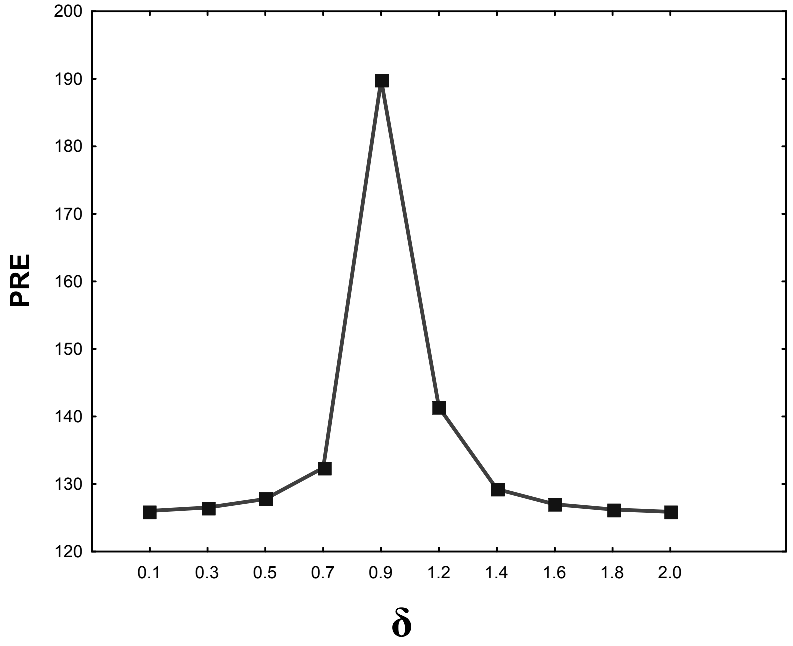
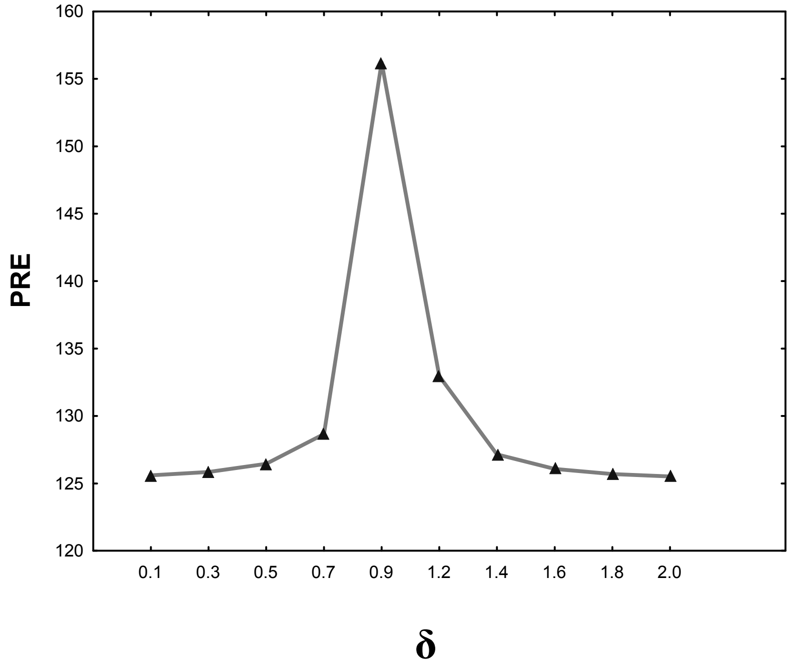


Figure 6. PRE of $\hat{T}$ for Population II for n = 25



From the tables and different graphs, it is strongly vindicated that the availability of the guess value of the median at the current occasion is highly rewarding in terms of precision. The gain on precision is highly significative when the value of δ approaches 1, i.e., guesses value is almost equal to the median at the current occasion.

9. Conclusion

Recapitulating what have been discussed in earlier sections, it can be noted that there are many situations of practical importance where the possibility of the distribution to be skewed is there together with the value of the study character may be time dependent. In such situation the estimation of median is quite fruitful at successive occasions. The analytic and empirical results support the fact that the proposed estimator for median estimation at current occasion is quite feasible.

Generalization of the above proposed estimator when more than two occasions are considered is achievable without a great complexity.

Acknowledgement

We appreciate the free access and acknowledge the use of data from the Statistical Abstract of the United States, 2001 that is available on internet.

REFERENCES

ALLEN, J., SAXENA, S., SINGH, H. P., SINGH, S. and SMARANDACHE, F. (2002) Randomness and optimal estimation in data sampling. American Research Press: 26—43.

ADHVARYU, D. E. (1978) Successive Sampling using multi-auxiliary information. Shankhya 40C: 167—173.

ARNAB, R and OKAFOR, F. C. (1992) A note on double sampling over two occasions. Pakistan JNL of statistics: 8(3), B

BIRADAR, R. S. and SINGH, H. P. (2001) Successive sampling using auxiliary information on both occasions. Cal. Stat. Assoc. Bull., 51: 243—251.

CHAMBER, R. L. and DUNSTAN, R. (1986) Estimating distribution functions from survey data. Biometrika 73: 579—604

CHEN, J. and QIN, J. (1993) Empirical likelihood estimation for finite population and the effective usage of auxiliary information. Biometrika 80: 107—116.

ECKLER, A. R. (1955) Rotation Sampling. Ann. Math. Statist.: 664—685

FENG, S. and ZOU, G. (1997) Sample rotation method with auxiliary variable. Commun.Statist. Theo-Meth. 26: 6, 1497—1509.

FRANSISCO, C. A. and FULLER, W. A. (1991) Quantile estimation with a complex survey design. Ann. Statist. 19: 454—469.

GORDON, L. (1983) Successive sampling in finite populations. The Annals of statistics 11(2): 702 — 706

GROSS, S. T. (1980) Median estimation in sample surveys. Proc. Surv. Res. Meth. Sect. Amer. Statist. Assoc.: 181—184.

JESSEN, R. J. (1942) Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Station Road Bulletin No. 304, Ames:,1—104.

KUK, A. Y. C. and MAK, T. K. (1989) Median estimation in presence of auxiliary information. J. R. Statit. Soc. B, 51: 261—269.

KUK, A. Y. C. and MAK, T. K. (1994) A functional approach to estimating finite population distribution function. Comm. Statist. Theory Methods 2:, 883—896.

KULLDORFF, G. (1963) Some problems of optimum allocation for sampling over two occasions. Rev. Inter. Statist. Inst. 31: 24—57.

LEVIN, S. G., YONG, R. W. and STOHLER, R. L. (1992) Estimation of median human Lethal radiation dose computed from data on occupants of reinforced concrete structures in Nagasaki, Japan. Health Phys. 63(5): 522—531.

MAK, T. K. and KUK, A. (1993) A new method for estimating finite population quintiles using auxiliary information. Canadian J. Statist. 21: 29—38.

MEEDEN, G. and VARDEMAN, S. (1991) A non informative Bayesian approach to interval estimation in finite population sampling. J. Amer. Statist. Assoc. 86: 972—980.

MEEDEN, G. (1995) Median estimation using auxiliary information. Survey Methodology 21(1): 71—77.

NARAIN, R. D. (1953) On the recurrence formula in sampling on successive occasions. Journal of the Indian Society of Agricultural Staistics 5: 96—99.

PATTERSON, H. D. (1950) Sampling on successive occasions with partial replacement of units. Jour. Royal Statist. Assoc., Ser. B, 12: 241—255.

RAO, J. N. K. and GRAHAM, J. E. (1964) Rotation design for sampling on repeated occasions. Jour. Amer. Statist. Assoc. 59: 492—509.

RAO, J. N. K., KOVAR, J. G. and MANTEL, H. J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, vol. 77: 2, 365—375.

RUEDA, M. and ARCOS, A. (2002) The use of quantiles of auxiliary variable to estimate medians. Biom. J., 44: 5, 619—632.

RUEDA, M., ARCOS, A. and ARTES, E. (1998) Quantile interval estimation in finite population using a multivariate ratio estimator. Metrika 47: 203—213.

SEDRANSK, J. and MEYER, J. (1978) Confidence intervals for quantiles of a finite population: Simple random and stratified simple random sampling. J. R. Statist. Soc., B, 40: 239—252.

SEN, A. R. (1971) Successive sampling with two auxiliary variables. Sankhya Ser. B 33: 371—378.

SEN, A. R. (1972) Successive sampling with p $(p \geq 1)$ auxiliary variables. Ann. Math. Statist. 43: 2031—2034.

SEN, A. R. (1973) Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics 29: 381—385.

SINGH, V. K., and SINGH, G. N. (1991) Chain-type regression estimators with two auxiliary variables under double sampling scheme. Merton 49: 279—289.

SINGH, G. N. and SINGH, V. K. (2001) On the use of auxiliary information in successive sampling. Jour. Ind. Soc. Agri. Statist. 54(1): 1—12.

SINGH, G. N. (2003) Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling. Statistics in Transition. 6(4): 523—532.

SINGH, G. N. (2005) On the use of chain-type ratio estimator in successive sampling. Statistics in Transition 7(1): 21—26.

SINGH, G. N. and PRIYANKA, K. (2006 a) On the use of chain-type ratio to difference estimator in successive sampling. IJAMAS S06 5: 41—49

SINGH, G. N. and PRIYANKA, K. (2006 b) Search of good rotation patterns to improve the precision of the estimates at current occasion. Accepted in Communications in Statistics (Theory and Methods).

SINGH, V. K. and SHUKLA, D. (1987) One parameter family of factor type ratio estimators, Metron 45: 1—2, 30, 273—283.

SINGH, H. P., SINGH, S. and UPADHYAYA, L. N. (2004) Chain ratio and regression type estimator for median estimation in survey sampling. Statistical papers (In press).

SINGH, S. (2003) Advanced Sampling Theory with Applications: How Michael 'selected' Amy. (Vol. 1 and 2) pp 1—1247, Kluwer Academic Publishers, The Netherlands.

SINGH, S., JOARDER, A. H. and TRACY, D. S. (2001) Median estimation using double sampling. Aust. N. Z. J. Statist. 43(1): 33—46.

SINGH, H. P., SINGH, S. and PUETAS, S.M. (2003) Ratio type estimator for the median of finite populations. Allgemeines Statistiches Archiv: 369—382.

SINGH, S. and JOARDER, A. H. (2002) Estimation of distribution function and Median in two-phase sampling. Pak J. Statist. 18(2): 301—319.

SMITH, P. and SEDRANSK, J. (1983) Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. Commun. Statist.—Theory Meth. 12: 1329—1344.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S. and ASHOK, C. (1984) Sampling theory of surveys with applications. Iowa State University Press, Iowa, USA, 3rd Revised Edition.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 233—258

EXISTENCE OF THE BLUE FOR FINITE POPULATION MEAN UNDER MULTIPLE IMPUTATION

Pulakesh Maiti

ABSTRACT

Missing values not only mean less efficient estimates because of reduction in sample size, but also mean that the standard complete data methods cannot be immediately used to analyse the data. Imputation, single or multiple, is a compensatory method for handling non-responses and takes care of the fact that once the values have been filled in, standard complete data methods of analysis can be used. Here, in this paper, using multiple imputation technique, an estimator for the finite population mean in the presence of unit non-response has been proposed and the estimator so proposed has been found to be the BLUE. A very general cost model has been discussed in the presence of non-response and an optimal solution of sample size for a given number of imputations or of number of imputations for a given sample size has been worked out.

1. Introduction

A survey may mean to include census which attempts to collect information from each member in the population, whereas a sample survey refers to a survey in which a scientific sample of the population is studies i.e., the same sort of information is sought only for some of the units, — those in the sample. The choice of the sample is carefully made in order to draw inferences about the parameters of the population under study, but in many censuses and sample surveys, some of the selected units may not be possible to be contacted, and even contacted do not respond to at least some of the items being asked. Such non-responses which are known as survey non-response, whether it arises from a census or a sample survey is common.

The problem created by non-responses is of course that of non-availability of the complete data i.e., some of the values intended by the sampling design to be observed are in fact missing, and these missing values not only mean less efficient estimates because of the reduction in size in the data base, but also mean that standard complete data methods can not be immediately used to analyse the data.

Non-response as a concept has been defined in a number of ways. Most definitions distinguish **unit non-response** from **item non-response**. In general, non-response has been attributed to failure to obtain a response to a particular unit and/or to particular item when the questionnaire has been canvassed and has been completed partially or not been responded at all. [Kendall and Buckland (1960), Kish (1965), Bureau of census (1957, 1976), Cochran (1979), Zarkovieh (1966), Ford (1976), Sudman (1976), Suchman (1962), Wark and Lilinger (1975), Deghton et al. (1978), Deming (1953)]. An extended definition of non-response includes in which missing data arise from the processing of information provided by units rather than refusal of units.

1.1.Non-response of different types at different levels

In household surveys with multistage design, non-responses can occur at different hierarchical stages singly or jointly, say, at the PSU (village), at the household and at the individual data item singly or jointly.

The first stage sampling unit, say, the village/urban block may be temporarily inaccessible or might have altered in character as a village might have become urbanized since last census. In spite of all the efforts, a few casualties do occur in a large scale survey operation like NSS. The extent of non-response at this level is of the order of 0.5 to 1% [Sarma, Rao and Ambe (1980)].

The existence of non-responses at the household level has been well exhibited through many surveys from both developed and developing Countries [Thompson I.b. and Siring E., Scott and Singh (1980), Verma (1980), US Current Population Survey (CPS) (1959—1978)., etc. There, non-response rates have been found to differ by reason in all the surveys conducted both in developing and developed Countries. Non-responses due to refusal are more in developed Countries, whereas non-responses due to non-contact are more in developing Countries. It has also been observed in Verma (1980), that in fertility surveys conducted in each of twenty developing Countries, each of non-response rates due to “vacant-dwelling unit”, “due to not at home” is greater than that due to refusal and in all the Countries except Malayasia, Jamaica, Costarica, Mexica, Panama, non-response rate due to “not possible to locate” is greater than non-response rate due to refusal. Also one may consult the article by Maiti (2007) for demonstration of existence of non-responses in personal interviews.

The third level at which non-response occurs is at individual data level, which is popularly known as **item non-response**. There may be many reasons behind **item non-response**. To mention a few socio-economic background as well as the sensitivity of the specific item of information may be responsible for such non-responses. In fact, item non-responses are very much sensitive to particular type of items. For a good account of the extent of item non-responses on different items in different surveys, one can refer to the work by Sarma et al. (1980). Also one may consult paper by Dhar, N.R. (1971) and Maiti P. (2007).

One can also find through many surveys conducted both in developed and developing Countries that non-response rates are gradually increasing over the years, and hence call for strong attention. The work of assessing non-responses, though dates back to the forties [Deming (1944), (1950); Mahalanobis (1940), (1944), (1964); Moser (1958), Zarkovich (1966), Dalenius (1977a), (1977b), (1977c); Kish (1965), Sarndal et al. (1992)], but handling them with mathematical rigour is a new addition [Rubin (1987), Lessler and Kalsbeek (1996)] to the literature.

1.2. Two views of non-response: Deterministic and Probabilistic

The mechanism which generates non-response/response may be deterministic or may be stochastic in nature. By preassuming that members of the population are other certain to respond $(p_i = 1)$ or $(p_i = 0)$, the deterministic view of non-response rules out any uncertainty on whether or not each member of the population would provide usable data for the survey, if selected. Thus decision on whether to response or non-response is pre-determined. Reviews by Ford (1976) and Kalton (1983) provide extensive analytical discussions of non-response and non-response compensation procedures developed from a deterministic procedure.

Under stochastic view, each R_i associated with i^{th} sample unit, is a random variable whose outcome is determined by an assumed chance element in the response process. Associated with each R_i is the response probability p_i which may differ among different members of the population. [Politz and Simmons (1949), Hartley (1946), Deming (1953), Platek et al. (1977), Lessler (1983) etc.]

1.3. Different methods of dealing with non-response: Compensatory as well as Preventive

Several methods have been tried to compensate for the effect of non-response on the survey result. Some of these methods are part of **data collection procedure**; for example, intensive follow up of a subsample of non-respondents [Hansen, Hurwitz (1946), Fellegi and Sunter (1974), Platek et al. (1977)] or the collection of limited data through proxy interview from neighbours (Roshwab (1982) or though Call-backs (Birbaum and Sirken (1950), Durbin (1954), Deming (1953), Kish (1965), Kendal and Buckland (1972), Moser and Kalton (1972), Cochran (1977), Deighton et al. (1978). The substitution of other units for non-responding units is a controversial practice.

Other procedures, generally less costly, are used during data processing. These come under the general headings of **imputation** and **estimation procedures** which attempt to compensate for missing data.

There are other types of measures which may be termed as **preventive measures**. The preventive measures are those that would be implemented for identification, solicitation and compilation of the questionnaires, so that after the sample member has agreed, at least, in principle to co-operate, relevant data can be made available smoothly. These methods include correcting the frame errors, if any; proper designing of the questionnaires / schedules, providing uniform training to the investigators etc. [Maiti, (2007)].

1.4.Imputation versus Revising the weights under estimation procedures

Adjustment of estimates by revising the weights is for the fact that on measured information, non-respondents differ from respondents. It has been empirically observed through a number of surveys [Lundberg and Larsen (1949), Reuss (1943), Politz and Simon (1949), Birbaum and Sirken (1950), Pan (1951), King and Chen (1957), Buckland (1960), Suchman (1962), Lubin, Levit and Zuckerman (1962), Skelton (1963), Bennet and Hill (1964), Dunn and Hawks (1966), Ognibene (1970), Lessier (1974), U.S. Bureau of Census (1974), Warwiek and Lininger (1975), Roy (1976—77, 1977—78, 1988—89), Sundman (1976), Deighton et al. (1978), Gower (1979), Kalton (1983), Madow et al. (1983), Maiti (1994—95, 1995—96) etc.] that who would form the set of respondents and who would be the non-respondents. Under this view of non-response, in revising the weights, the works, among others, due to Politz and Simon (1949), Hansen et al. (1953), Hartigan (1975), Platek et al. (1977), Kish and Anderson (1978), Bailar et al. (1978), Kohen and Kalsbeek (1981), Drew and Fuller (1980, 1981), Rizvi (1983), Madow (1983) may be mentioned.

In the broad sense, **imputation** means replacing missing or unusable information with usable data from other sources. These sources can include the same questionnaire, another questionnaire from the same survey or external sources, such as another survey or an administrative record.

In case of total/unit non-response, the choice of imputing the questionnaire has to be made from a large group of responding units. **It is at this level, one sees the similarity between weighting and imputation. Such imputation is equivalent to duplicating questionnaires. Duplicating a questionnaire to adjust for a missing unit is equivalent to giving that questionnaire an extra weighting factor of 2.**

From the view point of sampling error, adjusting for non-response by an estimation procedure is preferable to duplication of individual questionnaires. **The only reason for using the latter procedure would be to maintain a sample design that is self-weighting. This is accomplished by actually duplicating the computer record for the selected questionnaire rather than giving an extra weighty of 2**.

Operational simplicity and flexibility of procedures for automatic imputation make them attractive.

1.5.Single Imputation Versus Multiple Imputation

Imputation, whether single or multiple, takes care of the fact that once the values have been filled in, standard complete data methods of analysis can be used. The second advantage of imputation is that in many cases, imputation can be created by incorporating the knowledge of the data collector to reflect the uncertainty about which values to inputs.

Single imputation is useful when possibly substantial efforts are needed to create, but single value being imputed can reflect neither sampling variability about actual value when one model for non-response is being considered, nor additional uncertainty when more than one model is entertained. The obvious problem with single imputation is that the missing value is not known, but automatic application of complete data set treats missing values, as if they were known. Because of this, inferences based on the single imputed data set will be too sharp, since the extra variability due to unknown missing values is not being taken into account.

Multi imputation corrects the major flaws of single imputation. The idea behind multiple imputation is that for each missing value several values, say *m*, instead of just one, are computed. Thus *m* imputations for each missing datum create *m*-complete data sets. The practical difficulty with multiple imputations lies in the necessity of producing multiple data for each missing value. Where single set of imputed values is prohibited, repeating the process may be difficult. **Fortunately, there is some empirical evidence that the number of sets should not be large for multiple imputation to be effective (Rubin and Shanker 1986).**

Multiple imputation using modest *m*, say, $2 \leq m \leq 10$ is designed for situations with a modest fraction of missing information due to non-response.

The organization of this presentation is as follows.

Some results are presented in the next Section 2.1; under the assumption that responses/non-responses are purely random, followed in the next Section 2.2, assuming that each member of the population can be considered as having been labeled "respondent" or "non-respondent", or assigned to a respondent or non-respondent sub class prior to the survey. Finally, a non-linear cost-model has been considered to have an optimal solution of *n* or *m* in Section 3.

2. Existence of the BLUE for Population Mean

2.1.Non-responses occur randomly

Let us consider a finite population $U = \{1, 2, \ldots, N\}$ of *N* number of units labeled 1 through *N*. Let $y_i = y(i)$ be the value of the variable *y* for the i^{th} unit. Let the parameter to be estimated be $\overline{Y} = \Sigma\, y_i / N$.

Sampling Scheme under single imputation: Under SRSWOR (N, *n*), let a typical sample realized be $s(i_1, i_2, \ldots, i_n)$, which after the initial data collection, is partitioned into

$s_{(1)} = (i_1, i_2, \ldots, i_{n_1})$ and $s_{(0)} = (j_1, j_2, \ldots, j_{n_0})$, $n = n_1 + n_0$,where the suffixes *i* and *j* stand for the responding and non-responding units. Thus, after the initial field work, the following situation wises.

Data are available on n_1 number of units	Data are not available on n_0 number of units

To compensate for these unit non-responses, the following imputation method is adopted.

The incomplete data set is completed by imputing the missing values $y_{j_1}, y_{j_2}, \ldots, y_{j_{n_0}}$ through $Z_{j_1}, Z_{j_2}, \ldots, Z_{j_{n_0}}$, where $Z_k' s\,(k = 1, 2, \ldots, n_0)$ are realized through a sample of SRSWR (n_1, n_0) and the incomplete data is completed as

$$\left(y_{i_1}, y_{i_2}, \ldots, y_{i_{n_1}}; Z_{j_1}, Z_{j_2}, \ldots, Z_{j_{n_0}}\right);$$

The incomplete data can be completed by adopting any other imputation method also.

Let $\bar{y}_1$ and $\bar{y}_0$ be the sample means based on $y_{i_k}' s\,(k = 1, 2, \ldots, n_1)$ and $Z_{j_k}' s(k = 1, 2, \cdots, n_0)$.

Let

$$\bar{y}_* = \frac{n_1}{n} . \bar{y}_1 + \frac{n_0}{n} \bar{y}_0 \tag{2.1}$$

then we have the following result.

Theorem 2.1: Under the above sampling schemes, $\bar{y}_*$ is unbiased for $\overline{Y}$ for any given (n_1, n_0) and we have,

$$V(\bar{y}_*) = \begin{cases} \left(S^2/n\right)\left(p_0 + \dfrac{1}{p_1} - f\right), & \text{under } SRSWR(n_1, n_0) \\ \left(S^2/n\right)\left\{\left(p_0 + \dfrac{1}{p_1} - f\right) - p_0^2/p_1\right\}, & \text{under } SRSWOR(n_1, n_0) \end{cases} \tag{2.2}$$

up to the order of $1/n$,

where, $(N-1)S^2 = \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2, f = n/N,\ p_0 = \frac{n_0}{n}$ and $p_1 = \frac{n_1}{n}$.

Proof: The results follow immediately through the observations

$$E\ (\bar{y}_*) = E_{s_1} E_{z/s1}(\bar{y}_*) \text{and } V\ (\bar{y}_*) = V_{s_1} E_{z/s1}(\bar{y}_*) + E_{s_1} V_{z/s1}(\bar{y}_*)$$

and after routine calculations, where the variable *z* stands for imputation.

Remarks:

1. Clearly, $V(\bar{y}_*)\,|_{SRSWOR} \le V(\bar{y}_*)\,|_{SRSWR}$;
2. $\dfrac{V(\bar{y}_*)\mid SRSWR}{V(\bar{y}_*)\mid SRSWOR} = \dfrac{\left(p_0 + \dfrac{1}{n_1} - f\right)}{\left[\left(p_0 + \dfrac{1}{p_1} + f\right) - \left(p^2/p_1\right)\right]}$;
3. If each missing value $y_{jk}\left(k = 1,2,\ldots,n_0\right)$ be imputed by the single value $\bar{y}_1$, then $\bar{y}_*$ becomes equal to $\bar{y}_1$.

Sampling Schemes under multiple imputation.

Under *m*-fold independent $SRS(n_1, n_0)$ from *SRSWOR* (*N*, *n*) i.e., under *m*-tier imputation by *SRS* each time, tires being independent, let the estimator for population mean be defined by

$$\bar{\bar{y}}_* = \sum_{j=1}^{m} \bar{y}_*^{(j)} \Big/ m \tag{2.3}$$

where,

$$\bar{y}_*^{(j)} = \left(\frac{n_1}{n_0}\right)\bar{y}_1 + \left(\frac{n_0}{n}\right)\bar{y}_0^{(j)};$$

then we have the following result.

Theorem 2.2: Under the above sampling schemes and for a given sample (n_1, n_0), we have

(i) $\bar{\bar{y}}_*$ is BLUE for $\bar{Y}$

$$(ii)\;\; V(\bar{y}_*) = \begin{cases} \left(S^2/n\right)\left(\dfrac{1}{p_1} - f\right) + \dfrac{S^2}{nm} p_0; \text{under} m - \textit{fold} \text{independent } SRSWR(n_1, n_0) \\ \\ \left(S^2/n\right)\left(\dfrac{1}{p_1} - f\right) + \dfrac{S^2}{nm}\, \alpha, \text{under} m - \text{foldindependent} SRSWOR(n_1, n_0) \end{cases} \qquad (2.4)$$

where S^2, p_0, p_1, f are defined as before and $\alpha = p_0 . p^*$, $p^* = \left(1 - \dfrac{p_0}{p_1}\right)$, normally $p_0 \le p_1$ and hence $\alpha \le 1$.

Lemma 2.1: Let each of $T_j = \left(j = 1, 2, \ldots, m\right)$ be unbiased for θ and $T_j' s$ are correlated i.e., $E(\underset{\sim}{T}) = \theta \underset{\sim}{1}$, $D(\underset{\sim}{T},) = \Sigma$; $T = (T_1, T_2, \ldots, T_m)'$ and $1 = (1, 1, \ldots, 1)'$, then the BLUE for θ i.e., $\hat{\theta}_{BLUE} = 1'\Sigma^{-1}T \big/ 1'\Sigma^{-1}1$. In particular, when $\Sigma = \left[(a-b)I + bJ\right]$,

then $\hat{\theta}_{BLUE} = \sum\limits_{j=1}^{m} T_j \Big/ m$, where Σ is of $m\ x\ m$.

Proof of theorem 2.2: (i) follows, from Lemma 2.1 by setting $T = \left(\bar{y}_*^{(1)}, \bar{y}_*^{(2)}, \ldots, \bar{y}_*^{(m)}\right)'$ and observing a, b in Σ as $a = \sigma_*^2$ and $b = \rho^* \sigma_*^2$, where $\sigma_*^2 = V\left(\bar{y}_*^{(j)}\right)$ and $\rho^* \alpha_*^2 = Cov\left(\bar{y}_*^{(j)}, \bar{y}_*^{(k)}\right)$.

(ii) follows after routine calculations.

2.2.Occurrence of non-responses are non-random

One may consider the deterministic view to be one in which the outcome of the R-variable for each member of the population has been conditioned. The deterministic view then becomes a conditional form of the stochastic view. Here the population of N-units is assumed to be partitioned into two mutually exclusive and exhaustive subgroups, one with $N_0 = \sum\limits_{i=1}^{N} R_i$ units, which with certainty would respond and $N_0 = N - N_1$ units, which with certainty, would not. The proportions in the respondent subgroup, $P_1 = N_1 / N$ and non-respondent subgroup, $P_0 = N_0 / N$ would depend on the characteristics of the study as well as on some specific features of the population members. Therefore, the population U can be thought of as consisting of two domains, $u_{(1)}$, the domain of respondents and $u_{(0)}$, the domain of non respondents.

$$U = \left(U_{(1)}, U_{(0)}\right),\ N = N_1 + N_0 .$$

Since, here the respondents are systematically different from the non-respondents, biases exist, unless further assumptions, on equality of two group means and/or group variances are assumed.

Let $\bar{y}_*^{(j)} = (j = 1,2,\ldots,m),\ \bar{\bar{y}}_*$ be defined as before, then we have the following result.

Theorem 2.3.: Under the sampling schemes of *SRSWOR (N, n)* and under m-tier $SRS\,(n_1, n_0)$ each time, we have,

(i) $E\left(\bar{y}_*^{(j)}\right) = \bar{Y}$ under the assumption of $\bar{Y}_1 = \bar{Y}_0$, for $j = 1,2,\ldots,m$, $\bar{Y}_1$ and $\bar{Y}_0$ being to group means;

(ii) $$V\left(\bar{\bar{y}}_*\right) = \left(S_1^2/n\right)\left[P_0 + \frac{1}{P_1}(1-f)\left(1+\frac{1}{C_1^2}\right)\right]$$

$$+\left(S_1^2/n^2\right)\left[\left\{(1+f)+\frac{1}{C_1^2}\right\}+\frac{3P_0(1-f)}{P_1^2}\left(1+\frac{(n-1)P_1}{C_1^2(1-f)}\right)\right]$$

$$\simeq \frac{S_1^2}{n}\left[\frac{P_0}{m}+\frac{1}{P_1}(1-f)\left(1+\frac{1}{C_1^2}\right)\right], \text{ (up to the order of } 1/n) \qquad (2.5)$$

where, S_1^2, f are as before, C_1^2 is the square of coefficient of variation of y in the domain of respondents, and $P_1 = N_1/N$, $P_0 = N_0/N$. Here n_1, n_0 have been treated as random variables with $E(n_1) = np_1$ and $E(n_0) = n\,p_0$.

Proof: (i) we have

$$E\left(\bar{y}_*^{(j)}\right) = E_s E_{n_1|s} E_{z|n_{1,s}}\left(\bar{y}_*^{(j)}\right)$$

$$\simeq \bar{Y}_1$$

$$= \overline{Y} + \left(\frac{N_0}{N}\right)\left(\overline{Y}_1 - \overline{Y}_0\right).$$ (For calculation see the Appendix)

thus, under the assumption of $\overline{Y}_1 = \overline{Y}_0$, the result (i) follows and (ii) follows by observing the following fact and after routine calculations,

(ii) $$V(\overline{\overline{y}}_*) = E_{s}\, E_{n_{1|s}}\ V_{z|n_1,s}(\overline{\overline{y}}_*)$$

$$+ E_s\, V_{s_1|s}\ E_{z|s_1,s}\ (\overline{\overline{y}}_*)$$

$$+ V_s\, E_{s_1|s}\ E_{z|s_1,s}\ (\overline{\overline{y}}_*)$$

and

$$Cov\left(\overline{y}_*^{(j)}, \overline{y}_*^{(k)}\right) = E_s\, E_{s_1|s}\, Cov_{z|s_1,s}\ \left(\overline{y}_*^{(j)}, \overline{y}_k^{*}\right)$$

$$+ E_s\, Cov_{s_1|s}\ \left(E_{z|s_1,s}(\overline{y}_*^{(j)}), E_{z|s_1,s}(\overline{y}_*^{k}\right)$$

$$+ Cov\left(E_{s_1|s}\, E_{z|s_1,s}(\overline{y}_*^{(j)},\ E_{s_1|s}\, E_{z|s_1,s}(\overline{y}_*^{(k)}\right)$$

where, suffix *z* stands over imputation. (For calculation see the Appendix)

3. General Cost Model

The optimal solution is always conditional on the appropriate cost-error model. Costs do vary across different activities in the total survey design and are very much dependent on the extent of efforts needed for their execution. Efforts at any stage can be translated into time of operations and finally, total cost can be evaluated with the available knowledge on the rate of cost per unit of time, rates being different for different survey operations. Thus specification of a cost model reduces to modeling of time allocation into different components under the total survey design. One of the major activities lies with the Survey Management Group who are mainly engaged in the survey operation leading to response or non response. It may be noted that time of response for complete or partial information from a respondent depends along with others on the efficiency of an investigator. The efficiency of an investigator need not necessarily be uniform in his whole course of action. In fact, efficiency of an investigator increases with the number of interviews completed (Mahalanobis, 1944).

A general cost model can be specified as follows:

$$C_T = C_0 + C(D) + C(n) + \underbrace{C(n+d)}_{} \tag{3.1}$$

where, C_T = Total Cost:

C_0= Summary of fixed costs, which are independent of sampling design as well as sample size; It primarily includes costs of recruitment of human capital for administration as well as for technical work associated with programming as well as other computer job for data processing. It also includes machine capital. The cost of hardware/software and other equipment charges are also included here. Costs in specifying an association rule δ_{ik} between k^{th} frame units and i^{th} population unit and costs in preparing survey instruments for procuring and reconciliation of the survey results are also included here.

C (*D*) = Summary of fixed costs mainly related to some sampling office work including computation of multipliers, estimators and their variances etc.

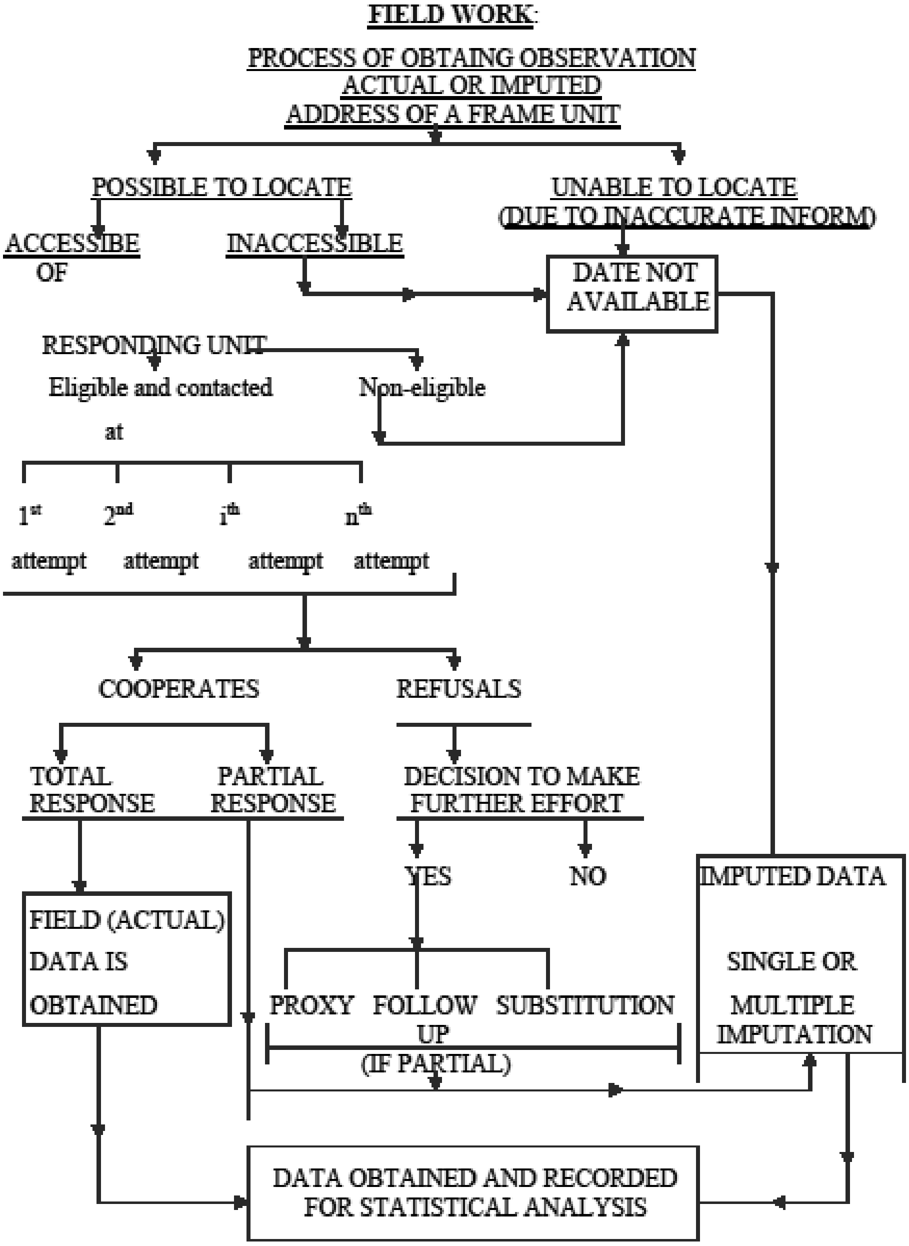
C (*n*) = *n C* = All those costs dependent on the number of sampling units alone, but unaffected by the change of the sampling design; These fixed costs include printing/photocopying of the schedules/questionnaires, coding, editing etc.;

C (*n*,*d*) = This is a **variable cost** dependent on the sample size and on the particular method of data collection adopted by the Survey Management Group;

It may be noted that though a response or a non-response is primarily the outcome of an interactive process between a respondent and an investigator under a given survey condition, such outcomes are dependent not only on interviewers and interviews, but also on all the instruments used in the whole system. The above general model tries to keep an account of all the costs.

The component *C* (n, *d*) which represents costs incurred during the data collection procedure will vary with different interviewers having different levels of efficiency. These efficiencies will further depend on different methods of data collation and collection. The component *C* (*n*, *d*) may require further specification, as it will have a larger share in the total cost. We present the different field conditions through the following schematic diagram 1.

Modeling of C (*n*, *d*) in the presence of no non – response



Since the efficiency of interviewers increases with the number of interviews completed, the relationship between sample size and cost may not be linear, but curvilinear so that cost/unit is a decreasing function of sample size. A possible cost model in the presence of no non-response would be

$$C(n,d) = C(t_r) + (K_2 + K_3 n)\, n \qquad (3.2)$$

where,

$C(t_r)$= the cost incurred in pertaining a training programme to the investigators;

K_2 = Base Cost

K_3= a measure reflecting to decline in cost for interviewer with increasing efficiency.

Costs incurred in all efforts leading to having an effective interview would form a part of the base cost k_2. Such cost arises because of some or all of the following field conditions.

(a) Some of the **dwelling units** may not be possible to be identified due to faulty information in the frame population and/or because of in accessibility due to natural calamities and / or political disturbances;

(b) **Dwelling units**, though accessible, may be found to be vacant;

(c) The **dwelling unit**, though may not be vacant, but may not have an eligible respondent;

(d) The eligible respondent may not be temporarily at home;

(e) When contacted after a number of call backs, the respondent may be turned out to be an '**initial non-respondent"** and efforts would be needed to convert him into a respondent, failure to which he becomes a "**permanent non-respondent**";

(f) A respondent may be contact at the first time, but refusal may take place, and the interviewer may proceed further with an attempt to meet another eligible interviewee.

The above efforts may be viewed as the amount of time need until an interviewee is reached and all the costs should be attributed to the cost of a successful completion of the schedule. This cost may be termed as the **cost of exploration leading to the discovery of an interviewee**. In some cases, these costs may even be zero.

Let

$$\delta_j(i) = \begin{cases} 1, \text{when } j^{th} \text{ investigator finds } i^{th} \text{ respondent ready for cooperation} \\ 0, \text{ otherwise;} \end{cases}$$

When $\delta_j(i) = 1$, then the time passed through the process of investigation may be termed as operating time $\theta_j(i)$ or exploitation time and may further be split into the following components.

(a) **Rapport-time** $\left(R_j(i)\right)$ taken by the j^{th} investigator in pursuing the i^{th} investigator for co-operation;

(b) **Enumeration time** $\left(E_j(i)\right)$ taken by the j^{th} investigator in actually collecting data from the i^{th} investigator;

(c) **Editing time** $\left(ED_j(i)\right)$ taken by the j^{th} investigator in assessing, if the information collected from i^{th} respondent needs to be monitored and re-interviewed;

(d) **Re-interview time** $\left(\mathrm{Re}_j(i)\right)$ is the time of re-interview and reconciliation time;

(e) **Break time** $\left(Br_j(i)\right)$ taken by the j^{th} investigator to depart from the i^{th} investigator;

(f) **Travel time** $\left(Tra._j(i)\right)$ is the time taken by the j^{th} investigator moving after the completion of the i^{th} schedule in search of another interviewee allotted to him.

Thus, operating time:

$$0_j(i) = Ra._j(i) + E_j(i) + Ed_j(i) + \mathrm{Re}_j(i) + Br._j(i) + Tra._j(i).$$

Total operating time taken by the j^{th} investigator would be ,

$$0_j = \sum_{i \in U} \delta_j(i) 0_j(i)$$

and the associated cost would be,

$C_j = 0_j R_j$, R_j being the rate of the j^{th} investigator.

Cost of operation for all the schedules by all the investigators combined would be $C = \sum_{j \in V} C_j$, where V is the set of investigators. In fact the rate of cost combined with the base cost would be a decreasing function of the sample size and takes the form as mentioned in the model specified by (3.2).

The filed-in schedules are finally scrutinized by the supervision staff, and in this process, let $t_K(i,j)$ be the time needed by the k^{th} supervisor for scrutinizing the schedules completed by the j^{th} investigator from the i^{th} respondent $(i = 1,2,\ldots u; j = 1,2,\ldots v$ and $k = 1,2,\ldots,L)$. Therefore, the total time needed by the k^{th} supervisor to supervise all the filled–in schedules allotted to him would be,

$$S(k) = \sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_k (i,j) t_K (i,j)\ ,\ k = 1,2,\ldots,L\ ;$$

where,

$$\lambda_k (i,i) = \begin{cases} 1, \text{ if } (i,j)^{th} \text{ schedule is supervised by the } k \text{ superviser;} \\ 0, \text{otherwise.} \end{cases}$$

and the related cost for the k^{th} supervisor would be

$$C(k) = S(k) \times R(k)\ ,$$

R (k) being the rate of the k^{th} supervisor.

Therefore, the total amount needed in supervision work would be

$$C(S) = \sum_{k=1}^{k} S(k) R(k)\ .$$

Normally, supervisory staff forms the permanent staff and hence these costs are included in the fixed cost in the form of recruiting human capital.

Modeling of *C* (*n*, *d*) is the presence of unit non-response.

The previous model may be reformulated as

$$C = C_0 + C_1 n + \left(K_2 + K_3\ (n - n_0)\right)(n - n_0) + \left(K_2 + C^* m\right) n_0$$

where,

C_0 = fixed cost (including the training cost);

C_1 = cost/unit in preparing schedules;

n_0 = number of non-respondents;

C^* = imputation cost/unit non-response;

m = number of imputations;

Therefore, the cost at expected number of non-respondents would be

$$C = C_0 + \left(C_1 + K_2 + C^* m P_0\right) n + K_3 (1 - P_0)^2 n^2 \qquad (3.3)$$

where, P_0 is the proportion of non-response in the population.

Let $\Phi = V(\bar{\bar{y}}_*) + \lambda \left\{C - C_0 - \left(C_1 + K_2 + C^* m P_0\right)n - K_3(1-P_0)^2 n^2\right\}$

$$= \frac{S_1^2}{n}\left[\frac{1}{P_1} + (1-f)\left(1 + \frac{1}{C_1^2}\right)\right] + \frac{S_1^2 P_0}{nm}$$

$$+ \lambda \left\{C - C_0 - \left(C_1 + K_2 + C^* m P_0\right) n - K_3(1-P_0)^2 n^2\right\}$$

$$\frac{\partial \Phi}{\partial n} = -\frac{S_1^2 A}{n^2} - \frac{S_1^2 P_0}{n^2 m} - \lambda \left[C_1 + K_2 + C^* m P_0 + 2n K_3 (1-P_0)^2\right]$$

where, $A = \frac{1}{p_1} + (1-f)\left(1 + \frac{1}{C_1^2}\right) > 0$

$$\frac{\partial \Phi}{\partial m} = -\frac{S_1^2 P_0}{n m^2} - \lambda\ C^* n\ P_0$$

$$\frac{\partial \Phi}{\partial \lambda} = C - C_0 - \left(C_1 + K_2 + C^* m P_0\right)n - K_3(1-P_0)^2 n^2$$

$$= (C - C_0) - \left(C_1 + K_2 + C^* m P_0\right)n - K_3 P_1^2 n^2$$

$$\frac{\partial \Phi}{\partial m} = 0 \Rightarrow -\frac{S_1^2 P_0}{n\ m^2} = \lambda\ C^* n P_0$$

$$\Rightarrow \qquad \lambda = -\frac{S_1^2 P_0}{n^2\ m^2 C^* P_0}$$

$$\frac{\partial \theta}{\partial n} = -\frac{S_1^2 A}{n^2} - \frac{S_1^2 P_0}{n^2 m} + \frac{S_1^2 P_0}{n^2 m^2 C^* P_0}\lambda \left[C_1 + K_2 + C^* m P_0 + 2n K_3(1-P_0)^2\right]$$

$$= -\frac{S_1^2 A}{n^2} - \frac{S_1^2 P_0}{n^2 m} + \frac{S_1^2 P_0}{n^2 m^2 C^* P_0}\left[\frac{C_1 - C_0 + K_3 + P_1^2 n^2}{n} + 2n K_3 P_1^2\right]$$

$$\left(\text{By } \frac{\partial\Phi}{\partial\lambda} = 0 \Rightarrow C_1 + K_2 + C^* m P_0 = \frac{C - C_0 - K_3 P_1^2 n^2}{n}\right)$$

$$= -\frac{S_1^2 A}{n^2} - \frac{S_1^2 P_0}{n^2 m} + \frac{S_1^2}{n^2 m^2 C^*}\left(\frac{a}{n} + bn\right), \; a = C - C_0 > 0, b = K_3 P_1^2 < 0,$$

as $k_3 < 0$

Thus, by $\frac{\partial\Phi}{\partial n} = 0$, we have,

$$\frac{S_1^2}{n^2 m^2 C^*}\left(\frac{a + bn^2}{n}\right) = \frac{\left(mS_1^2 A + S_1^2 P_0\right)}{n^2 m}$$

or, $S_1^2\left(a + bn^2\right) = \left(m^2 S_1^2 A + mS_1^2 P_0\right) nC^*$

$$n^2 b - C^* m\ (P_0 + m\ A)\ .n + a = 0$$

or, $n^2 b - nD + a = 0$, where $D = mC^*(P_0 + m)$

or, $n = \frac{D \pm \sqrt{D^2 - 4ab}}{2b} = \frac{D \pm \sqrt{D^2 + 4ab}}{2b}$,

Thus, for given value of *m*, $(2 \le m \le 10)$, we shall have different pairs (*n*, *m*) $(2 \le m \le 10)$ and from these choice of (n, m), the optimum pair, say, $(n, m)_0$ can be obtained by comparing the cost at expected number of non-respondents as specified in (3.3) for different pairs (*n*, *m*)

REFERENCES

BABBIE, EARL R. (1973): Survey Research Methods. Belmont, CA, Wadsworth.

BACKSTORM, CHARLES H. and GERALD HURSH-CESAR (1981). Survey Research, 2nd edition, New York, Wiley.

BOWLEY, A.L. (1906): Address to the Economic and Statistics Section of the British Association for the Advancement of Science, York, 1906, J. Roy Statist. SOC. 69, 540—558

---------(1926): Measurement of the Precision attained in Sampling. Bulletin of the International Statistical Institute, 22, 6—62.

BENNET, C.M. and HILL, R.E (1964): A comparison of selected personality characteristics of respondents and non-respondents to a mailed Questionnaire. Journal of Educational Research, 58, No. 4 178—180.

BIRBAUM, Z.W. and MONROE G. SIRKEN (1950): Bias due to non-availability in Sampling Survey. JASA, 45, 98—111.

BAILER, BARBARA A. (1979): Rotation Sampling Biases and their effects on estimates of changes 43rd session of the International Statistical Institute, Manita.

BERGMAN, L.R. HONVE, R. and RAPPA, J. (1978): Why do some people refuse to participate interview surveys? Statistik Tidskrift.

BROOKS, CAMILLA and BARBARA BAILAR (1978): An Error Profile Employment as Measured by current Population Survey Statistical Policy Working Paper 3, office Federal Statistical Policy and Standards. U.S. Department of Commerce.

BANDYOPADHYAY, S. CHAUDHURY, A., GHOSH, J.K. and MAITI, P. (1999): A Draft Proposal for an Enterprise Survey Scheme as a substitute for Economic Census. Indian Statistical Institute, Calcutta.

COCHRAN, W.G. (1979): Sampling Techniques, Wiley Eastern Limited, New Delhi, III edition.

COLE, D (1956): Field Work in Sample Surveys of Household Income and Expenditure, Applied Statistics, Volume 5, 49—61.

COBB, J.M., KING S., and CHEN, E. (1957): Differences between respondents and non-respondents in a morbidity survey involving clinical examination, Journal of Chronic Diseases, 6.

CHEVRY GABRIEL (1949): Control of General Census by means of an area sampling method, JASA, 44, 373—379.

CHAPMAN, DAVID D. and ROGERS CHARLES, E.(1978): Census of Agriculture- Area Sample design and methodology. Proceedings of the American Statistical Association Section on Survey Research Methods, 141—147.

DEMING, W (1960): Sampling Design and Business Research, New York, Wiley.

--------------(1944): On Errors in Surveys" American Sociological Review, 9, 359—369.

-------------(1950): Some Theory of Sampling, John Wiley and Sons, New York.

-------------(1953): On a Probability mechanism to attain an Economic Balance between in resultant error and the bias of non-response . JASA 48, 743—772.

DALENIUS, TORE (1974): The Ends and Means of Total Survey Design; Stockholm, The University of Stockholm.

---------------(1957): Sampling in sweden Contribution to the Methods and Theories of Sample Survey Practice, Stockholm, Almquist and Wicksell.

----------------(1962): Recent Advances in Sample Survey Theory and Methods, AMS, 33, 325—349.

----------------(1977a): Bibliography of non-sampling errors in Surveys. I(A—G), International Statistical Review, 3, 71—89.

----------------(1977b): Bibliography of non-sampling errors in Surveys II(A—Q), International Statistical Review, 45, 181—197.

-----------------(1977c): Bibliography of non-sampling errors in Surveys, III(R—Z), International Statistical Review, 45, 313—317.

DASGUPTA, A and MITRA, S.N. (1958): A Technical Note on Age Grouping. The National Sample Survey No.12, New Delhi.

DHAR, N.R. (1971): A note on non-sampling errors in NSS data. The National Sample Survey Working Paper No. 57/71/1.

DUNN, J.P. and HAWKES, R (1966): Comparison of non-respondents and respondents in a Periodic Health Examination Program to a mailed questionnaire, American Journal of Public Health, 56, 230—236.

DEMAIO, T.Y.(1980): Refusals, who where and why? Public Opinion Quarterly 44.

ERICKSON, W.A. (1967): “Optimal Sample Design with non-response”, JASA, 62, 63—78.

EMRICH, LAWRENCE (1983): “Randomised Response Technique” In William G. Madow and Ingram olkin eds. Incomplete data in Sample Surveys; Volume 2, Theory and Bibliographies, New York, Academic, 73—80.

FELLEGI, IVAN P. (1963): The Evaluation of the Accuracy of Survey Results Some Canadian Experiences. International Statistical Review, 41, 1—14.

---------------(1964): Response Variance and its Estimation, JASA, 59, 1016—1041.

FELLEGI, IVAN and SUNTER, A.B. (1974): Balance between Different Sources of Survey Errors, Some Canadian Experiences, Sankhya, 36, Series C), 119—142.

FERVER, REBERT (1966): Items non-response in a consumer survey, Public Opinion Quarterly, 12, 669—676.

FORD, BARRY L. (1976): Missing Data Procedures, A Comparative Study, American Statistical Association, Proceedings of the Social Statistics Section 1976, Pt. 1, 326—329.

GHOSH, J.K. and MAITI, PULAKESH (2003): The Indian Statistical System at cross roads an appraisal of Past, Present and Future, To be presented at the IMS meet during 2—3 January – 2004.

GHOSH, A (1953): Accuracy of Family Budget Data with reference to period of re-call, Calcutta Statistical Association Bulletin, 5, 16—23.

GOWER, A.R. (1979): Characteristics of non-respondents in the Labour Force Survey, Statistics Canada.

GROVES, ROBERT, M. and KAHN ROBERT LOUIS (1979): Surveys by Telephone, A national comparison with personal interview, New York; Academic.

GRAY, P. and GEE, F.E.N. (1972): A Quality check on the 1966 ten percent sample census of England Wales, office of the population census and surveys, London.

GHOSH, J.K., MAITI, P. MUKHOPADHYAY, A.C., PAL, M.P (1977): Stochastic Modeling and Forecasting of Discovery, Reserve and Production of Hydrocarbon-with an application, Sankhya, Series B, 59, pt. 3, 288—312.

HANSEN, M.H., MADOW WILLIAM G., and TEPPING B.J. (1983): An Evaluation of Model dependent and Probability Sampling inference in Sample Surveys, JASA, 78, 776—807.

HANSE, M.H., HURWITZ WILLIAM N. (1946): The Problem of non-response in Sample Survey, JASA, 41, 516—529.

---------and NISSELSON, H., STEINBERG, J. (1955): The redesign of the current population survey, JASA, 50, 701—719.

----------JUBINE, TOMAS B. (1963): The use of imperfect lists for Probability Sampling at U.S. Bureau of Census, Bulletin of the International Statistical Institute, 40(1), 497—517.

--------and PRITZKER, LENON (1964): The Estimation and interpretation of Gross differences and the simple response variance. In C.R. Rao with D.B. Lahiri, K—P.

NAIR, P. PANT and S.S. SHRIKHANDE eds. Contributions to Statistics Presented to Professor P.C. Mahalanobis on the occasion of his 70th birth day

Oxford, England, Pergaman, Calcutta Statistical Publishing Society, 111—136.

-----------and BERSHAD, MAX A. (1961): Measurement errors in censuses and surveys, Bulletin of the International Statistical Institute, 38, 359—374.

-----------MARKS, ELIS MAULDIN, PARKER W. (1951): Response Errors in Surveys, JASA, 46, 147—190.

-------------(1976): Some Important Events in the Historical Development of Sample Surveys in Donald Bruce Owen ed., on the History of Statistics and Probability, Statistics Text Books and Monographs, Volume 17, New York Dekker, 73—102.

HURSCGBERG, DAVID FREDERICK, J. SCHEUREN and YUSKAVAGE ROBERT (1977): the impact on Personal and Family income of adjusting the current population survey for under coverage, Proceedings of the Social Statistics Section, American Statistical Association, 70—80.

HUBBACK, J.A. (1927): Sampling for rice yields in Bihar and Orissa, Imp. Agr. Res. Inst. Bulletin, Pusha (reprinted in Sankhya (1946), 7, 282—294)

HALDEN, J.B.S. (1957): The Syadvada System of Prediction, Sankhay 18, 195—2000.

HACKING, J. (1965): Lobgic of Statistical Inference Cambridge University Press.

HOINVILLE, GERALD and ROBERT JOELL (1978): Survey Research Practic , London, Heinemann.

JESSEN, RAYMUND J. (1978): Statistical Survey Techniques, New York, Wiley.

KIAWER, A. (1895): Observations et experiences concernant des denombrements representatives, Bull. Int. Statist. Inst. 9, 176—183.

KRUSKEL, WILLIAM and FREDERICK MOSTELLER (1980): Representative Sampling, IV, the History of the concept in Statistics, 1895—1939, International Statistical Review, 48, 169—195.

KISH, L. (1965): Survey Sampling Wiley and Sons, New York.

------- and HESS I. (1958): on non-coverage of sampling dwellings, JASA, 54, 509—524.

KALTON, GRAHAM and DANIEL KASPRZYK (1982): Imputing of missing survey Response, American Statistical Association 1982, Proceedings of the Section on Survey Research Methods, 22—31.

KOOP, J.C. (1974): Notes for a unified theory of estimation for sample surveys taking into account response errors, Metrika, 21, 19—39.

KALTON, GRAHAM (1983): Compensating for missing survey data Research Report Series, Ann. Arbor M1, Institute for Social Research, University of Michigan.

KENDAL, MAURICE GEORGE and WILLIAM R. BUCKLAND (1960): A Dictionary of Statistical Terms, 2nd edition, London, Oliver and Boyd.

LUBIN, B. LEVITT, E. and ZUCKERMAN, M. (1962): Some personality differences between respondents and non-respondents in a survey questionnaire, Journal of Consulting Psychology, 26—192.

LUNDBERG, G.A. and LARSEN, O.A. (1949): Characteristics of Hard-to-reach individuals in field surveys, Public Opinion Quarterly, 13, 487—494.

LYBERG, L. and RAPAPORT, E. (1979): Unpublished non-response problems at the national central Bureau of Statistics, Sweden.

LITTLE, RODERICK J.A. (1982): Models for non-response in Sample Surveys, JASA, 77, 237—250.

------------------------(1983): Super Population models for non-response, Part IV. In William G. Madow and Ingram Olkin eds. Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies, New York, Academic, 337—413.

LESSLER, J.T. (1974): A double sampling scheme model for eliminating measurement process bias and estimating measurement errors in surveys, Institute of Statistics Mimeo Series No. 949, University of North Carolina, New Chapel.

-------------------(1980): Errors associated with the frames, Proceedings of the American Statistical Association Section on Survey Research Methods, 125—130.

MADOW, W.G. NISSELSON, HAROLD and OLKIN, INGRAM (1983): Incomplete data on Sample Survey, Volume 1, Report and Case studies; New York, Academic.

MC. NEIL, JOHN M. (1981): Factors affecting the 1980 census content and the effort to develop a post census disability survey. Presented at the annual meeting of the American Public Health Association.

MAHALANOBIS, P.C. and LAHIRI, D.B. (1961): Analysis of errors in censuses and surveys, Bulletin of the International Statistical Institute, 38(2), 359—374.

----------------and SEN, S.B. (1954): On some aspects of the Indian National Sample Survey, Bulletin of the International Statistical Institute, 34, pt. 2.

MAHALANOBIS, P.C.(1944): On Large Scale Sample Surveys, Philosophical Transactions of Royal Society, 231—(B), 329—451.

-------------------(1946): Recent Experiments in Statistical Sampling in the Indian Statistical Institute.

---------------------(1941): A Sample Survey of the Acre-age under jute in Bengal, 4, 511—30.

---------------------(1954): The Foundations of Statistics, Dialectica, 8, 95—111 (reprinted in Sankhya 18, 183—194)

MAITI, P (1983): unpublished Ph.D. Thesis entitled Some Contributions to the Sampling Theory using auxiliary information" submitted to the Indian Statistical Institute, Calcutta.

----------, PAL, M. and SINHA B.K. (1992): Estimating unknown Dimensions of a Binary matrix with application to the estimation of the size of a mobile population. Statistics and probability, 220—233.

MOSER, CLAYS ADOLF and GRAHAM KALTON (1972): Survey methods in Social investigation , 2nd edition, New York, Basic Books.

MOONEY, H. (1962): On Mahalanobis' contributions to the development of sample survey theory and method in C.R. Rao et al. (eds) contributions of statistics, Pergamon Press.

------------------- (1967): Sampling Theory and Methods, Statistical Publishing Society, Calcutta.

NEYMAN JERZY (1934): On the two different aspects of the representative method, the method of stratified sampling and the method of purposive selection, J. Roy, Statist. SOC. 97, 5589625.

NETER, J. and WAKSBERG, J. (1965): Response errors in collection of Expenditures data by household interview. An Experimental Study Technical Report No. 11 U.S. Bureau of the Census.

NEWMAN, S. (1962): Difference between early and late respondents in a mailed survey, Journal of Advertising Research, volume 2, 37—39.

OGNIBENE, P. Traits affecting questionnaire response, Journal of Advertising Research Volume 10, 18—20.

PAN, J.S. (1951): Social Characteristics of respondents and non-respondents in a questionnaire study of later maturity, Journal of Applied Psychology, 35, 780—781.

POLITZ, A.N. and SIMMONS, W.R. (1949): An attempt to get Not-at Homes into the sample without call-backs, JASA, 44, 9—31.

PLATEK, R. (1977): Some factors affecting non-response, Survey Methodology, 3.

PLAN, V.T. (1978): A Critical appraisal of household surveys in Malaysia Multipurpose household survey in developing Countries, Development Centre, OECD, Paris.

REUSS, C.F. (1943): Differences between persons responding and not responding to mail questionnaires, American Sociological Review, 8, 433—438.

RAO, V.R. and SASTRY, N.S. (1975): Evolution of a total survey design, The Indian Experience, Invited paper presented to the International Association of Survey Statisticians Warsaw.

Report of Research Projects

(1975—76): Cost Benefit Analysis of Rural Electrification, Project Leader Professor J. Roy, Computer Science Unit, ISI, Calcutta.

(1977—78): Calcutta Urban Poverty Survey, Project Leader Professor J. Roy, Computer Science Unit, ISI, Calcutta.

(1988—89): A Survey on Domestic Tourists in Orissa, Project Co-ordinator, P. Maiti, ISI, Calcutta.

(1994—95): An Enquiry into the Quality of Life in five communities in selected districts of Rural West Bengal, Project Co-ordinator, P. Maiti, ISI, Calcutta.

(1995): Community attitudes and Preferences pertaining to cemetery and cremated related issues in the East Rand in the Republic of South Africa, CENSIAT, HSRC, Pretoria, South Africa, Principal Statistician – P. Maiti.

(1995): Survey of family and Community life in the Selected Communities of the cape Peninsula of the Republic of South Africa, CENSTAT< HSRC, Principal Statistician – P. Maiti.

(1995): the Socio-economic demographic and cultural pattern of the female labour force participation in the North West and the Cape; CENSTAT, HSRC South Africa, Principal Statistician – P. Maiti.

(1996): Stanza-Bopape Project; CENSTAT, HSRC, South Africa, Principal Statistician – P. Maiti.

(1998): Mid. Term Review of IPP—VIII in Calcutta Metropolitan Area, ISI, Calcutta, Survey Statistician – P. Maiti.

(2001, August): National Statistics Commission , Government of India.

ROSHWALB, ALAN (1982): Respondent Selection Procedures within Households, American Statistical Association 1982 Proceedings of the section on Survey Research Methods, 93—98.

RUBIN DONALD B. (1983): Conceptual issues in the presence of non-responses, In William G. Madow and Ingram Olkin eds. Incomplete data in Sample Surveys 2, Theory and Bibliographies, New York, Academic, 123—142.

-------------------(1977): formalizing Subjective notions about the effect of non-respondents in Sample Surveys", JASA, 72, 538—543.

---------------(1978): Multiple imputations in Sample Surveys – A Phenomenological Bayesian Approach to non-response, American Statistical Association 1978 proceedings of the Section on Survey Research Methods, 20—28.

---------------(1987): Multi imputation for non-response in Surveys, New York, Wiley.

RIZVI, M. HASEEB (1983): Hot-Deck Procedures Imputation in William G. Madow and Ingram Olkin eds., Incomplete Data in Sample surveys, 3, Proceedings of the symposium, New York, Academic, 351—352.

SARMA, V.R.R., RAO, G. D., AMBE, V.N. (1980): Non-responses in household surveys of National Sample Survey.

STEPHEN, FREDERICK F. (1948): History of the uses of Modern Sampling Procedures, JASA, 43, 12—39.

SMITH, T.M.F. (1976): The Foundations of Survey Sampling, A Review, JRSS, 139A, 183—195.

SARNDAL, C.E., SWENSSON, B. and WRETMAN, J. (1992): Model Associated Survey Sampling , Springer Verlag, New York, Inc.

SCOTT CHRISTOPHER (1973): Experiments on recall error in African Budget Surveys, paper presented to the International Association of Survey Statisticians, Viena.

SHAH, NASRA M. (1981): Data from Tables used in the paper presented at Weekly Seminar of East West Population Institute, October 28, Honolulu.

SUDMAN, SEYMOUR (1976): Applied Sampling, New York, Academic.

SUCHMAN, EDWARD A. (1962): An analysis of bias in Survey Research, Public Opinion Quarterly, 26, 102—111.

SCHEAFFER, RICHARD, L. MENDENHALL, WILLIAM and OTT LYAN (1979): Elementary Survey Sampling, 2nd edition, North Scituate, MA: Duxbury Press.

SZAMEITAT, KLEUS and SCHAFFER, KARL AUGUST (1963): Imperfect Frames in Statistics and the consequences for their use in sampling, Bulletin of the International Statistical Institute, 40, 517—544.

SINGH, BAHADUR, SEDRANSK, JOSEPH (1978): A two phase sampling design for estimating the finite population mean when there is non-response, In N. Krishnan Namboodiri ed. Survey Sampling and measurement, New York, Academic, 143—155.

THOMPSON, IB and SIRING, E. (unpublished): On the causes and effect of non-response, Norwegian Experiences, Central Bureau of Statistics, Norway.

TUYGAN, KUTHAN and CAVADOR, TEVFIK (1975): Comparison of self and presale responses related to children of ever married women. In laboratories for population Statistics Scientific Report Series No. 17, 22—28.

TURNER, ANTHONY G., WALTMEN, HENRY. F, FAY ROBERT and CARLSON BEVERLY (1977): Sample Survey Design in developing Countries – three illustrations of methodology, Bulletin of the International Statistical Institute.

U.S. Bureau of the Census (1974): Standards for the discussion and presentation of errors in data, Technical Report No. 32.

--------------------(1976): An overview of population and housing census evaluation programmes conducted at the U.S. Bureau of Census, Census Advisory Committee of the American Statistical Association.

VERMA, VIJAY (1980): Sampling for national fertility surveys, World fertility survey conference, London.

WARWICK, DONALD P. and CHARTES A. LININGER (1975): The Sample Survey, Theory and Practice, New York, Mc. Graw Hill.

WOLTMAN, HENRY and BUSHERY, JOHN (1975): A panel bias study to the national crime survey. Proceedings of the Social Statistics Section. American Statistical Association.

WILLIAM, W.H. and MALLOWS, C.L. (1970): Systematic biases in panel surveys JASA, 65, 1338—1349.

WARNER, STANLEY L. (1965): Randomised Response, A Survey Technique for eliminating Evasive answer bias, JASA, 60, 63—69.

ZARKOVICH. S.S. (1966): Quality of Statistical data, Rome; FAO of the United Nations.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 259—271

FACTOR TYPE ESTIMATOR UNDER POST-STRATIFICATION IN SAMPLE SURVEYS

Diwakar Shukla, Jayant Dubey[1], Manish Trivedi[2]

ABSTRACT

Factor-type estimator, proposed by Singh and Shukla (1987), is useful in estimating the mean of a finite population. With some nice properties, it reduces the bias even at the optimal-level of MSE and reduces MSE when it is unbiased. The post stratification is effective in a set-up of stratified sampling when sizes of stratum are known but not the list of units (frame) of each strata. This is a real situation and most likely situation for survey practitioners and they are advised to draw a random sample by SRSWOR and stratify later the sample units according to strata which is the crux of post stratification technique. Sukhatme (1984) advocates that post-stratification procedure is as precise as the stratified sampling under proportional allocation if the sample size is large enough. This paper extends the properties of Factor-type estimator under the set-up and applicable situations for the post stratification.

1. Introduction

Let a finite population *U* consists of units labeled $\left(U_1, U_2, U_3 ....U_N\right)$ in some order. With each unit $U_a\left(a = 1, 2, 3, ...N\right)$ a study variable *Y* and an auxiliary variable *X* is associated. The population is stratified into *K* strata each of size $N_i\left(i = 1, 2, 3, ... K\right)$ with $\sum_{i=1}^{k} W_i = \sum_{i=1}^{k}\left[N_i / N\right] = 1$. The symbol Y_{ij} and X_{ij} denotes j^{th} value $\left(j = 1,2,3,.....N_i\right)$ of the i^{th} strata in the population and $\bar{Y}_i = \left(N_i\right)^{-1}\left[\sum_{i=1}^{N_i} Y_{ij}\right], \bar{X}_i = \left(N_i\right)^{-1}\left[\sum_{i=1}^{N_i} X_{ij}\right]$ are respective means

[1] Faculty Member, ICFAI National College, Pumma Sahu Complex, 5-Civil Lines, Sagar, M.P, e-mail: dubey.jayant@rediffmail.com

[2] Lecturer, Deptt of Applied Mathematics B.I.T., Mesra, Ranchi, Jharkhand

of the same strata for *Y* and *X*. Moreover, $\overline{Y}=\sum_{i=1}^{k} W_{i\overline{Y}_i}, \overline{X}=\sum_{i=1}^{k} W_{i\overline{X}_i}$ are population means and S_i^2, S^2 are population means squares defined in usual fashion. A sample of size *n* is drawn by SRSWOR and post stratified according to *K* stratums such that n_i represents $N_i\left(\sum_{i=1}^{k} n_i = n\right)$. Let y_{ij}, x_{ij} be the *j*th sample value from the *i*th strata with respective means $\overline{y}_i = (n_i)^{-1}\sum_{i=1}^{k} y_{ij}$, $\overline{x}_i = (n_i)^{-1}\sum_{i=1}^{k} x_{ij}$.

In order to estimate population means $\overline{Y}$ under SRSWOR set-up $\left(N, n, \overline{y}, \overline{x}, \overline{X}\right)$ Singh and Shukla (1987) have proposed a Factor-Type (F-T) estimator of the form

$$\overline{y}_{FT} = \overline{y}\left[\frac{(A+C)\overline{X} + fB\overline{x}}{(A+fB+C)\overline{X} + C\overline{x}}\right]$$

where $A=(d-1)(d-2); B=(d-1)(d-4); C=(d-2)(d-3)(d-4);$

$$f = n/N, 0 \le d < \infty$$

This estimator was observed more efficient than the usual ratio and product estimators. Moreover, it was having a property of control over bias too even at the optimum level of mean square error. Singh, Shukla and Singh (1991) have extended the applicability of $\overline{y}_{FT}$ to the case of negative correlation between *X* and *Y*. In another contribution, Singh and Shukla (1993) have modified the structure of the Factor-Type estimator by strengthening the sample mean $\overline{x}$ using the additional weight. The prerequisite for the stratified sampling is prior knowledge of (a) strata size and (b) frame of each stratum. The (a) is easy to manage but nothing sure is about (b), for example, a telephone directory could provide a frame of the whole population of a city and by some other sources, it is observed that x_1% are low-income, x_2% are middle income and x_3% are high income group subscribers of phones. Now, this does not provide the frame of each 'economic-strata' though the size. In such situations, the stratified sampling fails and survey practitioners are advised to use the post-stratification technique. As per Sukhatme (1984) the post-stratification is as precise as the stratified samplings with proportional allocation if the sample size is sufficiently large. This motive for the wider application of post-stratification because of being closer to the real life situation. Some useful contributions, in this area, are by Agrawal and Panda (1993 & 1995), Holt and Smith (1979), Jagers (1985, 1986), Smith (1991).

This paper presents an application of Factor-Type (F-T) estimator in the setup of post-stratification $\left(N = \sum_{i=1}^{k} N_i , n = \sum_{i=1}^{k} n_i , \bar{y}_i , \bar{x}_i , \bar{X}_i \right)$ for estimating the population mean.

2. Proposed estimator

Following Singh and Shukla (1987) the proposed Factor-Type estimator under post-stratification for estimating $\bar{Y}$ is:

$$\bar{y}_{PFT} = \sum_{i=1}^{k} W_i \bar{y}_i \left[\frac{(A_i + C_i)\bar{X}_i + fB_i \bar{x}_i}{(A_i + fB_i)\bar{X}_i + C_i \bar{x}_i} \right]$$

Where

$$A_i = (d_i - 1)(d_i - 2); B_i = (d_i - 1)(d_i - 4); C_i = (d_i - 2)(d_i - 3)(d_i - 4);$$

$$f = n/N, 0 \le d < \infty$$

2.1.Special cases

(a) At $d_i = 1, (\bar{y}_{PFT})_1 = \sum_{i=1}^{k} W_i \bar{y}_i (X_i / x_i) = \bar{y}_{PR}$

(b) At $d_i = 2, (\bar{y}_{PFT})_2 = \sum_{i=1}^{k} W_i \bar{y}_i (x_i / X_i) = \bar{y}_{PP}$

(c) At $d_i = 3, (\bar{y}_{PFT})_3 = \sum_{i=1}^{k} W_i \bar{y}_i \left(\frac{NX_i - n\bar{x}_i}{(N - n)\bar{x}_i} \right) = \bar{y}_{PST}$

(d) At $d_i = 4, (\bar{y}_{PFT})_4 = \sum_{i=1}^{k} W_i \bar{y}_i = \bar{y}_{PS}$

The estimator $\bar{y}_{PR}$ is a usual Ratio estimator, $\bar{y}_{PP}$ is a product estimator, $\bar{y}_{PST}$ is of the type Srivenketaramana & Tracy (1980) and $\bar{y}_P$ is a usual sample mean estimator when taken into the set-up of post-stratification.

Note 2.1: The proposed estimator looks like a class of estimators containing some well known estimators $\bar{y}_{PR}$, $\bar{y}_{PP}$, $\bar{y}_{PST}$ and $\bar{y}_{PS}$ as special cases.

3. Bias of estimator

We defined symbol $E\{(.)/n_i\}$, $E\{(.)/n_i\}^2$ and $V\{(.)/n_i\}$ as conditional expectation and variance given n_i.

Assume the quantity $\left|(fB_i)/(A_i+fB_i+C_i)\right|<1$ for all $d_i \geq 0$ and for large n which is justified too.

3.1.Setting approximations

Let $\bar{y}_i=\bar{Y}(1+e_{i1});\bar{x}_i=\bar{X}_i(1+e_{i2})$
such that $E\{(e_{i1})/n_i\}=E\{(e_{i2})/n_i\}=0$

$$\text{(i)}\quad E\{(e_{i1}^2)/n_i\}=V\{(e_{i1}^2)/n_i\}=[(1/n_i)-(1/N_i)](S_{iy}^2/\bar{Y}_i^2)$$

$$\text{(ii)}\quad E\{(e_{i2}^2)/n_i\}=V\{(e_{i2}^2)/n_i\}=[(1/n_i)-(1/N_i)](S_{ix}^2/\bar{X}_i^2)$$

$$\text{(iii)}\quad E\{(e_{i1}.e_{i2})/n_i\}=Cov\{(e_{i1}.e_{i2})/n_i\}=[(1/n_i)-(1/N_i)][(S_{ixy}^2/\bar{Y}_i^2.\bar{X}_i)]$$

$$\text{(iv)}\quad E\{(e_{i1}.e_{i2})/n_i\}=Cov\{(e_{i1}.e_{i2})/n_i\}=0 \text{ for } i\neq i$$

3.2.Some standard results and symbols

$$\text{(a)}\quad E[1/n_i]=\left[\frac{1}{nW_i}+\frac{(N-n)(1-W_i)}{(N-1)n^2W_i^2}\right]$$

$$\text{(b)}\quad C_{iY}=\frac{S_{iY}}{\bar{Y}_i};C_{iX}=\frac{S_{iX}}{\bar{X}_i};C_{iXY}=\frac{S_{iXY}}{\bar{Y}_i.\bar{X}_i}=\frac{\rho S_{iX}S_{iY}}{\bar{Y}_i.\bar{X}_i}=\rho_i C_{iX}.C_{iY}$$

$$\text{(c)}\quad V_i=\rho_i\left[\frac{C_{Yi}}{C_{Xi}}\right]$$

where ρ_i being the correlation coefficient between X and Y of the i^{th} strata.

Remark 3.1: As according to Reddy (1978), the quantity $V = \rho\left(C_Y / C_X\right)$ is a stable quantity over the span of time and could be guessed using the past experience or pilot surveys. We assume herein that the $V_i = \rho_i\left(C_{iY} / C_{iX}\right)$ is known for every i^{th} strata.

THEOREM 3.1: The expected value of $\bar{y}_{PFT}$, under above stated large sample approximation is

$$E\left[\bar{y}_{PFT}\right] = \bar{Y} + \sum_{i=1}^{k} P_i \bar{Y}_i \theta_i C_{iX}\left\{D_i C_{iX} - \rho_i C_{iY}\right\}$$

where

$$P_i = \left[\frac{\left(fB_i - C_i\right)}{\left(A_i + fB_i + C_i\right)}\right] \quad ; \qquad D_i = \left[\frac{C_i}{\left(A_i + fB_i + C_i\right)}\right]$$

$$\theta_i = \left[\frac{NW_i M + (N - n)}{(N - 1)n^2 W_i}\right] \; ; \; M = \left[(n+1)^2 - n^2 N - (n+2)\right]$$

Proof: $E\left[\bar{y}_{PFT}\right] = E\left[E\left\{\left(\bar{y}_{PFT}\right)/n_i\right\}\right]$

$$= E\left[E\left\{\sum_{i=1}^{k} W_i \bar{y}_i \left\{\frac{\left(A_i + fB_i\right)\bar{X}_i + C_i \bar{x}_i}{\left(A_i + C_i\right)\bar{X}_i + fB_i \bar{x}_i}\right\}\right\} \Big/ n_i\right]$$

$$= E\left[E\left\{\sum_{i=1}^{k} W_i \bar{y}_i \left(1 + e_{i1}\right)\left\{\frac{\left(A_i + fB_i\right)\bar{X}_i + C_i \bar{x}_i}{\left(A_i + C_i\right)\bar{X}_i + fB_i \bar{x}_i\left(1 + e_{i2}\right)}\right\}\right\} \Big/ n_i\right]$$

$$= E\left[E\left\{\sum_{i=1}^{k} W_i \bar{y}_i \left(1 + e_{i1}\right)\left\{1 + D_i e_{i2}\right\}\left\{1 + \left(P_i + D_i\right)e_{i2}\right\}^{-1}\right\} \Big/ n_i\right]$$

Expanding binomially, using $\left|\left(P_i + D_i\right)e_{i2}\right| < 1$, we get

$$E\left[\bar{y}_{PFT}\right] = E\left[E\left\{\sum_{i=1}^{k} W_i \bar{Y}_i \left\{1 - \left(P_i + D_i\right)\left(e_{i2}\right) + \left(P_i + D_i\right)^2\left(e_{i2}^2\right) + D_i\left(e_{i2}\right) - \left(P_i + D_i\right)D_i\left(e_{i2}^2\right) + e_{i1}\right.\right.\right.$$

$$\left.\left.\left. - \left(P_i + D_i\right)\left(e_{i1}.e_{i2}\right) + \left(P_i + D_i\right)^2\left(e_{i1}.e_{i2}\right)^2 + D_i\left(e_{i1}.e_{i2}\right) - \left(P_i + D_i\right)D_i\left(e_{i1}.e_{i2}\right)^2 + \ldots\ldots\ldots\infty\right\}\right\}/n_i\right]$$

$$= E\left[\sum_{i=1}^{k} W_i \bar{Y}_i \left\{1 + \left(P_i + D_i\right)^2 E\left\{\left(e_{i2}^2\right)/n_i\right\} - \left(P_i + D_i\right)D_i E\left\{\left(e_{i2}^2\right)/n_i\right\} - \left(P_i + D_i\right)E\left(e_{i1}.e_{i2}\right)/n_i\right\}\right.$$

$$+\ D_i E\{(e_{i1}.e_{i2})/n_i\}\}]$$

Above obtained using approximations and ignoring the terms $E[e_{i1}^{r}.e_{i2}^{s}]$ when $r+s\geq 2\ (r,s=0,1,2,3,......)$. Further use of approximation provides:

$$E[\bar{y}_{PFT}]=\sum_{i=1}^{k}W_i\bar{Y}_i\{1+\{E(1/n_i)-(1/N_i)\}P_iC_{iX}\{D_iC_{iX}-\rho_iC_{iY}\}\}$$

$$=\ \bar{Y}+\sum_{i=1}^{k}P_i\theta_i\bar{Y}_iC_{iX}\{D_iC_{iX}-\rho_iC_{iY}\}$$

3.3.Special results

$$\text{At } d_i=1, E[\bar{y}_{PR}]=\bar{Y}-\sum_{i=1}^{k}\theta_i\bar{Y}_iC_{iX}(C_{iX}-P_iC_{iY})$$

$$\text{At } d_i=2, E[\bar{y}_{PP}]=\bar{Y}+\sum_{i=1}^{k}\theta_i\bar{Y}_i(\rho_iC_{iX}C_{iY})$$

$$\text{At } d_i=3, E[\bar{y}_{PST}]=\bar{Y}+\{f(1-f)^{-1}\}\sum_{i=1}^{k}\theta_i\bar{Y}_iC_{iX}[\{f/(1-f)\}C_{iX}-\rho_iC_{iY}]$$

$$\text{At } d_i=4, E[\bar{y}_{PS}]=\bar{Y}$$

4. Unbiasedness condition

Obviously, d_i = 4 provides the usual unbiased poststratifed estimator. Moreover, $\bar{y}_{PFT}$ will be unbiased if $D_i=V_i=(\rho_iC_{iY}/C_{iX})$ which further reduce into equation

$$A_iV_i+fV_iB_i+C_i(V_i-1)=0 \tag{4.1}$$

Remark 4.1: Since we assume V$_i$ to be the equation (4.1) produces K different equations, each having cubic power in d_i and there shall be maximum of 3K values of $d_i^{\,s}$ making the estimator unbiased. One can use the selection criteria as under:

“For a given i, among maximum of three available values of d_i, choose that having the least mean square error.”

Remark 4.2: The estimator, at the level of maintaining unbiasedness, reduces the mean square error also as revealed from Remark 4.1.

5. Mean square error

THEOREM 5.1: The M.S.E. of $\bar{y}_{PFT}$ could be

$$M(\bar{y}_{PFT}) = \sum_{i=1}^{k} \bar{Y}_i^2 .\theta_i^{'} \left\{C_{iY}^2 + P_i^2 C_{iX}^2 - 2P_i C_{iXY}\right\}$$

where $\theta_i^{'} = \left[\dfrac{NW_i M + (N-n)}{(N-1)n^2}\right]$

Proof: Consider the terms up to order $O(n^{-2})$ only, we have

$$M(\bar{y}_{PFT}) = E\left[E\left\{\left(\bar{y}_{PFT} - \bar{Y}^2\right)/n_i\right\}\right]$$

$$= E\left[E\left\{\sum_{i=1}^{k} W_i \bar{Y}_i (e_{i1} - P_i e_{i2})\right\}^2 \Big/ n_i\right]$$

$$= E\left[\left\{\sum_{i=1}^{k} W_i^2 \bar{Y}_i^2 E\left\{\left(e_{i1}^2\right)/n_i\right\} + \sum_{i \neq i'}^{k} \sum_{=1}^{k} W_i W_j \bar{Y}_i \bar{Y}_j E\left\{\left(e_{1i} e_{1j}\right)/n_i, n_j\right\}\right\}\right.$$

$$+ \sum_{i=1}^{k} W_i^2 \bar{Y}_i^2 P_i^2 E\left\{\left(e_{2i}^2\right)/n_i\right\} + \sum_{i \neq i'}^{k} \sum_{=1}^{k} W_i W_j \bar{Y}_i \bar{Y}_j P_i P_j E\left\{\left(e_{2i}.e_{2j}\right)/n_i, n_j\right\}$$

$$\left. - 2\sum_{i \neq i'}^{k} \sum_{=1}^{k} W_i \bar{Y}_i P_i E\left\{\left(e_{1i} e_{2i}\right)/n_i\right\} - 2\sum_{i \neq i'}^{k} \sum_{=1}^{k} W_i \bar{Y}_i P_i W_j \bar{Y}_j P_j E\left\{\left(e_{2i} e_{2j}\right)/n_i, n_j\right\}\right]$$

$$= E\left[\sum_{i=1}^{k} W_i^2 \bar{Y}_i^2 \left[E\left\{\left(e_{i1}^2\right)/n_i\right\} + P_i^2 E\left\{\left(e_{i2}^2\right)/n_i\right\} - 2P_i E\left\{\left(e_{i1}.e_{i2}\right)/n_i\right\}\right]\right]$$

$$= \left[\sum_{i=1}^{k} W_i^2 \bar{Y}_i^2 \left[E\left\{(1/n_i) - (1/N_i)\right\}\left\{C_{iY}^2 + P_i^2 C_{iX}^2 - 2P_i C_{iXY}\right\}\right]\right]$$

$$= \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' \left\{ C_{iY}^2 + P_i^2 C_{iX}^2 - 2P_i C_{iXY} \right\}$$

Special cases

At $d_i = 1, M\left[\left(\bar{y}_{PFT}\right)_1\right] = M\left(\bar{y}_{PR}\right) = \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' \left\{ C_{iY}^2 + C_{iX}^2 - 2\rho_i C_{iX} C_{iY} \right\}$

At $d_i = 2, M\left[\left(\bar{y}_{PFT}\right)_2\right] = M\left(\bar{y}_{PP}\right) = \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' \left\{ C_{iY}^2 + C_{iX}^2 - 2\rho_i C_{iX} C_{iY} \right\}$

At

$d_i = 3, M\left[\left(\bar{y}_{PFT}\right)_3\right] = M\left(\bar{y}_{PS}\right) = \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' \left\{ C_{iY}^2 + \left(f/(1-f)\right)^2 C_{iX}^2 + \left(2f/(1-f)\right)\rho_i C_{iX} C_{iY} \right\}$

At $d_i = 4, M\left[\left(\bar{y}_{PFT}\right)_4\right] = M\left(\bar{y}_{PS}\right) = \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' C_{iY}^2$.

Remark 5.1: The optimum M.S.E. occurs at $P_i = V_i$ with the expression

$$M\left[\left(\bar{y}_{PFT}\right)_{opt.}\right] = \sum_{i=1}^{k} \bar{Y}_i^2 \theta_i' C_{iY}^2 \left(1 - \rho_i^2\right)$$

Proof: On differentiating $M\left(\bar{y}_{PFT}\right)$ with respect to P_i and equating to zero, the condition $P_i = V_i$ could be obtained and substituting this in the expression of M.S.E., the optimum expression is acquired.

Remark 5.2: The optimum M.S.E. is equivalent to M.S.E. of the linear regression estimator $\bar{y}_{1r}$ when expressed in the set of post-stratification.

Remark 5.3: Using Reddy (1978)**,** since *f* and V_i's are known for all *i*, the equation $P_i = V_i$ could be solved for d_i. This would provide utmost 3*K* values of d_i have in order to achieve the optimum level of M.S.E. For a given i, the equation $P_i = V_i$ generates a cubic equation with utmost three values of d_i in order to attain the optimum level of M.S.E. The selection criteria for best of them shall be as under: "For a given *i*, among three d_i, values, choose we that having the least $\left|\text{Bias}\left(\bar{y}_{PFT}\right)\right|$"

Remark 5.4: In view to above, the estimator $\left(\bar{y}_{PFT}\right)$ reveals having the control over bias also along with obtaining the optimum level of M.S.E.

Remark 5.5: For the choice of optimum values of d_i, the following equation needs to be solved for every i, for a known pair of(V_i, f):

$$A_i V_i + fB_i(1+V_i)+C_i(V_i-1)=0$$

(a) When $V_i < 1$, the equation has all the three real and positive roots.
(b) When $V_i = 1$, the equation has only two real and positive roots.
(c) When $V_i > 1$, the equation has two positive real and one negative real roots which is ignored since $d_i \geq 0$.

Note 5.1: For an easy and quick selection of d_i - values, a table 6.3 is prepared and attached at the end. One can further extend this table as per requirement.

6. Numerical illustration

Table 6.1. Population-I

Strata No.	N_i	$\overline{Y}_i$	$\overline{X}_i$	W_i	S_{iY}^2	S_{iX}^2
I	20	18.2	19.50	0.2325	53.92	48.76
II	49	127.79	103.14	0.5697	15023.62	10803.52
III	17	676.82	215.00	0.1976	1327757.9	56898.34
Total	$N = 86$	$\overline{Y} =$ 210.83	$\overline{X} =$ 105.80	-	-	-

Strata No.	ρ_i	V_i	C_{iX}	C_{iY}	n_i
I	0.9842	1.1118	0.3645	0.4117	8
II	0.9817	0.9343	1.0122	0.9634	15
III	0.8748	1.34238	1.1369	1.7446	7
Total	-	-	-	-	$n = 30$

From the above table we calculated the values as under;

For Population I:

$M(\bar{y}_{PR})$=613.31; $M(\bar{y}_{PP})$=916.44; $M(\bar{y}_{PST})$= 539.12; $M(\bar{y}_{PS})$=841.98

$M(\bar{y}_{PFT})_{opt.}$=416.19 at optimum choice of d_i – Values given below

Strata I :	d_{11}= 0.6813	d_{12}= 2.7116	$d_{13}=x$
Strata II :	d_{21}= 1.2421	d_{22}= 2.2048	d_{23}= 21.5121
Strata III :	d_{31}= 1.0098	d_{32}=2.0139	$d_{33}=x$

Table 6.2. Population-II

Strata No.	N_i	$\overline{Y}_i$	$\overline{X}_i$	W_i	S_{iY}^2	S_{iX}^2
I	20	7.45	12.75	0.3125	9.2626	21.7269
II	10	25.3	100.0	0.1562	70.7149	34.5396
III	34	196.0	208.0	0.5312	22846.13	22278.20
Total	N = 64	$\overline{Y}$ = 110.40	$\overline{X}$ = 130.10	-	-	-

Strata No.	ρ_i	V_i	C_{iX}	C_{iY}	n_i
I	0.8660	0.9692	0.3721	0.4165	10
II	0.7770	4.3943	0.0614	0.3476	4
III	0.9737	1.0452	0.7226	0.7758	14
Total	-	-	-	-	n =28

For Population II:

$M(\overline{y}_{PR})$=243.14; $M(\overline{y}_{PP})$=314.23; $M(\overline{y}_{PST})$= 191.69; $M(\overline{y}_{PS})$= 302.93

$M(\overline{y}_{PFT})_{opt.}$= 93.61 at optimum choice of d_i – Values given below

Strata I :	d_{11}= 1.0501	d_{12}= 2.9001	d_{13}= 47.3127
Strata II :	d_{21}= 0.0013	d_{22}= 0.0126	$d_{23}= x$
Strata III :	d_{31}= 0.9089	d_{32}=2.8813	$d_{33}= x$

Table 6.3. Values of d_i for different values of V_i and f.

V_i	Sampling Fraction $f\left(= n/N\right)$							
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.05	1.948924	1.906437	1.869907	1.837802	1.809140	1.783245	1.759631	1.737934
	3.004336	3.090815	3.169916	3.242959	3.310697	3.373560	3.431780	3.485478
	4.154634	4.165905	4.178598	4.192922	4.209109	4.227405	4.248063	4.271324
0.10	1.942750	1.896297	1.356985	1.822819	1.792575	1.765434	1.740822	1.718316
	2.927360	3.015233	3.094069	3.165710	3.231238	3.291345	3.346504	3.397076
	4.302112	4.321803	4.343390	4.367026	4.392854	4.420999	4.451561	4.484608
0.15	1.935694	1.884998	1.842824	1.806598	1.774807	1.746474	1.720925	1.697672
	2.858919	2.949986	3.030374	3.102562	3.167989	3.227587	3.282011	3.331760
	4.449504	4.476780	4.506214	4.537899	4.571909	4.608291	4.647064	4.688215
0.20	1.927573	1.872356	1.827262	1.788998	1.755716	1.726259	1.699844	1.675917
	2.796182	2.891592	2.974519	3.048211	3.114497	3.174542	3.229159	3.278962
	4.601245	4.636052	4.673219	4.712719	4.754786	4.799199	4.845997	4.895121
0.25	1.911854	1.858151	1.810113	1.769861	1.735164	1.704666	1.677472	1.652955
	2.737546	2.838152	2.924245	3.000000	3.067672	3.128667	3.183951	3.234244
	4.760967	4.803696	4.848975	4.896805	4.947163	5.000000	5.055244	5.112802
0.30	1.907139	1.842125	1.791161	1.749004	1.712993	1.681559	1.653687	1.628676
	2.682040	2.788526	2.878214	2.956385	3.025764	3.088006	3.144231	3.195258
	4.932249	4.983635	5.037768	5.094610	5.154099	5.216149	5.280654	5.347495
0.35	1.894150	1.823965	1.770152	1.726215	1.689021	1.656778	1.628347	1.602953
	2.629071	2.741989	2.835570	2.916389	2.987678	3.051352	3.108685	3.160597
	5.119086	5.180200	5.244278	5.311242	5.380993	5.453408	5.528353	5.605680
0.40	1.878701	1.803300	1.746790	1.701247	1.663037	1.630138	1.601288	1.575641
	2.578303	2.698071	2.795739	2.879350	2.952678	2.952678	3.076448	3.129338
	5.326329	5.398628	5.474137	5.552736	5.634285	5.634285	5.805597	5.895020
0.45	1.860176	1.779686	1.720727	1.673811	1.634794	1.601423	1.572321	1.546569
	2.529598	2.656472	2.758323	2.844804	2.920244	2.987088	3.046916	3.100843
	5.560225	5.645660	5.734587	5.826839	5.922235	6.020579	6.121673	6.225315
0.50	1.837792	1.752587	1.691545	1.643563	1.603996	1.570378	1.541218	1.515535
	2.482990	2.617008	2.723041	2.812412	2.889999	2.958501	3.019647	3.074646
	5.829218	5.930405	6.035414	6.144025	6.256005	6.371121	6.489136	6.609819
0.55	1.810570	1.721359	1.658747	1.610091	1.570292	1.536694	1.507709	1.482300
	2.438655	2.579577	2.689692	2.781922	2.861654	2.931829	2.994310	3.050399
	6.145219	6.265730	6.390450	6.519098	6.651387	6.787033	6.925759	7.067301
0.60	1.777311	1.685224	1.621725	1.572890	1.533252	1.500000	1.471468	1.446575
	2.396886	2.544142	2.658133	2.753142	2.834990	2.906829	2.970649	3.027834
	6.525803	6.670634	6.820142	6.973967	7.131758	7.293171	7.457884	7.625591

V_i	Sampling Fraction $f\left(=n/N\right)$							
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.65	1.736570	1.643222	1.579726	1.531337	1.492345	1.459837	1.432094	1.408008
	2.358041	2.510700	2.628260	2.725923	2.809832	2.883311	2.948459	3.006737
	6.998246	7.174649	7.356298	7.542739	7.733537	7.928280	8.126590	8.328112
0.70	1.686629	1.594155	1.531799	1.484642	1.446902	1.415630	1.389089	1.366165
	2.322475	2.479275	2.600000	2.700147	2.786039	2.861120	2.927577	2.986936
	7.607562	7.826570	8.051534	8.281877	8.517059	8.756583	9.000000	9.246899
0.75	1.625431	1.536485	1.476707	1.431782	1.396060	1.366643	1.341822	1.320498
	2.290454	2.449891	2.573293	2.675720	2.763497	2.840129	2.907868	2.968292
	8.434115	8.713624	9.000000	9.292498	9.590442	9.893228	10.20031	10.511210
0.80	1.550391	1.468154	1.412790	1.371390	1.338679	1.311916	1.289475	1.270307
	2.262092	2.422565	2.548093	2.652562	2.742110	2.820234	2.889219	2.950687
	9.637517	10.009281	10.389117	10.776048	11.169211	11.56785	11.971306	12.379006
0.85	1.457987	1.386277	1.337728	1.301579	1.273205	1.250154	1.230960	1.214676
	2.237324	2.397293	2.524359	2.630612	2.721798	2.801345	2.871537	2.934022
	11.588022	12.116431	12.654579	13.201142	13.754996	14.315167	14.880836	15.451302
0.90	1.342885	1.286538	1.248110	1.219625	1.197436	1.179558	1.164793	1.152362
	2.215920	2.374042	2.502052	2.609803	2.702493	2.783385	2.854739	2.918214
	15.391195	16.23942	17.099838	17.973947	18.850071	19.737057	20.630468	21.529424
0.95	1.196046	1.161963	1.138588	1.121375	1.108098	1.097511	1.088858	1.081642
	2.197546	2.352752	2.481127	2.590092	2.684134	2.766290	2.838757	2.903190
	26.556403	28.385284	30.230284	32.088532	33.957768	35.836198	37.722384	39.615168
1.00	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	2.181818	2.333333	2.461538	2.571430	2.666667	2.750000	2.823529	2.888889
1.05	0.713615	0.774124	0.814469	0.843004	0.864137	0.880362	0.893188	0.903566
	2.168352	2.315675	2.443234	2.553767	2.650044	2.734465	2.809003	2.875254
1.10	0.189041	0.408763	0.538420	0.623220	0.682586	0.726257	0.759624	0.785892
	2.156792	2.299652	2.426157	2.537063	2.634223	2.719640	2.795132	2.862239

REFERENCE

AGRAWAL, M.C. and PANDA, K.B. (1995): On efficient estimation in post-stratification, Metron 53, 107—115.

AGRAWAL, M.C. and PANDA, K.B. (1993): An efficient estimator in post stratification, Metron, 51, 3—4, 179—187.

HOLT, D. and SMITH, T.M.F. (1979): Post stratification, Jour. Roy. Stat. Soc., A, 142, 33—36.

JAGERS, P. (1986): Post stratification against bias in sampling, Int. Stat. Rev., 54,159—167.

JAGERS, P. ODEN, A. and TRULSSON, L. (1985): Post stratification and ratio estimation, Int. Stat. Rev., 53, 221—238.

REDDY, V.N. (1978): A study on the use of prior knowledge on certain population parameters in estimation, Sankhya, 40 C, 29—37.

SINGH, V.K. and SHUKLA, D. (1987): One parameter family of factor type ratio estimator, Metron, 45, 30, 1—2, 275—283.

SHUKLA, D., SINGH, V.K. and SINGH, G.N. (1991): On the use of transformation in factor type estimators, Metron, 49, 31, 1—4, 349—361.

SINGH, V.K. and SHUKLA, D. (1993): An efficient one-parameter family of factor- type estimators in sample surveys, Metron, 51, 30,1—2, 139—159.

SMITH, T.M. F. (1991): Post-stratification, The Statistician, 40, 315—323.

SRIVENKETARAMANA, T. (1980): A dual to ratio estimator in sample surveys, Biometrika, 67 (1), 199—204.

SUKHATME, P.V, SUKHATME, B.V., SUKHATME, S. and ASOK, C.(1984): Sampling Theory of Surveys With Applications, Iowa State University Press, Indian Society of Agricultural Statistics, New Delhi.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 273—296

EFFECT OF NON-RESPONSE IN SAMPLING OVER TWO SUCCESSIVE OCCASIONS USING AUXILIARY INFORMATION

Housila P. Singh[1], Sunil Kumar[2]

ABSTRACT

The problem of estimation of finite population in mail surveys for the current occasion in the context of sampling over two sampling occasion has been considered when there is non-response on current occasion. A general class of estimators has been suggested for population mean at current occasion in presence of non-response at current occasion in sampling on two occasions. It has been shown that the estimator reported by Singh and Priyanka (2007) is a member of the proposed estimator. The correct expression for the variance of one of the estimators of Singh and Priyanka (2007) is also given. The proposed estimator is more efficient than Singh and Priyanka (2007) estimator. The results obtained are illustrated with add of an empirical study. Both the theoretical and empirical results of the present study are encouraging and sound.

Key words: Non-response; Successive sampling; Mail surveys; Study variate; Auxiliary variate

1. Introduction

In most repeated surveys of the same population the current estimates are of primary interest if the characteristics of the population are likely to change rapidly with time. Jessen (1942), Yates (1949), Patterson (1950), Tikkiwal (1953), Eckler (1955) and Rao and Graham (1964) have contributed towards the development of the theory of successive sampling. Later various authors including Sen (1971), Sen (1972, 1973), Chaturvedi and Tripathi (1983), Das (1982), Singh et al (1991), Feng and Zou (1997), Birader and Singh (2001) and Singh and Singh (2001) have

[1] School of Studies in Statistics, Vikram University Ujjain – 456010, M. P., India, e-mail: hpsujn@rediffmail.com.

[2] School of Studies in Statistics, Vikram University Ujjain – 456010, M. P., India, sunilbhougal06@gmail.com.

used the information on single (or more) auxiliary variable(s) for estimating the current population parameters such as population mean or ratio in two occasions successive sampling. In many survey situations information on auxiliary variates may be readily available on the first as well as on the second occasions. Utilizing auxiliary information on both the occasions Singh (2005) and Singh and Priyanka (2006 a, b) have proposed chain type ratio, difference and regression type estimators for estimating the population mean at current (second) occasion in two occasions successive sampling which has been further generalized by Singh and Vishwakarma (2007).

In general during surveys, it is noted that information in most cases are not obtained at the first attempt even after some call-backs. An estimate obtained from such incomplete data may be misleading because of the biased estimator. This is the case of non-response and the usual approach to face the non-response is to recontact the non-respondents and obtained the information as much as possible. Hansen and Hurwitz (1946) envisaged a method tackling non-response in mail surveys. Cochran (1977), Rao (1986), Khare and Srivastava (1993, 1995, 1997), Fabian and Hyunshik (2000), Okafor and Lee (2000), Okafor (2001) and Tabasum and Khan (2004, 2006) extended the Hansen and Hurwitz technique to the case when besides the information on character under study, information is also available on auxiliary character. Choudhary et al. (2004) used the Hansen and Hurwitz technique for estimation of population mean on current occasion in the context of sampling on two occasions.

In two occasions successive sampling, a portion of sample is matched from the previous occasion and it is assumed that whole units respond at first occasion. So, we may think that as they are familiar with the questionnaire at first occasion therefore, they may not have any hesitation in responding at the second occasion for the units in the matched portion of the sample. At the current occasion a sample is drawn afresh from the remaining units, so there may be possibility of non-response at current occasion. Motivated with the above points and using Hansen and Hurwitz (1946) procedure Singh and Priyanka (2007) have studied the effect of non-response at current occasion in search of good rotation patterns on successive occasions. Two difference type estimators have been proposed by them for estimating the population mean at current occasion in presence of non-response in two occasions successive (rotation) sampling. It is found that the results of Singh and Priyanka (2007) are erroneous.

In this paper we have made an effort to suggest a general class of estimators for estimating the population mean at current occasion in presence of non-response in two occasion successive sampling. It is identified that the estimator suggested by Singh and Priyanka (2007) is a particular member of the proposed class. It has also been shown that the proposed estimator is more efficient then Singh and Priyanka (2007) estimator. Numerical illustration is given in support of the present study.

2. Notations and the proposed estimator

Consider a finite population $U = (U_1, U_2, ..., U_N)$ of N distinct and identifiable units. The character under study is denoted by $x(y)$ on the first (second) occasions respectively. Let the information on an auxiliary variable z with known population mean $\bar{Z}$, be available on both the occasions. A simple random sample (without replacement) of n units is taken on the first occasion. A random sub-sample of $m = n\lambda$ units is retained (matched) for use on the second occasion. We suppose that there is non-response at the current occasion, so that the population can be divided into two classes, those who will respond at the first attempt and those who will not. Let the sizes of these two classes be N_1 and N_2 respectively. Now, at the current occasion a simple random sample (without replacement) of $u = (n - m) = n\mu$ units is drawn afresh from the remaining $(N - n)$ units of the population so that the sample size on the second occasion is also n. λ and $\mu(\lambda + \mu = 1)$ are the fractions of matched and fresh samples respectively at the second (current) occasion. We assume that in the unmatched portion of the sample on the two occasions u_1 units respond and u_2 units do not respond. Let u_{2h} denote the size of sub sample drawn from the non-response class in the unmatched portion of the sample on the current occasion. In what follows we shall use the following notations for further use:

- $\bar{X}, \bar{Y}, \bar{Z}$: The population means of the variables x, y and z respectively.
- $\bar{x}_n, \bar{z}_n, \bar{y}_m, \bar{x}_m, \bar{z}_m, \bar{y}_{u_1}, \bar{y}_{u_{2h}}, \bar{z}_u$: The sample means of the respective variates of the sample sizes shown in suffixes.
- S_{xy}, S_{xz}, S_{yz} : The population covariances between the variables shown in suffixes.
- $\rho_{yx}, \rho_{xz}, \rho_{yz}$: The correlation coefficients between the variables shown in suffixes.
- S_x^2, S_y^2, S_z^2 : The population mean squares of x, y and z respectively.
- $W\left(\frac{N_2}{N}\right)$: The proportion of non- responding units in the population.

We assume that information on auxiliary character is readily available on both the occasions and is not changing rapidly over time. Under this assumption we suggest a class of estimators of population mean $\bar{Y}$ considering the problem of non-response only at current occasion. With this background utilizing the information on auxiliary character z, we define the following class of estimators for $\bar{Y}$ as

$$d = \phi d_1 + (1-\phi)d_2, \tag{2.1}$$

where the estimator

$$d_1 = \left[\bar{y}_u^*\left(\frac{\bar{z}_u}{\bar{Z}}\right)^{\alpha} + \delta\left(\bar{Z} - \bar{z}_u\right)\right], \tag{2.2}$$

based on u units, which are drawn afresh at current occasion such that out of these u units, u_1 units respond and u_2 units do not respond. Let u_{2h} be the size of sub sample drawn from the non-response class in the unmatched portion of the sample on current occasion and the estimator

$$d_2 = \left[\bar{y}_m + \alpha_1(\bar{x}_n - \bar{x}_m) + \alpha_2(\bar{z}_n - \bar{z}_m) + \alpha_3(\bar{Z} - \bar{z}_m)\right], \tag{2.3}$$

based on m units, which are retained from previous occasion.

- $\bar{y}_u^* = \dfrac{u_1\bar{y}_{u_1} + u_2\bar{y}_{u_{2h}}}{u}$, (2.4)
- is a Hansen and Hurwitz (1946) estimator based on sample of size $u(= n\mu)$ drawn afresh on the second occasion.
- (α, δ) and $(\alpha_1, \alpha_2, \alpha_3)$ are suitably chosen constants to be determined such that the variances of d_1 and d_2 are minimum respectively.
- ϕ is an unknown constant to be determined under certain criterion such that the variance of d is least.

The estimator d reduces to the following set of estimators for different values of $(\alpha, \delta, \alpha_1, \alpha_2, \alpha_3)$:

(i) $d_{(1)} = \phi d_{1(1)} + (1-\phi)d_{2(1)}$, (2.5)

for $(\alpha, \delta) = (0,0)$ and $(\alpha_1, \alpha_2, \alpha_3) = (\beta_{yx}, 0, 0)$, where $d_{1(1)} = \bar{y}_u^*$,

$d_{2(1)} = \bar{y}_m + \beta_{yx}(\bar{x}_n - \bar{x}_m)$. (2.6)

β_{yx} being the known population regression coefficient between the variates shown in suffixes.

(ii) $d_{(2)} = \phi d_{1(2)} + (1-\phi)d_{2(2)}$, (2.7)

for $(\alpha, \delta) = (0, \beta_{yz})$ and $(\alpha_1, \alpha_2, \alpha_3) = (\beta_{yx}, -\beta_{yx}\beta_{xz}, \beta_{yz})$,

where $d_{1(2)} = \bar{y}_u^* + \beta_{yz}(\bar{Z} - \bar{z}_u) = d_{1(0)}(say)$, (2.8)

$d_{2(2)} = \bar{y}_m + \beta_{yx}(\bar{x}_n - \bar{x}_m) - \beta_{yx}\beta_{xz}(\bar{z}_n - \bar{z}_m) + \beta_{yz}(\bar{Z} - \bar{z}_m)$

$= \bar{y}_m^* + \beta_{yx}(\bar{x}_n^* - \bar{x}_m^*)$ (2.9)

where

$\bar{x}_n^* = \bar{x}_n + \beta_{xz}\left(\bar{Z} - \bar{z}_n\right), \bar{x}_m^* = \bar{x}_m + \beta_{xz}\left(\bar{Z} - \bar{z}_m\right), \bar{y}_m^* = \bar{y}_m + \beta_{yz}\left(\bar{Z} - \bar{z}_m\right),$ $\left(\beta_{yx}, \beta_{yz}, \beta_{xz}\right)$ being the known population regression coefficients between the variates shown in suffixes. It is to be mentioned that the estimators $d_{(1)}$ and $d_{(2)}$ are due to Singh and Priyanka (2007). Thus the estimators $d_{(1)}$ and $d_{(2)}$ due to Singh and Priyanka (2007) are particular members of the proposed class of estimators d at (2.1).

(iii) $d_{(3)} = \phi d_{1(3)} + (1-\phi)d_{2(3)}$, (2.10)

for $(\alpha, \delta) = (-1,0)$ and $(\alpha_1, \alpha_2, \alpha_3) = \left(\beta_{yx}, -\beta_{yx}\beta_{xz}, \beta_{yz}\right)$,

where $d_{1(3)} = \bar{y}_u^*\left(\bar{Z}/\bar{z}_u\right)$, (2.11)

$d_{2(3)} = d_{2(2)}$. (2.12)

(iv) $d_{(4)} = \phi d_{1(4)} + (1-\phi)d_{2(4)}$, (2.13)

for $(\alpha, \delta) = (1,0)$ and $(\alpha_1, \alpha_2, \alpha_3) = \left(\beta_{yx}, -\beta_{yx}\beta_{xz}, \beta_{yz}\right)$,

where $d_{1(4)} = \bar{y}_u^*\left(\bar{z}_u^*/\bar{Z}\right)$, (2.14)

$d_{2(4)} = d_{2(2)}$. (2.15)

Many estimators for population mean $\bar{Y}$ can be generalized from the proposed class of estimators d for different values of $(\alpha, \delta, \alpha_1, \alpha_2, \alpha_3)$.

2.1.Search of optimum estimators in the classes d_1 and d_2

The bias and variance of d_1 to the first degree of approximation, are respectively given by

$$B(d_1) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{\alpha \bar{Y} C_z^2}{2}\right)\left[\alpha + 2\left(\frac{\beta_{yz}}{R_z}\right) - 1\right], \tag{2.16}$$

$$Var(d_1) = \left\{\frac{A}{u} - \frac{(A-B)}{N}\right\} \tag{2.17}$$

where $A = \left[\{1 + (f-1)W\}S_y^2 + \left\{(\alpha R_z - \delta)^2 + 2(\alpha R_z - \delta)\beta_{yz}\right\}S_z^2\right]$,

$$B = (f-1)WS_y^2 \text{ and } R_z = \left(\bar{Y}/\bar{Z}\right).$$

Putting the different values of (α, δ) in (2.16) and (2.17) one can easily obtain the biases and variances of different estimators generated from d_1.

Minimization of (2.17) yields the optimum value of $(\alpha R_z - \delta)$ as

$$(\alpha R_z - \delta) = -\beta_{yz} \tag{2.18}$$

which shows that asymptotically optimum estimator (AOE) in the class of estimators d_1 is not unique. Substitution of (2.18) in (2.17) yields the minimum variance of d_1 as

$$\min.Var(d_1) = \left[\frac{\{(1-\rho_{yz}^2)+(f-1)W\}}{u} - \frac{(1-\rho_{yz}^2)}{N}\right]S_y^2$$

or

$$\min.Var(d_1) = Var(d_{1(2)}) = Var(d_{1(0)}) = \left[\frac{Q^*}{u} - \frac{P}{N}\right]S_y^2 \tag{2.19}$$

where $Q^* = (1-\rho_{yz}^2)+(f-1)W$ and $P = (1-\rho_{yz}^2)$.

Thus we established the following theorem.

THEOREM 2.1.

To the first degree of approximation,

$$Var(d_1) \geq \left[\frac{Q^*}{u} - \frac{P}{N}\right]S_y^2$$

with equality holding if

$$(\alpha R_z - \delta) = -\beta_{yz}. \tag{2.20}$$

COROLLARY 2.1.

(i) For $\alpha = 0$, the optimum value of δ which minimizes variance d_1 is

$$\delta = \beta_{yz} \tag{2.21}$$

and hence the optimum estimator is

$$d_{1(0)} = \bar{y}_u^* + \beta_{yz}(\bar{Z} - \bar{z}_u) \tag{2.22}$$

which is due to Singh and Priyanka (2007). The variance of $d_{1(0)}$ is given by $Var(d_{1(0)}) = \min.Var(d_1)$ where $\min.Var(d_1)$ is given by (2.19).

(ii) For $\delta = 0$, the optimum value of α which minimizes variance of d_1 is

$$\alpha = -(\beta_{yz}/R_z) = -K \text{ (say)} \tag{2.23}$$

and hence we get the optimum estimator in the class d_1 as

$$d_{1(1)} = \bar{y}_u^*\left(\bar{Z}/\bar{z}_u\right)^K . \tag{2.24}$$

(iii) For $\alpha = 1$, we get the optimum value of δ as

$$\delta = \left(\beta_{yz} + R_z\right) = R_z\left(1 + K\right) \tag{2.25}$$

and hence we get the optimum estimator in the class d_1 as

$$d_{1(2)} = \bar{y}_u^*\left(\bar{z}_u/\bar{Z}\right) + R_z\left(1 + K\right)\left(\bar{Z} - \bar{z}_u\right) \tag{2.26}$$

(iv) For $\alpha = -1$, we get the optimum value of δ as

$$\delta = \left(\beta_{yz} - R_z\right) = R_z\left(K - 1\right) \tag{2.27}$$

and hence we get the optimum estimator in the class d_1 as

$$d_{1(3)} = \bar{y}_u^*\left(\bar{Z}/\bar{z}_u\right) + R_z\left(K - 1\right)\left(\bar{Z} - \bar{z}_u\right) \tag{2.28}$$

Proceeding in similar way one can obtain different optimum estimators for different values of (α, δ). Thus there is no optimum estimator in the class of estimators d_1 which is unique.

Now we will obtain the optimum estimator in the class of estimators d_2.

From (2.3) we have that

$$E\left(d_2\right) = \bar{Y} \tag{2.29}$$

and

$$Var\left(d_2\right) = \left(\frac{1}{m} - \frac{1}{N}\right)\left(S_y^2 + \alpha_3^2 S_z^2 - 2\alpha_3 S_{yz}\right)$$
$$+ \left(\frac{1}{m} - \frac{1}{n}\right)\left(\alpha_1^2 S_x^2 + \alpha_2^2 S_z^2 - 2\alpha_1 S_{xy} - 2\alpha_2 S_{yz} + 2\alpha_1\alpha_2 S_{xz} + 2\alpha_1\alpha_3 S_{xz} + 2\alpha_2\alpha_3 S_z^2\right)$$

Expression (2.29) shows that the estimator d_2 is unbiased. Putting the different values of $\alpha_i{}'s, i = 1,2,3$ in (2.30) one can obtain the variances of different estimators generated from d_2. Minimization of (2.30) gives the optimum values of $\alpha_i{}'s, i = 1,2,3$; as

$$\left.\begin{aligned} \alpha_1 &= \frac{\beta_{\text{yx}} - \beta_{\text{yz}}\beta_{\text{zx}}}{1 - \beta_{\text{xz}}\beta_{\text{zx}}} = \alpha_{10}(\text{say}), \\ \alpha_2 &= -\frac{\beta_{\text{xz}}\left(\beta_{\text{yx}} - \beta_{\text{yz}}\beta_{\text{zx}}\right)}{1 - \beta_{\text{xz}}\beta_{\text{zx}}} = \alpha_{20}(\text{say}), \\ \alpha_3 &= \beta_{\text{yz}} = \alpha_{30}(\text{say}) \end{aligned}\right\} \tag{2.31}$$

where β_{yx} : population regression coefficient between y on x,

β_{yz} : population regression coefficient between y on z,

β_{xz} : population regression coefficient between x on z,

β_{zx} : population regression coefficient between z on x.

Putting (2.31) in (2.30) yields the minimum variance of d_2 as

$$\min.Var(d_2)=\left[\left(\frac{1}{m}-\frac{1}{N}\right)\left(1-\rho_{yz}^2\right)-\left(\frac{1}{m}-\frac{1}{n}\right)\frac{\left(\rho_{yx}-\rho_{yz}\rho_{zx}\right)^2}{\left(1-\rho_{xz}^2\right)}\right]S_y^2$$

or

$$\min.Var(d_2)=\left[\left(\frac{1}{m}-\frac{1}{N}\right)P+\left(\frac{1}{m}-\frac{1}{n}\right)Q\right]S_y^2 \tag{2.32}$$

where

$$P=\left(1-\rho_{yz}^2\right) \text{ and } Q=-\frac{\left(\rho_{yx}-\rho_{yz}\rho_{zx}\right)^2}{\left(1-\rho_{xz}^2\right)}.$$

Thus we established the following theorem:

THEOREM 2.2.

To the first degree of approximation,

$$Var(d_2)\geq\left[\left(\frac{1}{m}-\frac{1}{N}\right)P+\left(\frac{1}{m}-\frac{1}{n}\right)Q\right]S_y^2$$

with equality holding if

$$\alpha_1=\alpha_{10},\alpha_2=\alpha_{20},\alpha_3=\alpha_{30}.$$

Further substitution of (2.31) in (2.3) yields the optimum estimator in the class of estimators d_2 as

$$d_{2(0)}=\left[\bar{y}_m+\frac{\beta_{yx}-\beta_{yz}\beta_{zx}}{1-\beta_{xz}\beta_{zx}}\left(\bar{x}_n-\bar{x}_m\right)-\frac{\beta_{xz}\left(\beta_{yx}-\beta_{yz}\beta_{zx}\right)}{1-\beta_{xz}\beta_{zx}}\left(\bar{z}_n-\bar{z}_m\right)+\beta_{yz}\left(\bar{Z}-\bar{z}_m\right)\right] \tag{2.33}$$

with the variance $Var\left(d_{2(0)}\right)=\min.Var(d_2)$,

where $\min.Var(d_2)$ is given by (2.32).

2.2.Search of best combined estimator

As advocated by Singh and Priyanka (2007) that β_{yx}, β_{yz}, β_{xz} and β_{zx} can be known, therefore, we define a combined estimator for $\overline{Y}$ as

$$\mathrm{d}_{(0)} = \phi \mathrm{d}_{1(0)} + (1-\phi)\mathrm{d}_{2(0)} \tag{2.34}$$

where $d_{1(0)}$ and $d_{2(0)}$ are respectively defined in (2.22) and (2.33).

The variance of $d_{(0)}$ is given by

$$\mathrm{Var}(\mathrm{d}_{(0)}) = \phi^2 \mathrm{Var}(\mathrm{d}_{1(0)}) + (1-\phi)^2 \mathrm{Var}(\mathrm{d}_{2(0)}) \tag{2.35}$$

which is minimum where

$$\phi = \frac{Var(d_{2(0)})}{Var(d_{1(0)}) + Var(d_{2(0)})} \tag{2.36}$$

$$\Rightarrow (1-\phi) = \frac{Var(d_{1(0)})}{Var(d_{1(0)}) + Var(d_{2(0)})},$$

where $Var(d_{1(0)})$ and $Var(d_{2(0)})$ are respectively given by (2.19) and (2.32).

Thus the resulting variance of $\mathrm{d}_{(0)}$ is given by

$$\mathrm{Var}(\mathrm{d}_{(0)})_{\mathrm{opt}} = \frac{\mathrm{Var}(\mathrm{d}_{1(0)})\mathrm{Var}(\mathrm{d}_{2(0)})}{\mathrm{Var}(\mathrm{d}_{1(0)}) + \mathrm{Var}(\mathrm{d}_{2(0)})} \tag{2.37}$$

or

$$\mathrm{Var}(\mathrm{d}_{(0)})_{\mathrm{opt}} = \frac{(\mathrm{b}_1\mu^2 + \mathrm{b}_2\mu + \mathrm{b}_3)\mathrm{S}_\mathrm{y}^2}{(\mathrm{b}_4\mu^2 + \mathrm{b}_5\mu + \mathrm{Q}^*)}, \tag{2.38}$$

where

$$b_1 = -\left(\frac{nP^2}{N^2} + \frac{PQ}{N}\right), b_2 = \left(\frac{nP^2}{N^2} + \frac{Q^*P}{N} + \frac{Q^*Q}{n} - \frac{P^2}{N}\right), b_3 = Q^*P\left(\frac{1}{n} - \frac{1}{N}\right),$$
$$b_4 = \left(Q + \frac{2nP}{N}\right), b_5 = \left\{W_2(1-f) - \frac{2nP}{N}\right\} and\ \mu\left(= \frac{u}{n}\right)$$

which is the fraction of fresh sample taken at second (current) occasion.

3. Optimum replacement policy

To obtain the optimum value of μ, we minimize $\mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}}$ in (2.38) with respect to μ, we get

$$\hat{\mu} = \frac{-r_2 \pm \sqrt{r_2^2 - 4r_1r_3}}{2r_1} = \mu_0^* \text{(say)} \tag{3.1}$$

where $r_1 = b_1b_5 - b_2b_4, r_2 = 2Q^*b_1 - 2b_3b_4$ and $r_3 = Q^*b_2 - b_3b_5$.

Since $r_2^2 - 4r_1r_3 \geq 0$, two values of $\hat{\mu}$ are possible, hence to choose a value of $\hat{\mu}$, one should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values are inadmissible. In case if both the values of $\hat{\mu}$ are admissible, we choose the minimum of these two μ_0^*. Substituting the value of $\hat{\mu}$ from (3.1) in (2.38), we have

$$\mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}^*} = \frac{\left(\mathrm{b}_1\mu_0^{*2} + \mathrm{b}_2\mu_0^* + \mathrm{b}_3\right)\mathrm{S}_\mathrm{y}^2}{\left(\mathrm{b}_4\mu_0^{*2} + \mathrm{b}_5\mu_0^* + \mathrm{Q}^*\right)} \tag{3.2}$$

where $\mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}^*}$ is the optimum value of $\mathrm{d}_{(0)}$ with respect to both ϕ and μ.

4. Efficiency comparisons

(i) Singh and Priyanka (2007) suggested the following combined estimator for $\overline{Y}$ as

$$d_{(1)} = \phi d_{1(1)} + (1-\phi)d_{2(1)}, \tag{4.1}$$

where $d_{1(1)} = \overline{y}_u^* = \dfrac{u_1\overline{y}_{u_1} + u_2\overline{y}_{u_{2h}}}{u}$ and $d_{2(1)}$ is given by (2.6).

The variance of $d_{(1)}$ is given by

$$Var\left(d_{(1)}\right) = \phi^2 Var\left(d_{1(1)}\right) + (1-\phi)^2 Var\left(d_{2(1)}\right) \tag{4.2}$$

which is minimum when

$$\phi = \frac{Var\left(d_{2(1)}\right)}{Var\left(d_{1(1)}\right) + Var\left(d_{2(1)}\right)} \tag{4.3}$$

where

$$Var\left(d_{1(1)}\right)=\left[\left(\frac{1}{u}-\frac{1}{N}\right)+\frac{(f-1)W}{u}\right]S_y^2 \tag{4.4}$$

and

$$Var\left(d_{2(1)}\right)=\left[\left(\frac{1}{m}-\frac{1}{n}\right)P+\left(\frac{1}{n}-\frac{1}{N}\right)\right]S_y^2 \tag{4.5}$$

Putting (4.3) in (4.2) we get the minimum variance of $d_{(1)}$ as

$$Var\left(d_{(1)}\right)_{opt}=\frac{Var\left(d_{1(1)}\right)Var\left(d_{2(1)}\right)}{Var\left(d_{1(1)}\right)+Var\left(d_{2(1)}\right)}=\frac{\left(A_1\mu^2+A_2\mu+A_3\right)S_y^2}{\left(A_4\mu^2+A_5\mu+A_0\right)} \tag{4.6}$$

where

$$\mathrm{A}_1=\left(\frac{\rho_{\mathrm{yx}}^2}{\mathrm{N}}-\frac{\mathrm{n}}{\mathrm{N}^2}\right),\mathrm{A}_2=\left(\frac{\mathrm{n}}{\mathrm{N}^2}-\frac{\mathrm{A}_0}{\mathrm{n}}\rho_{\mathrm{yx}}^2+\frac{\mathrm{A}_0-1}{\mathrm{N}}\right),\mathrm{A}_3=\mathrm{A}_0\left(\frac{1}{\mathrm{n}}-\frac{1}{\mathrm{N}}\right),$$

$$\mathrm{A}_4=\left(\frac{2\mathrm{n}}{\mathrm{N}}-\rho_{\mathrm{yx}}^2\right),\mathrm{A}_5=\left(1-\mathrm{A}_0-\frac{2\mathrm{n}}{\mathrm{N}}\right),\mathrm{A}_0=\left(1+(\mathrm{f}-1)\mathrm{W}\right)\text{and}\,\mu\left(=\frac{\mathrm{u}}{\mathrm{n}}\right)$$

is the fraction of fresh sample taken at second (current) occasion.

To obtain the optimum value of μ, we minimize $Var\left(d_{(1)}\right)$ with respect to μ, we get the optimum value of μ as

$$\hat{\mu}=\frac{-B_2\pm\sqrt{B_2^2-4B_1B_3}}{2B_1}=\mu_0\,(\text{say}) \tag{4.7}$$

where $B_1=A_1A_5-A_2A_4, B_2=2A_0A_1-2A_3A_4$ and $B_3=A_0A_2-A_3A_5$.

Since $B_2^2-4B_1B_3\geq 0$, two values of $\hat{\mu}$ are possible, hence to choose a value of $\hat{\mu}$, one should be remembered that $0\leq\hat{\mu}\leq 1$. All other values are inadmissible. In case if both the values of $\hat{\mu}$ are admissible, we choose the minimum of these two μ_0. Substituting the value of $\hat{\mu}$ from (4.7) in (4.6), we have

$$\mathrm{Var}\left(\mathrm{d}_{(1)}\right)_{\mathrm{opt}^*}=\frac{\left(\mathrm{A}_1\mu_0^2+\mathrm{A}_2\mu_0+\mathrm{A}_3\right)\mathrm{S}_{\mathrm{y}}^2}{\left(\mathrm{A}_4\mu_0^2+\mathrm{A}_5\mu_0+\mathrm{A}_0\right)} \tag{4.8}$$

where $Var\left(d_{(1)}\right)_{opt^*}$ is the optimum value of $d_{(1)}$ with respect to both ϕ and μ.

(ii) Further Singh and Priyanka (2007) proposed another combined estimator for population mean $\overline{Y}$ as

$$d_{(2)} = \phi d_{1(2)} + (1-\phi)d_{2(2)}$$

where $d_{1(2)}$ and $d_{2(2)}$ are defined by (2.8) and (2.9) respectively. For the sake of similarity in notations we take $d_{1(2)} = d_{1(0)} = \bar{y}_u^* + \beta_{yz}\left(\overline{Z} - \bar{z}_u\right)$. Thus

$$d_{(2)} = \phi d_{1(0)} + (1-\phi)d_{2(2)} \tag{4.9}$$

The variance of $d_{(2)}$ is given by

$$Var\left(d_{(2)}\right) = \phi^2 Var\left(d_{1(0)}\right) + (1-\phi)^2 Var\left(d_{2(2)}\right) \tag{4.10}$$

which is minimum when

$$\phi = \frac{Var\left(d_{2(2)}\right)}{Var\left(d_{1(0)}\right) + Var\left(d_{2(2)}\right)} \tag{4.11}$$

where

$$Var\left(d_{1(0)}\right) = \left[\frac{Q^*}{u} + \frac{P}{N}\right]S_y^2 \tag{4.12}$$

and

$$Var\left(d_{2(2)}\right) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)P + \left(\frac{1}{m} - \frac{1}{n}\right)C\right]S_y^2 \tag{4.13}$$

where $C = 2\rho_{yx}\rho_{yz}\rho_{xz} - \rho_{yx}^2\left(1+\rho_{xz}^2\right)$. Here it is to be mentioned that the variance expression for the estimator $\mathrm{d}_{2(2)} = \Delta_2$ obtained by Singh and Priyanka (2007, eq. (23), p. 284) is incorrect. The expression (4.13) is the correct version of the variance of $d_{2(2)}$ obtained by Singh and Priyanka (2007).

Putting (4.11) in (4.10) we get the minimum variance of $d_{(2)}$ as

$$Var\left(d_{(2)}\right)_{opt} = \frac{Var\left(d_{1(0)}\right)Var\left(d_{2(2)}\right)}{Var\left(d_{1(0)}\right) + Var\left(d_{2(2)}\right)} = \frac{\left(d_1\mu^2 + d_2\mu + b_3\right)S_y^2}{\left(d_4\mu^2 + b_5\mu + Q^*\right)} \tag{4.14}$$

where

$$d_1 = -\left(\frac{nP^2}{N^2} + \frac{PC}{N}\right), d_2 = \left(\frac{nP^2}{N^2} + \frac{Q^*P}{N} + \frac{Q^*C}{n} - \frac{P^2}{N}\right), d_4 = \left(C + \frac{2Pn}{N}\right)$$

and $\mu\left(=\frac{u}{n}\right)$ is the fraction of fresh sample taken at second (current) occasion.

To obtain the optimum value of μ, we minimize $Var\left(d_{(2)}\right)$ with respect to μ, we get the optimum value of μ as

$$\hat{\mu} = \frac{-a_2 \pm \sqrt{a_2^2 - 4a_1a_3}}{2a_1} = \mu_0' \text{(say)} \tag{4.15}$$

where $a_1 = d_1b_5 - d_2d_4, a_2 = 2Q^*d_1 - 2b_3d_4$ and $a_3 = Q^*d_2 - b_3b_5$.

Since $a_2^2 - 4a_1a_3 \geq 0$, two values of $\hat{\mu}$ are possible, hence to choose a value of $\hat{\mu}$, one should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values are inadmissible. In case if both the values of $\hat{\mu}$ are admissible, we choose the minimum of these two μ_0'. Substituting the value of $\hat{\mu}$ from (4.15) in (4.14), we have

$$\mathrm{Var}\left(\mathrm{d}_{(2)}\right)_{\mathrm{opt}^*} = \frac{\left(\mathrm{d}_1\mu_0'^2 + \mathrm{d}_2\mu_0' + \mathrm{b}_3\right)\mathrm{S}_\mathrm{y}^2}{\left(\mathrm{d}_4\mu_0'^2 + \mathrm{b}_5\mu_0' + \mathrm{Q}^*\right)} \tag{4.16}$$

where $Var\left(d_{(2)}\right)_{opt^*}$ is the optimum value of $d_{(2)}$ with respect to both ϕ and μ.

From (2.35), (4.6) and (4.14) we have

$$\mathrm{Var}\left(\mathrm{d}_{(1)}\right)_{\mathrm{opt}} - \mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}} = \frac{1}{\mathrm{D}}\left\{\mathrm{Var}\left(\mathrm{d}_{1(0)}\right)\mathrm{Var}\left(\mathrm{d}_{1(1)}\right)\mathrm{D}_1 + \mathrm{Var}\left(\mathrm{d}_{2(0)}\right)\mathrm{Var}\left(\mathrm{d}_{2(1)}\right)\mathrm{D}_2\right\} \tag{4.17}$$

$$\mathrm{Var}\left(\mathrm{d}_{(2)}\right)_{\mathrm{opt}} - \mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}} = \frac{1}{\mathrm{D}^*}\left\{\mathrm{Var}\left(\mathrm{d}_{1(0)}\right)\mathrm{Var}\left(\mathrm{d}_{1(2)}\right)\mathrm{D}_1^*\right\} \tag{4.18}$$

where

$$D_1 = Var\left(d_{2(1)}\right) - Var\left(d_{2(0)}\right),\ D_2 = Var\left(d_{1(1)}\right) - Var\left(d_{1(0)}\right),$$
$$D_1^* = Var\left(d_{2(2)}\right) - Var\left(d_{2(0)}\right),$$

$$D = \left[\left\{Var\left(d_{1(1)}\right) + Var\left(d_{2(1)}\right)\right\}\left\{Var\left(d_{1(0)}\right) + Var\left(d_{2(0)}\right)\right\}\right],$$

$$D^* = \left[\left\{Var\left(d_{1(0)}\right) + Var\left(d_{2(2)}\right)\right\}\left\{Var\left(d_{1(0)}\right) + Var\left(d_{2(0)}\right)\right\}\right].$$

It can be shown that

$$D_1 = \left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{\rho_{yz}^2\left(1-\rho_{yz}^2\right)+\frac{\left(\rho_{yx}-\rho_{yz}\rho_{zx}\right)^2}{\left(1-\rho_{xz}^2\right)}\right\}+\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{yz}^2\right]S_y^2 \geq 0 \quad (4.19)$$

$$D_2 = \left(\frac{1}{u}-\frac{1}{N}\right)\rho_{yz}^2 S_y^2 \geq 0 \quad (4.20)$$

$$D_1^* = \left(\frac{1}{m}-\frac{1}{n}\right)\frac{\rho_{xz}^2\left(\rho_{yz}-\rho_{yx}\rho_{xz}\right)^2}{\left(1-\rho_{xz}^2\right)}S_y^2 \geq 0 \quad (4.21)$$

Thus from (4.17), (4.18), (4.19), (4.20) and (4.21) we have the following inequalities:

$$\mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}} \leq \mathrm{Var}\left(\mathrm{d}_{(1)}\right)_{\mathrm{opt}} \quad (4.22)$$

and

$$\mathrm{Var}\left(\mathrm{d}_{(0)}\right)_{\mathrm{opt}} \leq \mathrm{Var}\left(\mathrm{d}_{(2)}\right)_{\mathrm{opt}} \quad (4.23)$$

It follows from (4.22) and (4.23) that the proposed estimator $d_{(0)}$ is better than Singh and Priyanka (2007) estimators $d_{(1)}$ and $d_{(2)}$.

5. Estimation of mean in successive sampling on two occasions when there is no non-response

Here we consider the following combined estimator for $\overline{Y}$ as

$$t = \phi d^* + (1-\phi)d_{2(0)} \quad (5.1)$$

where

$$d^* = \bar{y}_u + \beta_{yz}\left(\overline{Z}-\bar{z}_u\right) \quad (5.2)$$

and

$$d_{2(0)} = \left[\bar{y}_m + \frac{\left(\beta_{yx}-\beta_{yz}\beta_{zx}\right)}{\left(1-\beta_{xz}\beta_{zx}\right)}\left(\bar{x}_n-\bar{x}_m\right)-\frac{\beta_{xz}\left(\beta_{yx}-\beta_{yz}\beta_{zx}\right)}{\left(1-\beta_{xz}\beta_{zx}\right)}\left(\bar{z}_n-\bar{z}_m\right)+\beta_{yz}\left(\overline{Z}-\bar{z}_m\right)\right] \quad (5.3)$$

which is same as defined in (2.33).

It is obvious that the estimator t is an unbiased estimator of $\overline{Y}$. The variance of t is given by

$$Var(t) = \phi^2 Var(d^*) + (1-\phi)^2 Var(d_{2(0)}) \tag{5.4}$$

where

$$Var(d^*) = \left(\frac{1}{u} - \frac{1}{N}\right) PS_y^2 \tag{5.5}$$

and $Var(d_{2(0)})$ is defined at (2.32). Since t is unbiased estimator of $\overline{Y}$, the optimum variance of t is given by

$$Var(t)_{opt^*} = \frac{Var(d^*)Var(d_{2(0)})}{Var(d^*) + Var(d_{2(0)})} = \frac{(b_1\mu_1'^2 + t_2\mu_1' + t_3)S_y^2}{(b_4\mu_1'^2 + t_5\mu_1' + P)} \tag{5.6}$$

where

$$\mu_1' = \frac{-s_2 \pm \sqrt{s_2^2 - 4s_1s_3}}{2s_1}, t_2 = \left(\frac{PQ}{n} + \frac{nP^2}{N^2}\right), t_3 = P^2\left(\frac{1}{n} - \frac{1}{N}\right), t_5 = -\frac{2nP}{N},$$

$s_1 = b_1t_5 - t_2b_4, s_2 = 2Pb_1 - 2t_3b_4$ and $s_3 = Pt_2 - t_3t_5$.

REMARK 5.1.

μ_1' being the optimum value of μ, so for certain situations there may be two values of μ_1' so while choosing μ_1' we must remember $0 \le \mu_1' \le 1$. However, if both the values of μ_1' are admissible, we choose the minimum of the two values as μ_1'.

When there is no non-response (i.e. in absence of non-response) Singh and Priyanka (2006 b) suggested an estimator for population mean $\overline{Y}$ (on current occasion in sampling over two successive occasion) as

$$d^{**} = \psi d^* + (1-\psi)d_{2(2)} \tag{5.7}$$

where d^* and $d_{2(2)}$ are defined in (5.2) and (2.9) respectively and ψ is an unknown constant to be determined under certain criterion. The variance of d^{**} is given by

$$Var(d^{**}) = \psi^2 Var(d^*) + (1-\psi)^2 Var(d_{2(2)}) \tag{5.8}$$

where $Var(d^*)$ and $Var(d_{2(2)})$ are given in (5.5) and (4.13) respectively.

Thus variance of d^{**} is minimum when

$$\psi=\frac{Var\left(d_{2(2)}\right)}{Var\left(d^{*}\right)+Var\left(d_{2(2)}\right)} \tag{5.9}$$

$$\left(1-\psi\right)=\frac{Var\left(d^{*}\right)}{Var\left(d^{*}\right)+Var\left(d_{2(2)}\right)} \tag{5.10}$$

Thus the resulting minimum variance of d^{**} is given by

$$\mathrm{Var}\left(\mathrm{d}^{**}\right)_{\mathrm{opt}^{*}}=\frac{\mathrm{Var}\left(\mathrm{d}^{*}\right)\mathrm{Var}\left(\mathrm{d}_{2(2)}\right)}{\mathrm{Var}\left(\mathrm{d}^{*}\right)+\mathrm{Var}\left(\mathrm{d}_{2(2)}\right)}=\frac{\left(q_1\mu_1^2+q_2\mu_1+t_3\right)S_y^2}{\left(q_4\mu_1^2+t_5\mu_1+P\right)} \tag{5.11}$$

where

$$\mu_1=\frac{-p_2\pm\sqrt{p_2^2-4p_1p_3}}{2p_1},q_1=-\left(\frac{nP^2}{N^2}+\frac{PC}{N}\right),q_2=\left(\frac{PC}{n}+\frac{nP^2}{N^2}\right),q_4=\left(\frac{2nP}{N}+Q\right),$$
$p_1=q_1t_5-q_2q_4, p_2=2Pq_1-2t_3q_4$ and $p_3=Pq_2-t_3t_5$.

REMARK 5.2.

μ_1 being the optimum value of μ, so for certain situation there may be two values of μ_1 so while choosing μ_1 we must remember $0\le\mu_1\le1$. However, if both the values of μ_1 are admissible, we choose the minimum of the two values as μ_1.

From (5.5) and (5.11) we have

$$\mathrm{Var}\left(\mathrm{d}^{**}\right)_{\mathrm{opt}^{*}}-\mathrm{Var}\left(\mathrm{t}\right)_{\mathrm{opt}^{*}}=\left\{D_3\mathrm{Var}\left(\mathrm{d}^{*}\right)\right\}/D^{**} \tag{5.12}$$

where

$$D_3=Var\left(d_{2(2)}\right)-Var\left(d_{2(0)}\right),$$

$$D^{**}=\left\{\mathrm{Var}\left(\mathrm{d}^{*}\right)+\mathrm{Var}\left(\mathrm{d}_{2(2)}\right)\right\}\left\{\mathrm{Var}\left(\mathrm{d}^{*}\right)+\mathrm{Var}\left(\mathrm{d}_{2(0)}\right)\right\}.$$

It can be shown that

$$D_3=\left(\frac{1}{m}-\frac{1}{n}\right)\frac{\left(\rho_{yz}\rho_{zx}-\rho_{yx}\rho_{xz}^2\right)^2S_y^2}{\left(1-\rho_{xz}^2\right)}\ge0 \tag{5.13}$$

Thus from (5.12) and (5.13), we have following inequality as

$$\mathrm{Var}\left(\mathrm{t}\right)_{\mathrm{opt}^{*}}\le\mathrm{Var}\left(\mathrm{d}^{**}\right)_{\mathrm{opt}^{*}} \tag{5.14}$$

Thus, it follows from (5.14) that the proposed estimator t is better than Singh and Priyanka (2006 b) estimator d^{**}.

REMARK 5.3.

It is to be mentioned that the estimators $d_{1(0)}, d_{2(0)}, d_{1(2)}, d_{2(2)}, d^*$ and hence $\mathrm{d}_{(2)}, \mathrm{d}_{(0)}, \mathrm{d}_{(1)}, \mathrm{t}$ and d^{**} defined by (2.7), (2.34), (4.1), (5.1) and (5.7) respectively utilize prior information on regression coefficients $\beta_{yx}, \beta_{yz}, \beta_{xz}$ and β_{zx}. However in many situations it may be happened that these regression coefficients $\beta_{yx}, \beta_{yz}, \beta_{xz}$ and β_{zx} are not known in advance and hence it restricts the practical utility of the estimators $d_{1(0)}, d_{2(0)}, d_{1(2)}, d_{2(2)}, d^*, \mathrm{d}_{(2)}, \mathrm{d}_{(0)}, \mathrm{d}_{(1)}, \mathrm{t}$ and d^{**}. In such situations it is advisable to use the estimators based on 'estimated values' of regression coefficients from the data available in hand. Thus one may define the following estimators:

$$\hat{d}_{1(0)} = \bar{y}_u^* + \hat{\beta}_{yz}\left(\bar{Z} - \bar{z}_u\right) = \hat{d}_{1(2)} \text{ (say)}, \tag{5.15}$$

$$\hat{d}_{2(0)} = \left[\bar{y}_m + \frac{\left(b_{yx} - b_{yz}b_{zx}\right)}{\left(1 - b_{xz}b_{zx}\right)}\left(\bar{x}_n - \bar{x}_m\right) - \frac{b_{xz}\left(b_{yx} - b_{yz}b_{zx}\right)}{\left(1 - b_{xz}b_{zx}\right)}\left(\bar{z}_n - \bar{z}_m\right) + b_{yz}\left(\bar{Z} - \bar{z}_m\right)\right], \tag{5.16}$$

$$\hat{d}_{2(2)} = \left[\bar{y}_m + b_{yx}\left(\bar{x}_n - \bar{x}_m\right) - b_{yx}b_{xz}\left(\bar{z}_n - \bar{z}_m\right) + b_{yz}\left(\bar{Z} - \bar{z}_m\right)\right], \tag{5.17}$$

$$\hat{d}^* = \bar{y}_u + b_{yz}\left(\bar{Z} - \bar{z}_u\right), \tag{5.18}$$

$$\hat{\mathrm{d}}_{(0)} = \phi^*\hat{\mathrm{d}}_{1(0)} + \left(1 - \phi^*\right)\hat{\mathrm{d}}_{2(0)}, \tag{5.19}$$

$$\hat{d}_{(1)} = \phi^*\hat{d}_{1(1)} + \left(1 - \phi^*\right)\hat{d}_{2(1)}, \tag{5.20}$$

$$\hat{d}_{(2)} = \phi^*\hat{d}_{1(2)} + \left(1 - \phi^*\right)\hat{d}_{2(2)}, \tag{5.21}$$

$$\hat{t} = \phi^* d^* + \left(1 - \phi^*\right)\hat{d}_{2(0)}, \tag{5.22}$$

$$\hat{d}^{**} = \psi^* d^* + \left(1 - \psi^*\right)\hat{d}_{2(2)}, \tag{5.23}$$

where $\hat{\beta}_{yz}=\frac{\hat{S}_{yz}}{\hat{s}_z^2}, \hat{s}_z^2=\frac{1}{(u-1)}\sum_{i=1}^{u}(z_i-\bar{z}_u)^2$ and the estimate $\hat{S}_{yz}$ is based on the available data under the given sampling design [see Khare and Srivastava (1995), p. 197],

$$b_{yx}=\frac{s_{yx}}{s_x^2}, b_{yz}=\frac{s_{yz}}{s_z^2}, b_{zx}=\frac{s_{xz}}{s_x^2}, b_{xz}=\frac{s_{xz}}{s_z^2}, s_{yx}=\frac{1}{(m-1)}\sum_{i=1}^{m}(y_i-\bar{y}_m)^2,$$

$$s_{yz}=\frac{1}{(m-1)}\sum_{i=1}^{m}(y_i-\bar{y}_m)(z_i-\bar{z}_m), s_{xz}=\frac{1}{(m-1)}\sum_{i=1}^{m}(x_i-\bar{x}_m)(z_i-\bar{z}_m),$$

$$s_x^2=\frac{1}{(m-1)}\sum_{i=1}^{m}(x_i-\bar{x}_m)^2, s_z^2=\frac{1}{(m-1)}\sum_{i=1}^{m}(z_i-\bar{z}_m)^2.$$

Cochran (1977,p.193-194) defined the regression estimators $\bar{y}_{lr}=\left[\bar{y}_n+\hat{\beta}_{yx}(\bar{X}-\bar{x})\right]$ and $\bar{y}_{lro}=\left[\bar{y}_n+\beta_{yx}(\bar{X}-\bar{x})\right]$; (where $\hat{\beta}_{yx}$ is the estimate of population regression coefficient of y on x based on n observations) and showed that if $\hat{\beta}_{yx}$ is the least squares estimate of β_{yx} then the first order variance of $\bar{y}_{lr}$ (MSE of $\bar{y}_{lr}$) is same as the variance of $\bar{y}_{lro}$ i.e. $MSE(\bar{y}_{lr})\cong Var(\bar{y}_{lro})$, [see Ahmed (1998)]. Thus the optimum mean squared errors of the combined estimators $\hat{d}_{(0)}, \hat{d}_{(1)}, \hat{d}_{(2)}, \hat{t}$ and $\hat{d}^{**}$ will be the same as the optimum variance of the combined estimators $d_{(0)}$, $d_{(1)}, d_{(2)}, t$ and d^{**} respectively.

6. Numerical illustration

To have tangible idea about the performance of the proposed estimator $d_{(0)}$ with respect to t under their respective optimality conditions is given by

$$L^*=\frac{Var(d_{(0)})_{opt^*}-Var(t)_{opt^*}}{Var(d_{(0)})_{opt^*}}\times 100 \tag{6.1}$$

where $Var(d_{(0)})_{opt^*}$ and $Var(t)_{opt^*}$ are given in (3.2) and (5.6) respectively.

The percentage loss in precision of $d_{(0)}$ and t for different choices of W, ρ_{yx}, f, n and N are indicated in tables 1and 2. Table 3 shows the percent relative loss in precision for different randomly chosen values of $\mu_0^*, \mu_1', \mathrm{W}, \mathrm{f}, \rho_{\mathrm{yz}}$ and ρ_{yx}.

Table 1. The percentage loss in precision of $d_{(0)}$ over t for different choices of $\mathrm{W}, \rho_{\mathrm{yx}}$. $(N = 3000, n = 300, f = 1.5)$.

$\rho_{yx} \rightarrow$		0.3			0.7			0.9		
W ↓	ρ_{yz} ↓	μ_0^*	μ_1'	L^*	μ_0^*	μ_1'	L^*	μ_0^*	μ_1'	L^*
0.2	0.3	0.89	0.51	7.42	0.42	0.57	4.49	0.65	0.69	5.11
	0.4	0.86	0.50	7.77	0.37	0.57	4.59	0.63	0.68	5.45
	0.5	0.86	0.50	8.56	0.26	0.56	4.48	0.61	0.67	5.96
	0.6	0.93	0.50	10.36	*	-	-	0.57	0.65	6.71
	0.7	*	-	-	*	-	-	0.51	0.63	7.78
0.4	0.3	*	-	-	0.28	0.57	7.02	0.61	0.69	8.97
	0.4	*	-	-	0.19	0.57	6.74	0.59	0.68	9.47
	0.5	*	-	-	*	-	-	0.56	0.67	10.16
	0.6	*	-	-	*	-	-	0.50	0.65	11.08
	0.7	*	-	-	*	-	-	0.39	0.63	12.09
0.6	0.3	*	-	-	0.16	0.57	8.29	0.57	0.69	11.96
	0.4	*	-	-	0.03	0.57	7.43	0.55	0.68	12.49
	0.5	*	-	-	*	-	-	0.51	0.67	13.20
	0.6	*	-	-	*	-	-	0.44	0.65	14.04
	0.7	*	-	-	*	-	-	0.30	0.63	14.54
0.8	0.3	*	-	-	0.03	0.57	8.75	0.54	0.69	14.32
	0.4	*	-	-	*	-	-	0.51	0.68	14.83
	0.5	*	-	-	*	-	-	0.46	0.67	15.47
	0.6	*	-	-	*	-	-	0.38	0.65	16.07
	0.7	*	-	-	*	-	-	0.21	0.63	15.91

** Indicates μ_0^* does not exist.*

Table 2. The percentage loss in precision of $d_{(0)}$ over t for different choices of W, ρ_{yx}. $(N = 3000, n = 300, f = 2.5)$.

$\rho_{yx} \rightarrow$		0.3			0.7			0.9		
W ↓	$\rho_{yz} \downarrow$	μ_0^*	μ_1'	L^*	μ_0^*	μ_1'	L^*	μ_0^*	μ_1'	L^*
0.2	0.3	*	-	-	0.16	0.57	8.29	0.57	0.69	11.96
	0.4	*	-	-	0.03	0.57	7.43	0.55	0.68	12.49
	0.5	*	-	-	*	-	-	0.51	0.67	13.21
	0.6	*	-	-	*	-	-	0.44	0.65	14.04
	0.7	*	-	-	*	-	-	0.30	0.63	14.54
0.4	0.3	*	-	-	*	-	-	0.48	0.69	17.71
	0.4	*	-	-	*	-	-	0.44	0.68	18.09
	0.5	*	-	-	*	-	-	0.38	0.67	18.46
	0.6	*	-	-	*	-	-	0.27	0.65	18.48
	0.7	*	-	-	*	-	-	0.04	0.63	16.92
0.6	0.3	*	-	-	*	-	-	0.39	0.69	20.81
	0.4	*	-	-	*	-	-	0.34	0.68	20.92
	0.5	*	-	-	*	-	-	0.26	0.67	20.80
	0.6	*	-	-	*	-	-	0.12	0.65	19.92
	0.7	*	-	-	*	-	-	*	-	-
0.8	0.3	*	-	-	*	-	-	0.31	0.69	22.57
	0.4	*	-	-	*	-	-	0.25	0.68	22.41
	0.5	*	-	-	*	-	-	0.15	0.67	21.84
	0.6	*	-	-	*	-	-	*	-	-
	0.7	*	-	-	*	-	-	*	-	-

** Indicates μ_0^* does not exist.*

Table 3. The percentage relative loss in precision for different randomly chosen values of $\mu_0^*, \mu_1', \mathrm{W}, \mathrm{f}, \rho_{\mathrm{yz}}$ and ρ_{yx} . $(N = 3000, n = 300)$.

		$f = 1.5, \mu_0^* = 0.1, \mu_1' = 0.1$			$f = 1.5, \mu_0^* = 0.3, \mu_1' = 0.3$			$f = 1.5, \mu_0^* = 0.5, \mu_1' = 0.5$		
$\rho_{yx} \rightarrow$		0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
W	$\rho_{yz} \downarrow$	L^*	L^*	L^*	L^*	L^*	L^*	L^*	L^*	L^*
0.2	0.5	1.09	1.06	1.02	3.43	3.22	2.87	5.97	5.51	4.64
	0.6	1.26	1.23	1.17	3.99	3.77	3.34	6.97	6.49	5.43
	0.7	1.53	1.51	1.43	4.89	4.70	4.13	8.55	8.15	6.80
	0.8	1.99	2.02	1.92	6.26	6.44	5.67	10.84	11.23	9.51
	0.9	2.34	3.11	3.14	-	9.43	9.71	-	16.10	16.71
0.4	0.5	1.94	1.89	1.82	6.12	5.74	5.12	10.63	9.81	8.26
	0.6	2.21	2.16	2.06	7.02	6.63	5.86	12.22	11.38	9.52
	0.7	2.63	2.58	2.46	8.38	8.05	7.07	14.61	13.93	11.62
	0.8	3.28	3.32	3.16	10.23	10.54	9.28	17.69	18.31	15.51
	0.9	3.47	4.62	4.67	-	13.95	14.36	-	23.74	24.63
0.6	0.5	2.64	2.57	2.47	8.28	7.78	6.93	14.37	13.26	11.16
	0.6	2.96	2.89	2.76	9.38	8.86	7.83	16.31	15.19	12.71
	0.7	3.45	3.39	3.23	10.99	10.56	9.26	19.12	18.23	15.20
	0.8	4.17	4.22	4.01	12.99	13.37	11.77	22.39	23.19	19.65
	0.9	4.14	5.50	5.57	-	16.61	17.09	-	28.19	29.26
0.8	0.5	3.21	3.13	2.99	10.07	9.45	8.42	17.44	16.09	13.55
	0.6	3.57	3.49	3.32	11.29	10.66	9.42	19.59	18.25	15.27
	0.7	4.09	4.03	3.83	13.02	12.51	10.97	22.61	21.56	17.98
	0.8	4.83	4.89	4.64	15.01	15.45	13.59	25.84	26.76	22.67
	0.9	4.58	6.09	6.16	-	18.36	18.89	-	31.12	32.29

7 Conclusion

It is observed from tables 1,2,3 that for all cases the relative percentage loss in precision is observed wherever the optimum value of μ exists when non-response is taken into account at current occasion. Table 3 exhibits that the loss in precision decreases for different values of ρ_{yx} and increases for various values of ρ_{yz} . But for $(\rho_{yz} = 0.9)$ the loss in precision increases with the increase in the value of ρ_{yx} and for $(\rho_{yz} = 0.8)$, no particular pattern is available. For fixed values of $f, \mu_0^*, \mu_1', \rho_{yz}$ and ρ_{yx} the values of L^* increases with the increase in the value of W. From table 1 and 2 it is observed that for fixed values of $\mathrm{f}, \mathrm{W}, \rho_{\mathrm{yz}}$ and ρ_{yx}, the loss in precision increases with increase in the values of μ_0^* and μ_1' i.e., more the fraction of sample to be drawn at current occasion more loss in precision is observed due to non-response. From the tables it is obvious that loss is seen due to the presence of non-response at current occasion, but the formulation of estimator is such that the loss is very marginal. Hence in the presence of non-response the performance of the proposed estimator is well, so it may be recommended for its use in practice.

REFERENCES

AHMED, M. S. (1998): *A note on regression-type estimators using multiple auxiliary information*. Austral. and New Zealand Jour. Statist., 40,(3), 373—376.

BIRADAR, R. S. and SINGH, H. P. (2001): *Successive sampling using auxiliary information on both occasions*. Cal. Stat. Assoc. Bull., 51, (203-204), 243—251.

CHATURVEDI, D. K. and TRIPATHI, T. P. (1983): *Estimation of population ratio on two occasions using multivariate auxiliary information*. Jour. Ind. Statist. Assoc.,21, 113—120.

CHOUDHARY, R. K., BATHAL, H. V. L. and SUD, U.C (2004): *On Non-response in Sampling on Two Occasions.* Jour. Ind. Soc. Agril.Statist., 58(3) 331—343.

COCHRAN, W. G. (1977): Sampling Techniques, 3rd edition, John Wiley & Sons, New York, p. 193-194.

DAS, A. K. (1982): *Estimation of population ratio on two occasions*. Jour Ind. Soc. Agril. Statist., 34, 1—9.

ECKLER, A. R. (1955): Rotation Sampling. Ann. Math. Statist., 26, 664—685.

FABIAN, C. O. and HYUNSHIK, L. (2000): *Double Sampling for ratio and regression estimation with sub-sampling the non-respondents*. Survey Methodology, 26 (2), 183—188.

FENG, S. and ZOU, G. (1997): *Sample rotation method with auxiliary variable.*Commun. Statist. Theo-Meth., 26, 6, 1497-1509.

HANSEN, M. H. and HURWITZ, W. N. (1946): *The problem of the non-response in sample surveys*. Jour. Amer. Statist. Assoc., 41, 517—529.

JESSEN, R. J. (1942): *Statistical investigation of a sample survey for obtaining farm facts*. Iowa Agricultural Experiment Station Road Bulletin No. 304, Ames, USA, 1—104.

KHARE, B.B and SRIVASTAVA, S. (1993). *Estimation of population mean using auxiliary character in presence of non-response.* Nat. Acad. Sc. Letters, India, 16(3), 111—114.

KHARE, B.B and SRIVASTAVA, S. (1995). *Study of conventional and alternative two- phase sampling ratio, product and regression estimators in presence of non-response.* Proc. Nat. Acad. Sci. India, 65(A), II, 195—203.

KHARE, B.B and SRIVASTAVA, S. (1997). *Transformed ratio type estimators for the population mean in the presence of non-response*. Comm. Statist. — Theory Methods, 26(7), 1779—1791.

PATTERSON, H. D. (1950): *Sampling on successive occasions with partial replacement of units*. Jour. Royal Statist. Assoc., Scr. B, 12, 241—255.

OKAFOR, F. C. (2001): *Treatment of non-response in successive sampling.* Statistica,. LXI,(2), 195—204.

OKAFOR, F. C. and LEE, H. (2000): *Double sampling for ratio and regression estimation with sub sampling the non-respondent.* Survey Methodology, 26, (2), 183—188.

RAO, J. N. K. and GRAHAM, J. E. (1964): *Rotation design for sampling on repeated occasions.* Jour. Amer. Statist. Assoc., 59, 492—509.

RAO, P. S. R. S. (1986). *Ratio estimation with sub sampling the non-respondents*. Survey Methodology, 12(2), 217—230.

SEN, A. R. (1971): *Successive sampling with two auxiliary variables*. Sankhya, Ser. B, 33, 371—378.

SEN, A. R. (1972): *Successive sampling with* $\mathrm{p}(\mathrm{p} \geq 2)$. *auxiliary variables*. Ann. Math. Statist., 43, 2031—2034.

SEN, A. R. (1973): *Theory and application of sampling on repeated occasions with several auxiliary variables*. Biometrics, 29, 381—385.

SINGH, V. K., and SINGH, G. N. (1991): *Chain-type regression estimators with two auxiliary variables under double sampling scheme*. Merton, 49, 279—289.

SINGH, G. N. and SINGH, V. K. (2001): *On the use of auxiliary information in successive sampling*. Jour. Ind. Soc. Agri. Statist., 54(1), 1—12.

SINGH, G. N. (2005): *On the use of chain-type ratio estimator in successive sampling.* Statistics in Transition, 7(1), 21—26.

SINGH, G. N. and PRIYANKA, K. (2006 a): *On the use of chain-type ratio to difference estimator in successive sampling*. IJAMAS,(5),,41—49.

SINGH, G. N. and PRIYANKA, K. (2006 b): *Search of good rotation patterns to improve the precision of the estimates at current occasion*. Accepted in Communications in Statistics-Theory and Methods.

SINGH, G. N. and PRIYANKA, K. (2007): *Effect of non-response on current occasion in search of good rotation patterns on successive occasions*. Statistics in Transition-new series,8,(2), 273—292.

SINGH, H. P. and VISHWAKARMA, G., K. (2007): *A general class of estimators in successive sampling*. METRON, LXV, (2), 201—227.

TABASUM, R. and KHAN, I.A. (2004). *Double sampling for ratio estimation with non-response.* Jour. Ind. Soc. Agril. Statist., 58, (3), 300—306.

TABASUM, R. and KHAN, I.A. (2006). *Double Sampling Ratio Estimator for the Population Mean in Presence of Non-Response.* Assam Statist. Review, 20, (1), 73—83.

TIKKIWAL, B. D. (1953): Theory of successive sapling. Unpublished Diploma thesis submitted to ICAR, New Delhi.

YATES, F. (1949): *Sampling methods for censuses and surveys*. Charles Griffin and Co., London.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 297—310

FIGARCH MODELS AND LONG MEMORY

Henryk Gurgul[1], Tomasz Wójtowicz[2]

1. Introduction

Stock market researchers usually concentrate on stock prices and their behavior over time. The current stock price reflects investor opinion about the future development of a firm. New pieces of public information, or information which is privately acquired, are the main source of price movement. This is due to investors adjusting their expectations about prices in the light of currently available information.

It is widely accepted that stock price returns are heteroscedastic. Thus in empirical applications to finance it is common to treat stock returns as a GARCH process. However, squared stock returns that describe return volatility are widely known to display high and lagged autocorrelation and thus can be described as long-memory processes, as in, for example, an ARFIMA model. On the other hand, to properly model long memory in conditional variance, the class of FIGARCH models seems to be more adequate in modeling stock return series.

Formally, a FIGARCH model for returns can be viewed as an ARFIMA model for squared returns. Hence, there exist two models describing the long memory of squared returns. In both, the classical ARFIMA and FIGARCH model the long memory parameter, *d*, appears. Moreover, in both classes of models the fractional differencing operator $(1\text{-}L)^d$ is applied to squared returns[3]. Thus, a natural question arises: does the application of these models make any difference to the estimation of the long memory of squared returns? Is there any difference in the estimators of parameter *d*? If there is, what relation between both estimators can be established? How the long memory of conditional variance and squared returns are connected? In this paper we instance the problem in the case of the returns of stocks quoted on the Warsaw Stock Exchange. The rest of the paper is organized as follows. The next section briefly reviews the concept of long

[1] Faculty of Management, University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland; E–mail: h.gurgul@neostrada.pl.

[2] Faculty of Management, University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland; E–mail: twojtow@agh.edu.pl.

[3] where *L* is the lag operator, i.e $Lx_t = x_{t-1}$

memory and estimating methods for the long-memory parameter. The data basis is characterized in section 3, while empirical results are presented in section 4. The final section concludes the paper.

2. Long memory

The concept of long memory is closely related to existence of persistence in the autocorrelation of the process. A covariance stationary stochastic process exhibits long memory with memory parameter d when its spectral density function $f(\lambda)$ satisfies:

$$f(\lambda) \sim c\lambda^{-2d} \text{ as } \lambda \to 0^+, \tag{1}$$

where c is a finite positive constant and the symbol "~" means that the ratio of the left- and right-hand sides tends toward one at the limit. When the process satisfies condition (1) and $d > 0$ its autocorrelation ρ_k dies out at a hyperbolic rate (Granger and Joyeux (1980), Hosking (1981), Beran (1994), i.e.

$$\rho_k \sim ck^{2d-1} \text{ as } k \to \infty, \tag{2}$$

where c is a finite constant. If $d > 0$ then the spectral density is unbounded near the origin, and the process exhibits long memory. When $d<0.5$ the process is also stationary. If $d = 0$ the spectral density is bounded at 0 and the process is called short memory.

The long memory of a process can be estimated by means of both parametric and semiparametric methods. One of the most popular classes of semiparametric estimators are the local Whittle estimators introduced by Künsch (1987), and developed by Robinson (1995) and Lobato (1999). In the univariate case the local Whittle long memory estimator $\hat{d}_{LW}$ is defined as a maximizer of the likelihood function:

$$Q(g,d) = -\frac{1}{m}\sum_{j=1}^{m}\left[\ln\left(g\lambda_j^{-2d}\right) + \frac{I(\lambda_j)}{g\lambda_j^{-2d}}\right], \tag{3}$$

where λ_j are the Fourier frequencies, $I(\lambda)$ is the periodogram of a given sample $x_1,...,x_T$ and the second parameter g in above function corresponds to constant c in formula (1). A local Whittle estimator uses only the m first values of the periodogram, where $m = m(T)$ is a bandwidth parameter. The local Whittle estimator is consistent for $d\in(-1/2, 1)$ and has an asymptotically normal limit distribution for $d\in(-1/2, 3/4)$ (see Velasco (1999) and Phillips and Shimotsu, (2004)):

$$\hat{d}_{LW} \sim N\left(d, \frac{1}{4m}\right) \tag{4}$$

For further modifications of the local Whittle estimator, see for example Shimotsu and Phillips (2002) or Andrews and Sun (2004).

The local Whittle estimator can be defined in the multivariate case. For *N*-dimensional process the corresponding (concentrating) likelihood function is:

$$Q(\mathbf{d}) = \frac{2}{m}\sum_{i=1}^{N} d_i \sum_{j=1}^{N} \ln\lambda_j - \ln\left|\hat{\mathbf{R}}(\mathbf{d})\right| \tag{5}$$

where

$$\hat{\mathbf{R}}(d) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{\Lambda}_j \operatorname{Re}\{\mathbf{I}(\lambda_j)\}\mathbf{\Lambda}_j \tag{6}$$

with $\mathbf{\Lambda}_j = diag\left(\lambda_j^{d_1},\ldots,\lambda_j^{d_N}\right)$ and a crossperiodogram matrix $\mathbf{I}(\lambda)$. The estimator $\hat{\mathbf{d}} = \left(\hat{d}_1,\ldots,\hat{d}_N\right)$ can be computed in two ways: by the numerical maximizing of (5) or using the two-step procedure proposed by Lobato (1999). As showed by Lobato (1999) the two-step estimator has the same asymptotic distribution as the QMLE estimator and under reasonable assumptions is normally distributed with parameters:

$$\hat{\mathbf{d}} \sim N\left(\mathbf{d}, \frac{1}{\sqrt{m}}\mathbf{E}^{-1}\right) \tag{7}$$

where $\mathbf{E} = 2\left(\mathbf{I}_N + \mathbf{R} \circ \mathbf{R}^{-1}\right)$ and $\circ$ denotes the Hadamard product of two matrices.

Based on these asymptotic properties a test for the null hypothesis of a set of *q* linear restrictions on **d** is available. Consider a **P** which is *q*×*N* matrix, *N*×1 vector **ρ** and the null hypothesis

$$H_0: \quad \mathbf{Pd} = \boldsymbol{\rho} \tag{8}$$

Then the test statistic:

$$m\left(\mathbf{P}\hat{\mathbf{d}} - \boldsymbol{\rho}\right)^T \left(\mathbf{P}\hat{\mathbf{E}}^{-1}\mathbf{P}^T\right)^{-1}\left(\mathbf{P}\hat{\mathbf{d}} - \boldsymbol{\rho}\right) \tag{9}$$

is asymptotically χ_q^2 distributed under the null hypothesis. In case of testing for the existence of a common long-memory parameter, **ρ** is a vector of zeroes and $\mathbf{P} = \left(\mathbf{I}_{N-1} \vdots \mathbf{0}\right) - \left(\mathbf{0} \vdots \mathbf{I}_{N-1}\right)$ is a (*N*-1)×*N* matrix.

Where a common long memory of processes exists, the problem of a common long run, i.e. fractional cointegration, could be considered. We describe the special but simplest case of the cointegration of two processes. Several definitions of fractional cointegration can be found in the literature (see Robinson and Yajima (2002)). The most common definition is as follows. We say that two fractionally integrated series x_t and y_t are cointegrated of order d if:

- x_t and y_t share the same long memory, i.e. $d_x = d_y$;
- there exists a constant β such that the process $\varepsilon_t = y_t - \beta x_t$ has the long-memory parameter $d < d_y$.

An estimation of parameter β can be done, among other ways, by means of the frequency domain least squares (FDLS) method. Alternatively, based on local Whittle estimation methods, another way to check the existence of fractional cointegration between x_t and y_t is to test the necessary condition that the coherency between both series is 1 at zero frequency. Given the estimate matrix $\hat{\mathbf{R}}$, the squared coherency estimate is expressed by:

$$\left[\hat{H}_{xy}(0)\right]^2 = \frac{\hat{\mathbf{R}}_{xy}^2}{\hat{\mathbf{R}}_{xx}\hat{\mathbf{R}}_{yy}} \tag{10}$$

The most well-known class of long-memory processes are autoregressive fractionally integrated moving average (ARFIMA) processes introduced into econometrics by Granger and Joyeux (1980).

Process y_t is said to be ARFIMA(p, d, q) if:

$$\phi(L)(1-L)^d(y_t - \mu) = \theta(L)e_t, \tag{11}$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ and $\theta(L) = 1 - \theta_1 L - \ldots - \theta_q L^q$ are lag polynomials of order p and q respectively, in the lag operator L with roots outside the unit circle, e_t is i.i.d. with zero mean and variance σ^2, and the fractional differencing operator $(1\text{-}L)^d$ is defined by binomial expansion:

$$(1-L)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} L^j , \tag{12}$$

If $d > -0.5$ the ARFIMA process is invertible and possesses a linear Wold representation and if $d < 0.5$ it is covariance stationary. Thus, if $0 < d < 0.5$ the process is stationary and exhibits long memory. The parameters of the ARFIMA model can be estimated by the maximum likelihood method. Sowell (1992) proved that the exact maximum likelihood estimator (EML) is consistent and asymptotically normal. Other properties of MLE and methods of solving some computational problems are discussed in details by Sowell (1992) and Doornik and Ooms (2003).

The class of ARFIMA models is designed to model long memory processes with assumed constant variance. However, in practice it is quite common to observe the long memory phenomenon in squared residuals, e.g. squared returns. When a changing conditional variance of the process is allowed, the most popular models are modifications of GARCH or EGARCH models – the fractionally integrated GARCH (FIGARCH) or the fractionally integrated EGARCH called FIEGARCH, respectively. The EGARCH (FIEGARCH) models additionally take into account asymmetric responses of stock market to bad and good news.

In order to construct a FIGARCH model for conditional variance consider first a discrete time real-valued stochastic process, $\{\varepsilon_t\}$,

$$\varepsilon_t = z_t \sigma_t \ , \tag{13}$$

where z_t are i.i.d. with $E_{t-1}(z_t) = 0$ and $VAR_{t-1}(z_t) = 1$, while σ_t^2 stands for time-varying conditional variance.

In the classic ARCH(q) model of Engle (1982), the conditional variance σ_t^2 is a linear function of the lagged squared innovations $\{\varepsilon_t\}$, i.e. it is defined by the formula:

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 , \tag{14}$$

where $\alpha(L) = \alpha_1 L + \alpha_2 L^2 + \cdots + \alpha_q L^q$ is a lag polynomial of order q.

The GARCH(p,q) specification of Bollerslev (1986) provides a more flexible structure of conditional variance:

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 , \tag{15}$$

where additionally $\beta(L) = \beta_1 L + \beta_2 L^2 + \cdots + \beta_p L^p$.

The GARCH(p,q) process $\{\varepsilon_t\}$ can be conventionally rewritten as an ARMA(m,q) process in squared residuals ε_t^2 as

$$[1 - \alpha(L) - \beta(L)]\varepsilon_t^2 = \omega + [1 - \beta(L)]\upsilon_t , \tag{16}$$

where $m = \max\{p, q\}$and terms$\upsilon_t = \varepsilon_t^2 - \sigma_t^2$ can be viewed as innovations since they have zero mean and are serially uncorrelated. Hence, for the stability and covariance stationarity of the process, all roots of $1 - \alpha(L) - \beta(L)$ and $1 - \beta(L)$ are assumed to lie outside the unit circle.

When the autoregressive lag polynomial,$1 - \alpha(L) - \beta(L)$, contains a unit root, the GARCH(p,q) process becomes integrated in the variance (Engle and Bollerslev (1986)). The corresponding IGARCH(p,q) process can be written as an ARIMA model

$$\varphi(L)(1 - L)\varepsilon_t^2 = \omega + [1 - \beta(L)]\upsilon_t , \tag{17}$$

where φ(L) is a lag polynomial of order m-1.

Hence, in order to include long-range dependence in squared residuals $\{\varepsilon_t^2\}$and by analogy to the ARFIMA class of models, the Fractionally Integrated GARCH, or the FIGARCH, model can be defined by replacing the difference operator (1-L) of the IGARCH model by the fractional differencing operator (1-L)^d defined analogously to the ARFIMA case.

Formally, according to Baillie, Bollerslev and Mikkelsen (1996) (hereafter BBM), the FIGARCH(p,d,q) process is defined by the formula:

$$\varphi(L)(1-L)^d \varepsilon_t^2 = \omega + [1-\beta(L)]\upsilon_t, \tag{18}$$

where0<d<1 and all the roots of φ(L) and 1-β(L) lie outside the unit circle.

There are some deficiencies in the above specification of the FIGARCH(p,d,q) process. As pointed out by Chung (1999) its connections and analogies with the ARFIMA class of processes of the conditional mean are not perfect. In particular, the constant term ω is structurally different from the μ in ARFIMA models: the fractional differencing operator applies to μ while it does not apply to ω. Moreover, given the unconditional variance σ^2 the parameter ω in BBM parameterization should be zero irrespective of the value of σ^2.

In order to avoid these drawbacks of the BBM parameterization of the FIGARCH model, Chung (1999) proposed a slightly different formula:

$$\varphi(L)(1-L)^d \left(\varepsilon_t^2 - \sigma^2\right) = \omega + [1-\beta(L)]\upsilon_t, \tag{19}$$

where σ^2 is the unconditional variance of ε_t.

3. Data

The data consists of the daily returns and squared return series for 74 stocks continuously traded on Warsaw Stock Exchange in the whole period from January 2001 to October 2007. Additionally, data concerning the two main stock indices, WIG and WIG20, are included. The considered period was chosen for two reasons. The first one is the introduction in November 2000 of the new Warsaw Stock Exchange Trading System (WARSET), which has stimulated internet trading. Secondly, it is a sufficiently long period to examine the long memory properties of time series.

Continuously compounded stock returns are calculated from daily reference and closing prices announced in the Official Quotation of the Warsaw Stock Exchange. Some descriptive statistics of the considered time series are provided in Table 1.

Table 1. Descriptive statistics for return and squared return series.

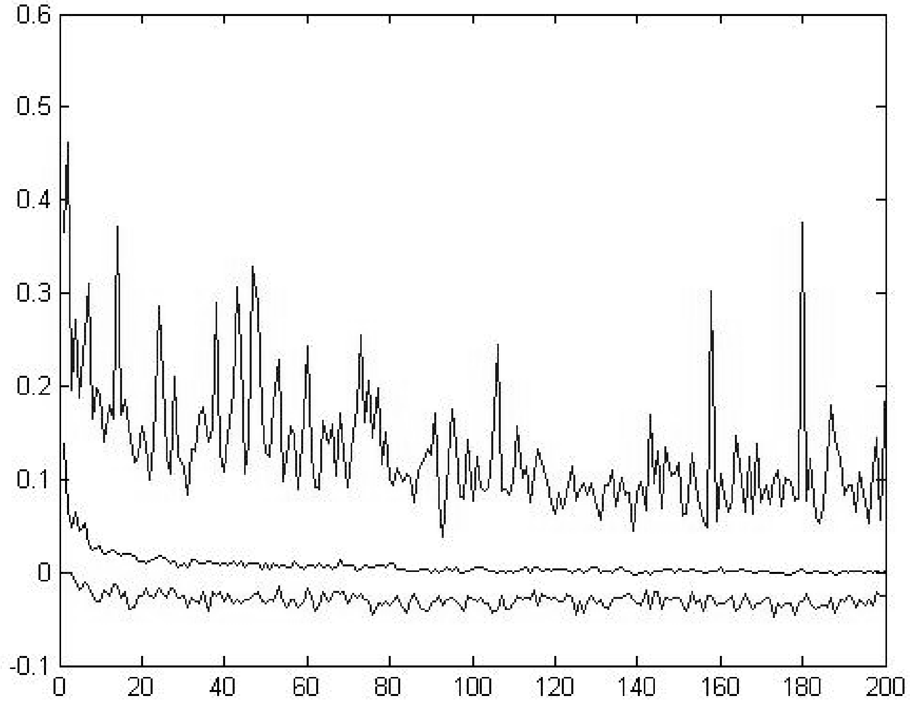
	Returns			
	Mean	Std. dev.	Skewness	Kurtosis
Min	-0,0017	0,012	-24,150	3,870
1st Quartile	0,0001	0,021	-0,223	7,309
Median	0,0006	0,028	0,205	11,365
3rd Quartile	0,0010	0,034	0,825	22,529
Max	0,0027	0,060	5,420	842,03
	Squared returns			
Min	0,0001	0,0003	3,13	15,59
1st Quartile	0,0005	0,0012	7,56	94,39
Median	0,0008	0,0024	11,24	191,29
3rd Quartile	0,0012	0,0056	18,11	466,03
Max	0,0036	0,0618	41,02	1685,5

4. Empirical results

We begin our study with an investigation of the autocorrelation properties of squared return series. As an example, Figure 1 plots the maximum, median and minimum autocorrelations of the considered volatility series. It suggests for most volatility series that the autocorrelations are of hyperbolic rather than exponential decay. This is a typical property of long memory time series.

The existence of persistent serial dependence in squared returns series is confirmed by the standard Ljung-Box test of serial autocorrelation as performed for all samples. The test statistics Q(20), not shown here, reject the null hypothesis of no serial autocorrelation at a 5% significance level in 64 cases. Volatility series with an insignificant Ljung-Box statistic also have the lowest variation, skewness and kurtosis. The above-mentioned autocorrelation properties indicate the existence of long memory in the considered squared return time series.

Figure 1. Returns autocorrelations. From top to bottom there are: maximum, median and minimum autocorrelation coefficients.



The rest of this section is concerned with a detailed analysis of the memory properties of volatility series. The degree of autocorrelation persistence of a time series is characterized by memory parameter, *d*. We estimate this by means of both semiparametric and parametric methods. The results of long memory estimation are collected in Table 2. As a first we use the semiparametric local Whittle estimation method. The advantage of this estimator is that it allows for quite general forms of short range dynamics, while parametric estimators are sensitive to any misspecifications of short run behavior. In a local Whittle estimation procedure the bandwidth parameter *m* must be specified. According to the results of Henry and Robinson (1996), *m* is chosen as $T^{0.65}$, where *T* is the sample length.

Table 2. Comparison of long memory estimates of squared returns and conditional variance.

Security	LW	ARFIMA(1,d,1)	FIGARCH(1,d,1)	LWE of conditional variance	coherency
01N	0,08	0,07	0,36	0,22	0,99
04N	0,25	0,43	0,38	0,56	0,94
05N	0,07	0,06	0,58	0,40	0,80
06N	0,22	0,25	0,57	0,51	0,94
08N	0,01	0,00	0,00	0,26	0,96
10N	0,13	0,11	0,23	0,31	0,97
13N	0,06	0,05	0,46	0,43	0,73
14N	0,19	0,40	0,46	0,43	0,89
ABG	0,05	0,04	0,13	0,27	0,97
ACP	0,10	0,07	0,36	0,41	0,90
AGO	0,24	0,28	0,36	0,64	0,81
AMC	0,16	0,17	0,12	0,37	0,97
ATS	0,13	0,43	0,45	0,21	0,98
BBC	0,10	0,08	0,49	0,44	0,91
BBD	0,12	-0,01	0,48	0,33	0,94
BDX	0,17	0,31	0,17	0,41	0,95
BHW	0,29	0,26	0,30	0,46	0,98
BPH	0,11	0,09	0,15	0,37	0,94
BRE	0,29	0,29	0,37	0,83	0,76
BRS	0,19	0,12	0,17	0,34	0,98
BSK	0,15	0,17	0,35	0,34	0,97
BZW	0,23	0,28	0,35	0,66	0,78
CMR	0,20	0,14	0,53	0,59	0,85
CRS	0,10	0,09	0,45	0,36	0,86
CSS	0,08	0,06	0,37	0,41	0,89
DBC	0,11	0,07	0,10	0,27	0,98
ECH	0,00	0,00	0,70	0,75	0,01
ELB	0,15	0,16	0,35	0,32	0,96
ELE	0,22	0,21	0,25	0,44	0,97
ELZ	0,10	0,08	0,52	0,26	0,91
EMF	0,14	0,09	0,72	0,35	0,95
EPD	0,15	0,13	0,26	0,42	0,95
FCL	0,00	-0,01	0,84	0,15	0,80
IGR	0,08	0,06	0,36	0,18	0,99
IPX	0,10	0,09	0,53	0,23	0,95
IRE	0,14	0,14	0,25	0,32	0,96
JPR	0,14	0,13	0,58	0,37	0,90
JTZ	0,14	0,09	0,14	0,41	0,92

Security	LW	ARFIMA(1,d,1)	FIGARCH(1,d,1)	LWE of conditional variance	coherency
KGH	0,28	0,31	0,42	0,88	0,31
KGN	0,19	0,23	0,22	0,53	0,86
KRB	-0,02	0,11	0,15	0,45	0,89
KTY	0,13	0,26	0,36	0,41	0,66
KZS	0,22	0,25	0,30	0,41	0,97
LTX	0,25	0,35	0,48	0,74	0,78
MCL	0,10	0,08	0,41	0,17	0,98
MDS	0,22	0,17	0,74	0,65	0,89
MIL	0,13	0,10	0,48	0,44	0,93
MNC	0,07	0,03	-0,06	0,10	1,00
MNI	0,21	0,19	0,51	0,33	0,98
MPP	0,02	0,03	0,94	0,70	0,61
MSO	0,08	0,03	0,64	0,36	0,68
MSW	0,13	0,18	0,24	0,31	0,98
MSX	0,18	0,24	0,22	0,40	0,97
MSZ	0,18	0,20	0,24	0,44	0,92
NET	0,28	0,36	0,63	0,50	0,95
ORB	0,24	0,26	0,46	0,78	0,78
PEO	0,10	0,07	0,24	0,81	0,66
PGF	0,18	0,26	0,43	1,20	0,40
PKM	0,22	0,12	0,32	0,71	0,73
PKN	0,20	0,22	0,24	0,66	0,48
PLE	0,11	0,07	0,15	0,31	0,98
PSP	0,23	0,09	0,21	0,43	0,97
PUE	0,14	0,10	0,25	0,37	0,95
RFK	0,12	0,12	0,31	0,42	0,94
SGN	0,57	-0,19	0,48	1,12	0,77
SME	0,16	0,07	0,22	0,34	0,98
STF	0,16	0,13	0,04	0,24	0,99
STP	0,11	0,20	0,42	0,27	0,98
STX	0,20	0,11	0,17	0,33	0,99
TIM	0,24	0,25	0,34	0,63	0,63
TLX	0,19	0,17	0,27	0,36	0,98
TPS	0,29	0,33	0,45	0,82	0,32
WLB	0,09	0,09	0,23	0,47	0,73
ZWC	0,24	0,24	0,65	0,55	0,90
WIG	0,25	0,31	0,54	0,73	0,08
WIG20	0,25	0,33	0,56	0,73	0,00

The estimated long memory parameter (column LW in Table 2) ranges form -0.022 (KRB) to 0.573 (SGN). It should be noted here that, except for SGN, all memory estimates are lower than 0.3, which means they are in the stationary region. Moreover, a closer look at the squared returns of SGN reveals that incomparably high memory estimates are strongly influenced by the last 50 terms of the sequence. Without them the local Whittle estimator of the memory parameter is close to 0.23. Long memory estimates of the squared returns of the stock indices of WIG and WIG20 are close to 0.25 and are one of the highest.

Based on the asymptotical normality of local Whittle estimators we conduct a test for the existence of long memory. The hypotheses are:

$$H_0: d=0$$

$$H_1: d>0,$$

which is equivalent to a one sided significance test. The long memory parameter is significantly greater than 0 for 67 samples. All of them except MNC indicate also a significant Ljung-Box statistic. On the other hand there are four stocks with insignificant serial autocorrelation indicated by the Q(20) statistic and with a significant local Whittle long memory estimator. In these cases the estimators of *d* are between 0.08 and 0.1.

The above results indicate that the majority of the considered squared return series can be seen as stationary long memory processes. For the sake of economy we use in our analysis ARFIMA(1,d,1) models. The estimators of long memory parameter *d* in the ARFIMA models, collected in the third column of Table 2, and of the local Whittle estimators appear quite consistent with one another. The *t*-test of parameter equality rejects the null hypothesis only for seven samples (04N, 14N, ATS, BDX, KRB, PSP, SGN) indicating a significant difference between long memory estimates. Hence, we can assume that both long memory estimation methods give quite similar results for the squared returns series.

On the other hand, squared returns are a measure of volatility and hence they are proxies for the conditional variance of return series. Due to the presence of a time varying conditional variance, it is common to use GARCH models to describe stock return series. To verify the assumption about the presence of the ARCH effect we perform an Engle test for lag length one. The null hypothesis of homoscedasticity of a return series is rejected in 64 cases. This means that a majority of the considered stock return display time varying conditional variance. Here two interesting things should be noted. First, one of the samples for which the null hypothesis is not rejected is the return series of the WSE main index WIG, where the *p*-value is equal to 0.135. Second, for seven series the lack of rejection of the null hypothesis in Engle test coincides with the lack of rejection of the null hypothesis of no serial autocorrelation in squared return series.

The tests performed above leave 62 samples of stock returns that display both serial dependence in squared returns and heteroscedasticity. Thus, FIGARCH models could be applied to describe their behavior more adequately – we will not consider the asymmetry properties of returns. As with ARFIMA we estimate one of the most popular and simple FIGARCH(1,d,1) models. To make a comparison between estimates of *d* in ARFIMA and FIGARCH models, we use a Chung (1999) parameterization. This is because in its ARFIMA form for squared returns the fractional differencing operator is applied to ω parameter while in BBM parameterization it is not. Moreover, the equal length of the AR and MA polynomials in the considered ARFIMA and FIGARCH models will enable us to compare their long memory estimates. Since we are mainly interested in memory properties, in Table 2, as with to ARFIMA models, only fractional differencing parameter estimates of FIGARCH models are presented. For 67 samples the differencing parameter estimates of the FIGARCH model are greater than those of ARFIMA models. The difference between estimates of long memory is significant for 38 return series.

In a FIGARCH model it is assumed that difference parameter *d* lies between 0 and 1 while in an ARFIMA model it is assumed to be lower than 0.5. Thus, it is interesting to check whether the estimates of *d* in FIGARCH models fulfill the assumption of ARFIMA models. In fact, 59 estimates of the difference parameter turn out to be lower than 0.5, whereas only 9 estimates are significantly greater than 0.5.

The above results about the differences between the memory estimates of ARFIMA and FIGARCH models contradict an assumption made in the definition of the FIGARCH process which was part of the basis for ARFIMA models of squared returns. If the assumption were correct, estimates would be close one to each other. The observed discrepancies have their source in the different assumptions of the models. In an ARFIMA model it is assumed that innovations are independent and normally distributed, while the main purpose of FIGARCH model is to incorporate the long memory of squared returns into conditional variance. Thus, its ARFIMA form is a rather formal representation.

Let us now have a look at the so-called "innovations" in the ARFIMA form of the FIGARCH model defined as the difference between squared returns and conditional variance. Together with the parameters of the model we estimated for each stock its conditional variance in FIGARCH model. For all samples the conditional variance series display longer memory than squared returns series[1] and for 23 stocks they have long memory estimates greater than 0.5. To compare memory of squared returns and conditional variance we use the multivariate local Whittle estimation procedure suggested by Lobato (1999). The obtained long memory estimates differ slightly from the univariate case but in general are in line with them. Based on the asymptotical normality of the estimates a Wald-type test of the existence of common long memory parameter is performed. The null hypothesis cannot be rejected only in one case (MNC). This means that a standard analysis of the fractional cointegration of squared returns and conditional variation cannot be conducted because its main assumption is the equality of the

[1] The memory parameter is estimated by means of the local Whittle method.

long memory of any considered time series. On the other hand, the high values of squared coherency, reported in the last column of the Table 2, indicate the presence of a common long memory factor in squared returns and conditional variance series.

5. Concluding remarks

Our analysis shows that when conditional variance is introduced and the FIGARCH model is considered, the long memory estimates of squared return series increase in comparison with the standard ARFIMA model of the same order. The estimates of the long memory of conditional variance are greater than estimators of memory of squared return series. Moreover, squared returns and conditional variance series do not share the same fractional differencing parameter. Thus, the assumption that innovations are not correlated in the FIGARCH model is frequently violated. On the other hand, examination of the coherency between squared returns and conditional variances indicates the presence of a long-run comovement of the considered time series.

REFERENCES

ANDREWS, D. W. K., SUN, Y. (2004): Adaptive local polynomial Whittle estimation of long-range dependence. *Econometrica,* vol. 72, pp. 569—614.

BAILLIE, R. T., BOLLERSLEV, T. – MIKKELSEN, H. O. (1996): Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics,* vol.74, pp. 3—30.

BERAN, J. A. (1994): *Statistics for Long-Memory Processes*. Chapman and Hall.

BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, vol. 31, pp. 307—327.

CHUNG, C.-F. (1999): Estimating the fractionally integrated GARCH model. *Working Paper*, National Taiwan University.

DOORNIK, J. A., OOMS, M. (2003): Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics and Data Analysis*, vol. 42, pp. 333—348.

ENGLE, R, (1982): Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, vol. 50, pp 987—1007.

ENGLE, R, BOLLERSLEV, T. (1986): Modeling the persistence of conditional variances. *Econometric Reviews*, vol. 5, pp. 1—50.

GRANGER, C. W. J., JOYEUX, R. (1980): An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, vol. 1, pp. 15—29.

HENRY, M., ROBINSON, P.M. (1996): Bandwidth choice in Gaussian semiparametric estimation of long range dependence, [in] Robinson, P. M., Rosenblatt, M. (eds.): *Athens Conference on Applied Probability and Time Series Analysis, Volume II: Time Series Analysis, In Memory of E. J. Hannan*, New York, Springer, pp. 220—232.

HOSKING, J. R. M. (1981): Fractional differencing. *Biometrika*, vol. 68, pp. 165—176.

KÜNSCH, H. R. (1987): Statistical aspects of self-similar processes. In: Prokhorov, Y. – Sazanov, V. V. (eds.): *Proceedings of the First World Congress of the Bernoulli Society. Utrecht*, VNU Science Press, pp.67—74.

LOBATO, I. N. (1999): A semiparametric two-step estimator in a multivariate long memory model. *Journal of Econometrics,* vol. 90, pp.129—153.

PHILLIPS, P. C. B. – SHIMOTSU, K. (2004): Local Whittle estimation in nonstationary and unit root cases. *Annals of Statistics,* vol. 34 (2), pp. 656—692.

ROBINSON, P. M. (1995): Gaussian semiparametric estimation of long range dependence. *Annals of Statistics,* vol.23, pp. 1630—1661.

ROBINSON, P. M. – YAJIMA, Y. (2002): Determination of cointegrating rank in fractional systems. *Journal of Econometrics,* vol. 106 (2), pp. 217—241.

SHIMOTSU, K. – PHILLIPS, P. C. B. (2002): Exact local Whittle estimation of fractional integration. *Cowles Foundation Discussion Paper* 1367.

SOWELL, F. B. (1992): Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics,* vol. 53, pp. 165—188.

VELASCO, C. (1999): Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis,* vol. 20, pp. 87—127.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 311—326

CONJOINT ANALYSIS WITHIN THE FIELD OF CUSTOMER SATISFACTION PROBLEMS – A MODEL OF COMPOSITE PRODUCT/SERVICE

Piotr Tarka[1]

ABSTRACT

The article describes how the benefits of conjoint analysis can be adapted to measuring performance criteria in the customer service area. It is also explained how a single composite model can be built, incorporating a wide range of customer key choice criteria, including service. Derived from a face to face personal interview, the data collected can be used to address key research objectives; including the measurement of the relative importance of service criteria versus other key choice criteria, the integration of customer perceptual data, the identification of market segments based on customer needs.

1. Introduction

Conjoint Analysis, since its inception in the early 1960's, has proved itself to be a reliable and useful methodology, and now has the status of a respected technique for product development in many markets. Moreover, with the development of computer assisted interviewing, it is also currently undergoing a reawakening of interest among practitioners.

Conjoint analysis has often been core to the better understanding of the relevant key drivers of demand. However, there are a number of problems with the approach that have now become apparent. One of these is the apparent distortion in computed utility values that arises in circumstances where global ‘macro’ variables are traded-off against more ‘micro’ topics. This can lead to dramatic underestimation of the overall contribution or importance of ‘macro’ issues. To address this concern, author discuss an approach known as ‘dual scaling’ for eliminating the bias.

Another drawback to the approach in customer service studies is the limited number of variables that can be addressed by a typical conjoint study. This makes

[1] University of Economics, Department of Marketing Research in Poznan, Al. Niepodległości 10, 60-967 Poland, POLAND and also Research International in Poznan; e-mail:piotr.tarka@op.pl.

it difficult to cover the large range of service topics typically examined in a customer satisfaction study. The paper argues that this limits the scope of both classical conjoint studies and current customer satisfaction approaches; since conjoint can typically handle service only at a global level, while customer satisfaction approaches do not measure the importance of service in the context of other (product) variables such as branding and pricing. The paper describes a way of merging data collected using both methods, and constructing an overall composite model containing both product and service variables.

2. The growth of attention in customer satisfaction

The past 10 years has seen the area of customer satisfaction taking an increasingly important role in marketing thinking. This has taken place, not just in the traditional arena of 'service' industries, but also in FMCG markets with a historic orientation towards the selling of 'products'. A new view concerning the relationships between product and service has emerged, and some writers speak of the incorporation of service as part of branding equity (King, 1991). This has led, in turn, to manufacturers and service industries examining afresh the frameworks of service expectation among managers and customers, and current perceptions of service delivery among those customers. Research initiatives, based on interviews with staff and customers alike, have become commonplace marketing tools; and formal monitors of 'objective' service delivery have become standard in many industries. The days of measuring customer satisfaction by the level of complaints received alone are long gone, and Total Quality Management (TQM) programmes are now widely implemented.

However, there is more to customer service than simply doing what we say we'll do — customer satisfaction is not product quality alone. In general, there is considerable evidence in the management consultancy world that TQM for its own sake is not sufficient. We also have to respond to demands that customers have, and configure the organisation to deliver them. Failure to understand this principle, or to act upon it effectively, has been held to explain (at least in part) the waning enthusiasm for TQM programs among Western companies — "an over-concentration of process and mechanics on the one hand coupled with an almost arrogant neglect of the buyer's needs, wants, and desires on the other" (Biel, 1992).

Moreover, many operators in this area have reported that the effect of improved customer service is greater in retaining existing customers than growing new ones, except in cases where the industry operating system has undergone a quantum change (e.g. the growth of direct phone-based insurance providers). Nevertheless, the paradigms that exist for categorising the various components of service delivery are complex, and organisations' responses to them varied. Some key workers in this field (Berry, Parasuraman, Zeithaml, 1984), approach the problem by identifying 'gaps' between expectation and delivery as the causes of customer dissatisfaction — they have listed at least five critical gaps in the service delivery process; but it has been noted that many firms address only some of these gaps, and these not fully (Biel, 1992).

3. Customer service and market research

The role of effective market research is to measure such gaps, and provide strategic tools for decision making. We would argue that when service problems are translated into research projects, they invariably require (preferably quantitative) answers to the questions:

- What are the *criteria* that customers use to evaluate my organisation?
- How *important* are they, in terms of generating a latent climate of satisfaction/ dissatisfaction, as well as stimulating behavioral change (i.e. choices)?
- How am I *perceived* by customers as delivering on those criteria, both absolutely, and relative to competitors?
- What are the attitudes of my *management/staff* — do they fully understand the reality of customer demand?

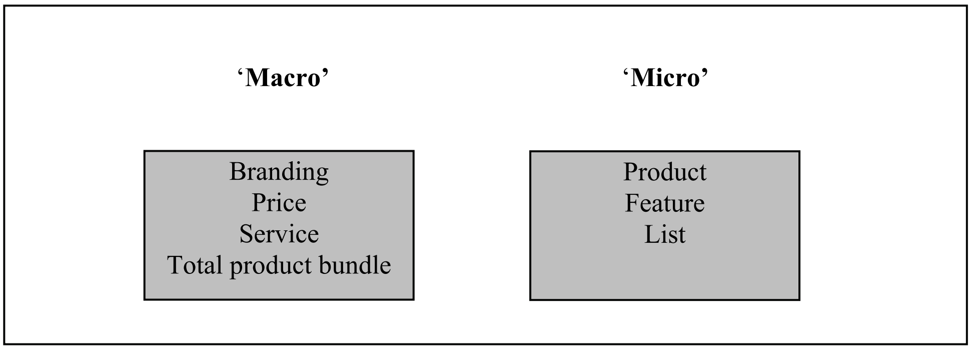
Moreover, our experience is that what clients really need is an understanding of the 'breakpoints' that exist with choice criteria, and a means of performing cost/benefit analysis on them. At the same time, service improvements need to be understood in a context relative to competition, and not just in isolation.

Unless research programmes have the ability to answer *all* of these topics, it might be safely said that any data emerging is incomplete. Moreover, these issues are inter-related in such a way that the full picture is only apparent when all aspects are considered. Thus, a firm that embarks on a quality tracking programme without some valid understanding of the importance or significance of the aspects being measured is getting only part of the story. Equally, a firm that concentrates on measuring the attractiveness of improving service dimensions does not obtain actionable information until the relative positions of competitors are also identified.

4. Identification and measurement of 'key decision criteria'

In order to be able to understand the determinants of consumer choice, of which customer service is but one, we need to understand the full range of variables involved, and their hierarchical relationships. One way of classifying choice criteria is to separate them into the 'macro' issues that transcend the product, and the 'micro' issues that represent the specification of the product itself.

Graph 1. Key decision criteria



'Macro' criteria — One of the most obvious criteria of consumer choice is that of the **branding** attached to the product or service. In market research, this is normally either the traditional notion of a manufacturer's brand (Marlboro, Persil, Sony etc.) or the name of an organisation providing service (British Airways, Taco Bell, Shell etc.). The characteristics of the branding process have lent these names an equity based upon their symbolic value, where the symbolism acts as a shorthand for other qualities such as dependability, personality, advertising recall etc.

Another macro issue must be **price** itself. This, after all, is a bottom line topic against which all the other components must be weighed. Price can be simple and unitary (e.g. the price tag attached to most grocery products), or can be complex with elements that trade off against each other. Examples of the latter might be many financial products. Another might be mobile phones, where at least three pricing elements (handset cost, monthly charge, and call rates) operate. It is against these topics (or topics like them) that **service** stacks up. In research terms, it is useful to keep service separate from issues of product specification (the 'micro' variables) since service is often a context within which products (possibly changing over time) are evaluated.

'Micro' product issues — In addition to the macro issues that operate at a global level, the specific characteristics of products must also be taken into account as variables that consumers use for decision making. A wide variety of criteria are encountered in product design, ranging from the emotional to the physical or functional. These have a role to play even in service industries — for example, type of aircraft for airlines, or regularity of statements in banking. However, these operate at the macro level as a *bundle of benefits*, which consumers trade-off against other macro topics. In our operationalisation of the model, individual aspects of the product profile never trade-off directly with macro variables. In fact, our experience (which has been replicated elsewhere) is that conventional conjoint models that treat these categories of variables as similar result in gross underestimation of the contribution of the most important ones (e.g. price). Therefore, models that aim to measure macro and micro variables simultaneously must adapt the basic conjoint approach. Firstly, however, a description of conjoint analysis is appropriate.

5. Conjoint analysis

There are a variety of ways of measuring the relative importance of choice criteria to consumers or customers. One of the more recent methods involves using a particular technique known as ‘conjoint analysis’. Conjoint analysis is a particular method of analysing experimental data in such a way that the contributions of different experimental conditions can be systematically separated and quantified. Its origins derive from a classical form of statistical analysis known as 'Analysis of Variance', and factorial designs. A full technical description of the method can be found elsewhere (Green, Srinivasan, 1990) In market research, it can be argued that the main use of conjoint analysis is in measuring the degree of 'importance' that consumers (i.e. respondents) attach to the characteristics of choices. In fact, the measurement of importance has been a problem that has bedevilled researchers for many years. In contrast to other areas of research requiring measurement — such as behaviour, brand image, core attitudes, and so on — it has proved to be a considerably more difficult type of data to collect than other types.

Direct measures of importance — It is sometimes difficult to see why this should be. After all, nothing could be simpler than asking customers to consider a range of product or service characteristics and communicating directly how important each of these is to them. A number of collection systems might be used — importance rating scales, ranking exercises, and so on. However, researchers who attempt this approach — sometimes known as the 'direct' approach — find themselves running into a number of problems.

Discrimination — It is frequently found that importance rating scales (e.g. for service attributes) exhibit poor discrimination — i.e. all scores tend to be the same, or not statistically differentiated to a level of significance. This is understandable, since there is nothing to stop respondents making the task easier for themselves by declaring all attributes to be 'important' to them. This ignores the fact that in the real world, obtaining good performance on one attribute frequently has to be at the expense of performance on another attribute.

'Social attributes' — Another problem arises when 'softer' attributes are used in studies that possess a social content — in other words, telling the interviewer how important you rate a topic 'says something' about the sort of person you are. In service areas, this tends to mean that some respondents can find it difficult, for example, to admit that they find ‘staff politeness’ more important than (say) ‘staff efficiency’. This effect does not arise in all studies, but it can occur where high image attributes are used that possess a large degree of social or emotional content.

Indirect measures — In order to avoid the problems of the 'direct' approaches, researchers have frequently adopted other, more statistical, forms of approaching the problem more indirectly. One way of doing this is to ask respondents to rate a number of products or brands on the attributes concerned (i.e. obtain brand image ratings), and examine the statistical correlations between these and some overall rating for preference, using techniques such as multiple regression. These issues can be product specific, or they could be service related. The argument is that a strong positive correlation — arising when the image scores given to brands for a particular attribute rise and fall in step with overall rating — suggests that the overall rating is strongly associated with the opinion on the attribute — i.e. caused by it.

Table 1. Example conjoint attributes, with their levels

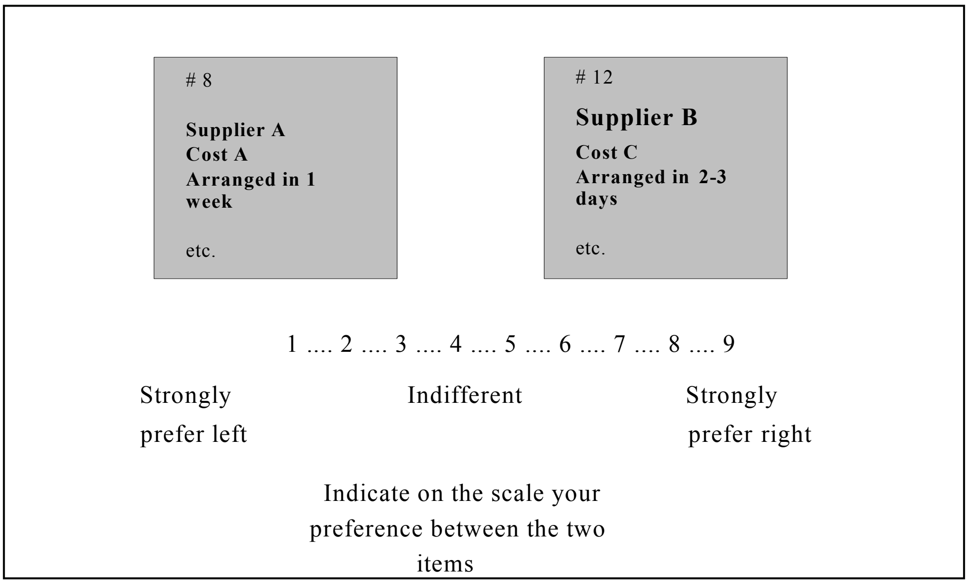
Brand (‘macro’)	**Price/fare** (‘macro’)	**Plane type** (‘micro’)	**Departure** (‘micro’)
British airways	100	Boeing 747	Morning
Lufthansa	120	Boeing 767	Afternoon
Air france	130	Airbus 320	Evening
Alitalia	140		
Iberia	150		

Source: own construction

Conjoint 'importances' — In many senses, therefore, conjoint analysis is the best technique for measuring attribute importances — partly because it is an 'indirect' method (that is, it does not ask the respondent to think about what is important, only what is preferred) that obtains a measure of importance by 'decomposing' stated preferences so that the importance of contributory factors can be inferred. At the same time, conjoint analysis also scores over other methods of measuring importance by specifically dealing with *stated levels of attributes*, rather than dealing with them as complete entities. This is extremely important in customer satisfaction measurement, where it is often insufficient to indicate that a particular service area is ‘important’ — it is also necessary to provide an operational definition of any breakpoints that apply.

The most modern form of collecting data for conjoint analysis involves the use of a laptop computer, so that the respondent interacts directly. It is an example of an ‘adaptive’ interview, in the sense that the computer program identifies the unique set of importances (or ‘utilities’) for each respondent by a series of iterative questions, the nature of which are determined by the preceding answers. As well as providing accurate measures, the task also has the benefit of sustaining interest among respondents.

Graph 2. An example of the type of question posed by the computer



8

Supplier A
Cost A
Arranged in 1 week

etc.

12

Supplier B
Cost C
Arranged in 2-3 days

etc.

1 .... 2 .... 3 .... 4 .... 5 .... 6 .... 7 .... 8 9

Strongly prefer left | Indifferent | Strongly prefer right

Indicate on the scale your preference between the two items

Source: own construction

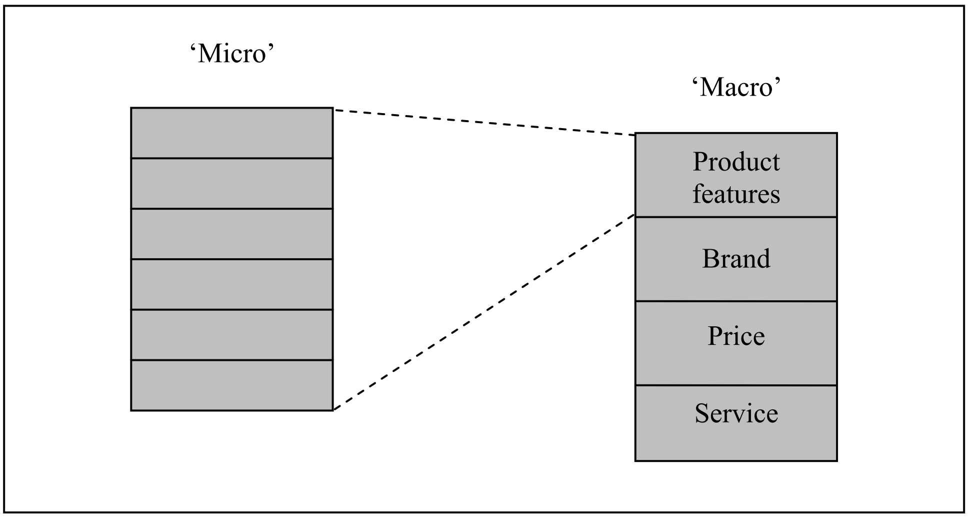
However, for many researchers the most appealing feature of conjoint analysis is that it can be used as a **model**, in which different product configurations can be simulated to provide a rich source of management information. Using a conjoint simulation model, organisations can identify those strategies which appear to offer the greatest benefit, in terms of providing customer satisfaction, product feature optimisation, or even optimal pricing.

6. Applying conjoint analysis to measure key choice criteria

Historically, conjoint analysis has most often been applied to product features, although other macro issues such as branding and price have sometimes been included in studies. However, recent thinking has tended to move away from using classical conjoint approaches in studies that mix ‘macro’ and ‘micro’ variables, on the grounds that micro variables do not naturally trade-off against macro ones; and that the process tends to underestimate the importance of the macro issues, given the many micro ones in a typical study.

The solution to the problem, is **dual scaling**. This essentially presents the task to the respondent as two conjoints — the first being a conventional conjoint exercise among product features (the ‘micro’ set); and the second being an exercise among the ‘macro’ set, but including pre-determined ‘bundles’ of micro attributes. These two models are then linked, using the utilities generated for the product features as common data. In effect, this then replaces the single attribute in the macro conjoint representing the bundles of product features. A single set of data across both sets of 'macro' and 'micro' attributes is thus created for each respondent, but without the distortion that would be incurred if they had been included in a common task. This is shown below.

Graph 3. Dual scaling



Source: own construction based on RI International report works

7. Expanding the service set — the smart[1] approach

Having effectively 'expanded' the 'macro' attributes dealing with product features into its constituent parts, it might be thought that the same could be done for the service attribute, which up till now has been measured as an aggregate topic, probably in terms of an attribute ranging from "very good service" to "poor service". However, although conjoint analysis works well for topics that have explicit descriptions, such as brand, price, and product features, it is not so easy to apply it directly to service areas. There are two key reasons for this:

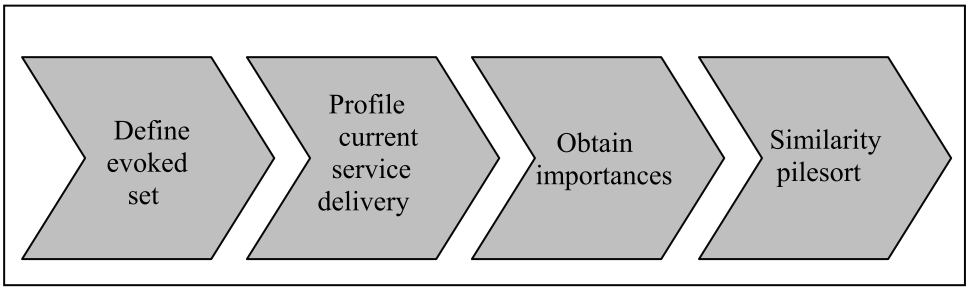
1. Service issues tend to be concerned with very many points of small detail. Unlike, say, product specification attributes, service areas tend to generate much larger sets of attributes. This means that a conventional conjoint task would be rapidly overwhelmed by the numbers of service issues that require measurement.
2. Service issues are qualitatively different to conventional conjoint methods in that they can incorporate topics with a substantially higher content of

[1] SMART is a registered service mark of Research International.

‘soft’ or emotional themes, especially where personal relationships are concerned. This means that a classical conjoint exercise would tend to allow people to ‘over-rationalise’ their responses, and provide answers that emphasise functional issues at the expense of ‘softer’ ones.

In view of this, it has been developed a proprietary extension of conjoint analysis to deal with service areas. Known as ‘smart’ (**S**alient **M**ulti-**A**ttribute **R**esearch **T**echnique), this method is a direct descendent of traditional conjoint approaches, but is adapted for the circumstances of measuring service variables. A full description of the method can be found elsewhere (Baird, Banks, Smith, Morgan, 1988).

Graph 4. Stages of smart elicitation



Source: own construction based on RI International report works

The main stages of the process are as follows:

1. Elicitation of ‘evoked set’ — Respondents are shown a range of service attributes, which are expressed (like conjoint variables) in terms of delivery levels. These levels are progressive (unlike conjoint variables) in that moving from one level to the next represents an increase in service delivery.
2. Obtaining perceived delivery levels — The second task is for the respondent to identify for each attribute the service level most closely associated with suppliers. This will depend on the market, but it is usual for a respondent to profile a number of suppliers. The respondent will consider each attribute for each supplier, and select the appropriate level (in their opinion), or indicate that they ‘don’t know’.
3. Obtaining importance weights (‘utilities’) for attributes — the respondent is asked to indicate the relative importance he or she attaches to each of the service levels. This is done by means of a ‘choice game’, which works as follows: respondents are shown an array of the attributes in their evoked set, all initially expressed in terms of the ‘worst’ level. He or she is then asked to imagine that this was the current situation, and invited to indicate the attribute they would prefer to improve first. This attribute is then advanced to the next level, and the question repeated. Each time the

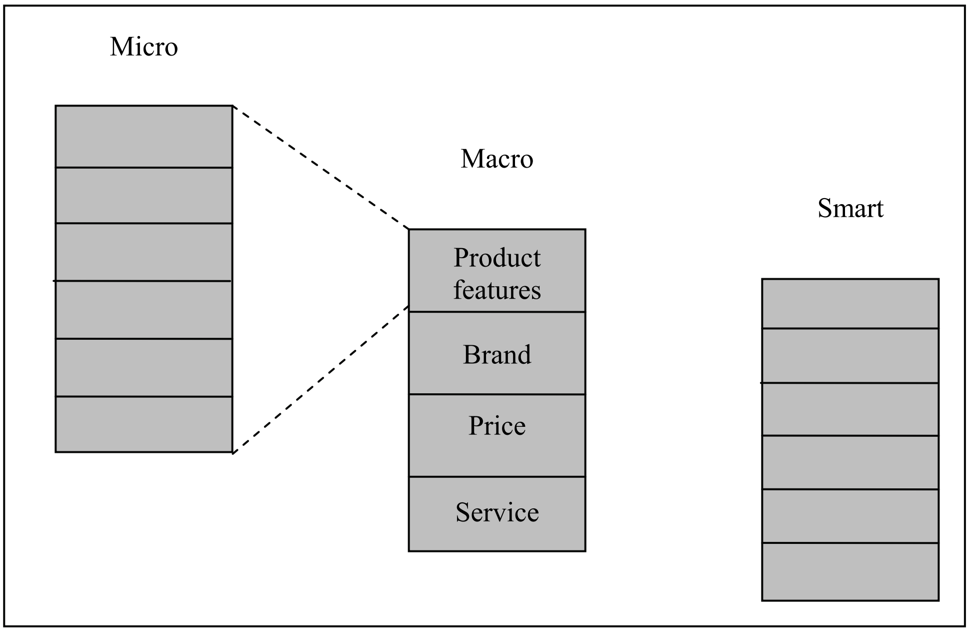
respondent selects an attribute, this is 'improved' to the next level. When an attribute is 'improved' to the highest level, it is removed. The game ends when all attributes have been 'improved' to their 'best' levels.

4. The 'pilesort' exercise — The final part is the elicitation procedure that deals with the technical problem of multi-collinearity within the attribute set. This is necessary, since unlike conjoint variables (which are ordinarily assumed to be statistically independent of one another), service variables are quite likely to 'overlap' in terms of meaning. Since these are really dealing with the same underlying theme, we need to understand this, and realise that we are dealing with one problem, and not ten or more. The data is collected quite simply by asking respondents to sort the attributes (which are written on cards) into piles, each pile representing a 'theme' in their own mind. They can have as many piles as they like. From this data, a similarity matrix derived from co-occurance measures (Burton, 1975) is decomposed using Principal Components Analysis to indicate the main service 'themes' or factors.

8. Linking all key choice criteria together – the composite model

The final stage in the modelling process is the integration of the 'smart' data with the other conjoint data. Essentially, this is performed in the same way as for product features noted earlier, using the macro variable for 'service' as the linking variable.

Graph 5. Composite model



Source: own construction based on RI International report works

Since we have created a merged set of utilities for both the conjoint and the ‘smart’ data, the next step is to use the utilities in a choice simulation model. In principle this is straightforward, since in replacing the original conjoint service attribute with the ‘smart’ utilities (appropriately scaled), we have not changed the properties of the original conjoint model, we have simply given it more attributes.

In practice, however, it is not straightforward. The reason is that unlike the conjoint attributes, which since they are global are much more likely to act independently as we have noted, the ‘smart' attributes will inevitably exhibit high levels of multi-collinearity — i.e. they have a considerable degree of overlap. In other words, the 30—40 or so attributes may be covering a much smaller number (say, 8 to 10) of themes. This means that if we were to treat the ‘smart’ attributes in a merged model in the same way as the conjoint attributes (i.e. as independent variables) we would run a considerable risk of double counting. This requires a modification to the conjoint model. To do this, the model is based on a fixed number of *factors* F, consisting of C conjoint variables plus S – ‘smart’ themes:

$$F = C + S$$

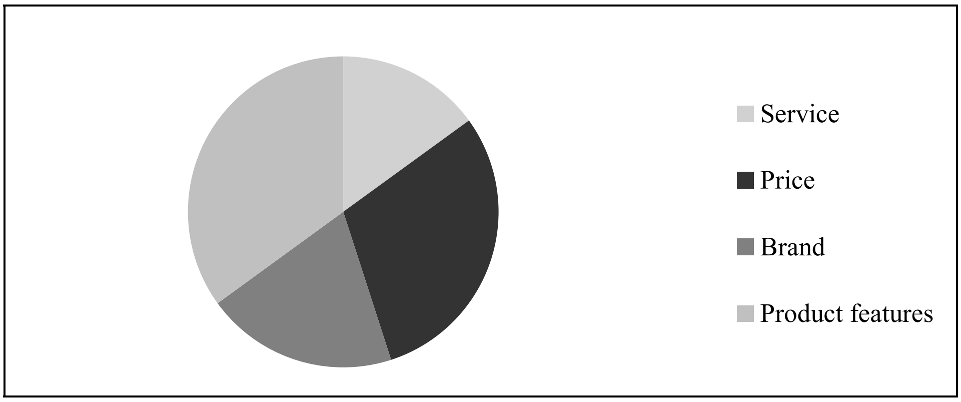
where: S is less than the original number of ‘smart’ variables.

9. Using the data

The data can be examined in three main ways, discussed below:

1. Understanding 'importances'. The complex procedure described above results in a common set of 'utilities' for all the attributes, whether conjoint in origin ('micro' as well as 'macro') or service attributes. These can then be compared, and an overall understanding of the relative impact of service and other attributes to the choice itself can be understood.

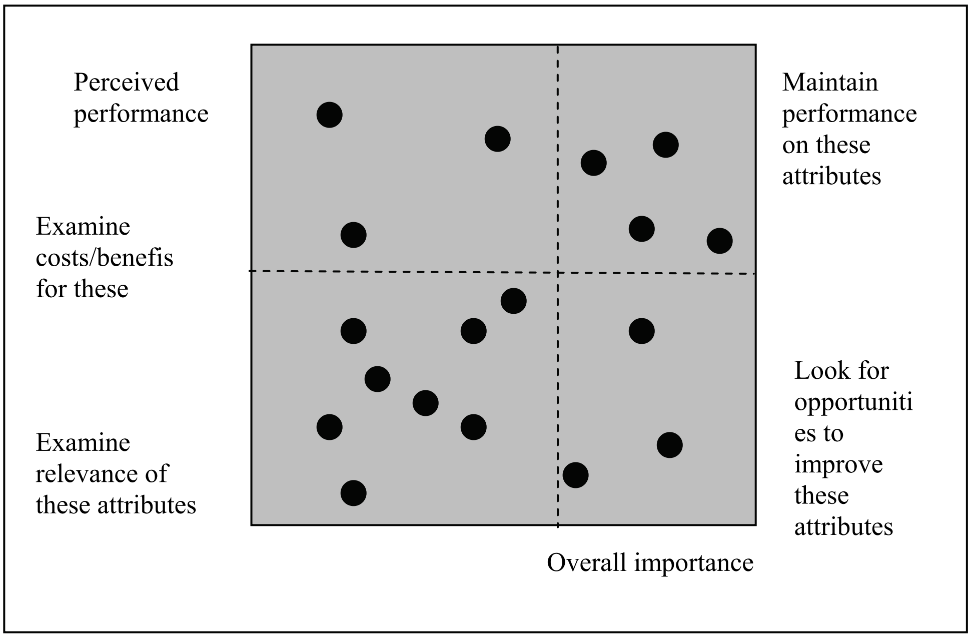
Graph 6. Illustration of relative contributions



Source: own construction

As with conjoint variables, the values themselves are notional, and are used to index the relative importance of topics. However, simple knowledge of the importance of service criteria is insufficient unless it is linked to knowledge of how customers *perceive* current delivery. Since the process outlined above has recorded this, we can therefore combine both data sets. One way of doing this is to examine individual organisations (perhaps competing suppliers) and generate a 'quadrant' plot of service attributes on two axes: firstly, the relative importance of the topics as measured by the 'smart' utilities, and secondly the degree to which the organisation is performing on them.

Graph 7. Quadrant for organisation X


Perceived performance
Examine costs/benefis for these
Examine relevance of these attributes
Maintain performance on these attributes
Look for opportuniti es to improve these attributes
Overall importance

Source: own construction

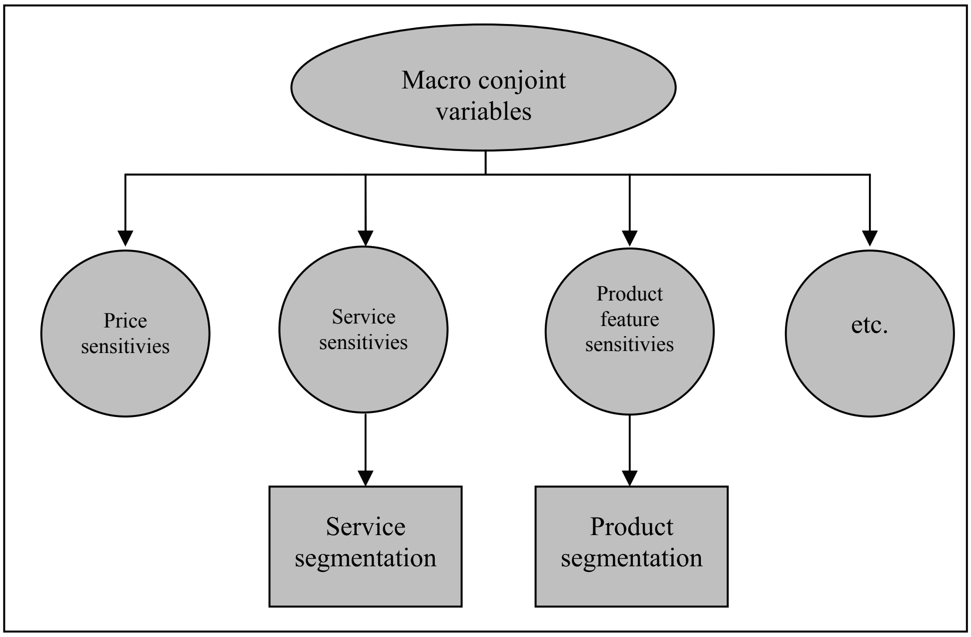
The ‘quadrants’ thus formed have direct marketing significance. The most significant one is formed in the bottom right of the place, which is an area of high service importance, but poor delivery from the organisation. Attributes falling in this area therefore would have the biggest impact, were they to be improved. At the same time, the organisation should not let its attention be drawn from topics falling into the top right quadrant, since these are issues which are important to customers, and on which the organisation is performing well. The aim should be therefore to maintain performance. The left hand quadrants refer to topics that are of lower importance to customers. However, the upper of these quadrants refer to issues that customers perceive that the organisation is performing well — this is fine, provided that there are no costs attached to this. Consequently, topics in this area deserve some cost/benefit analysis. In general, the strategic aim of the organisation is to shift resource (where appropriate) from topics in the top left of the display, to the bottom right.

2. Customer segmentation Another feature of conventional conjoint measurement can also be applied here. Since conjoint utilities are typically computed at the individual level, they can be analysed by various multivariate techniques such as Cluster Analysis to provide an indication of the natural groupings or segments of consumers that occur within a market.

This approach is commonly undertaken in customer service exercises. However, the procedure outlined here has a extra benefit. A conventional exercise would tend to cluster all respondents in a study, and possibly include those to whom service is a minor issue with those to whom it is very important indeed. This leads to the less than satisfactory outcome that service clusters derived in this way are 'diluted' by those who are less interested in service anyway.

A better approach is to identify those customers who are 'service sensitive' and then segment only these people. With the data we have, we can perform a multi-stage segmentation by firstly clustering the 'macro' conjoint variables, which contain service as an aggregate issue. This would lead us to identify service sensitive individuals, who would then be subjected to a secondary segmentation.

Graph 8. Macro conjoint variables



Source: own construction

3. Simulation modelling — Finally, the third way of making use of this data is in the construction of a simulation model. The principles of modelling conjoint material are well described elsewhere (Green, Srinivasan, 1990) but the core of the process involves constructing a set of decision rules based upon the total sum of utility generated by hypothetical product profiles. The model computes the likely choice or preference of every individual in the sample, and aggregates these to provide a summary of market share changes. This is an example of 'micro-modelling' since data is handled in this way at the individual level.

The power of this procedure lies in the ability to simulate the likely customer reaction to a very large number of possible initiatives, which can be defined in terms of the input variables. In this case, since we have all data on a common metric, we can define our input variables to be the full range of key choice criteria — brand, price, product features, service features, and so on. A full model of this sort may well (in our experience) have 30—40 such variables. In addition, unlike many conventional conjoint models, one can include the individual imagery that individuals associated with the service delivery as part of the model data itself. This means that we are not obliged to 'specify' the profile of current market brands or services — a drawback of conventional models dealing with attributes that are non-factual (i.e. the level of delivery is a matter of opinion, not fact).

Given this, in the case of service variables, a common use of the model is to perform a 'sensitivity analysis' of service variables, in terms of simulating the outcomes in scenarios where the organisation 'improves' its delivery on specific issues (or indeed, 'worsens' its performance). This enables us to identify key service areas that lead to the greatest degree of impact in customer satisfaction. At the same time, the process also enables us to identify areas that, while currently performing well, would cause great harm if allowed to deteriorate.

9. Performance tracking

Organisations undertaking quality programmes invariably require feedback from the customer base that changes initiated by management are impacting on customers. At the same time, there is a continuous need to monitor customer expectations, since it has been widely reported that the growth of a 'service culture' in many industries itself leads to a growth in customer expectations of service quality. It would be unrealistic to expect an organisation to conduct studies of the sort described above on anything other than a periodic basis, principally to cost constraints. In fact, it can be argued that it is not necessary to collect all the information described above on a regular basis, since it is likely that customer's *perceptions* of service quality from suppliers are likely to change more rapidly than their own intrinsic *needs*. Consequently, a continuous tracking programme can concentrate on measuring these perceptions (which is easy and cost effective to do), and only update on measuring needs at planned intervals. These periodic measurements of the full data set are called 'needs assessment' studies.

11. Conclusions

Conjoint analysis has proved over the years to be a very useful and reliable research tool to organisational management, provided that care and effort is taken to manage the interface between the results that are generated by the modelling and the decisions to be taken. In general, its extension into the area of customer satisfaction represents a major enhancement of its appeal to both researchers and management alike.

REFERENCES

BANKS, R., DE KORT, E., FERRER-VIDAL, J., MARTIN-ONRAET, B. (1988), *Customer Service in Europe: One Market or Many?*, Esomar Congress, Stockholm.

BAIRD, C., BANKS, R., SMITH, P., MORGAN, R. (1989), *The SMART^{sm} Approach to Customer Service*, Market Research Society Conference, Brighton.

BIEL, A. (1992), *Anticipating Expectations: What Will Tomorrow's Customers Want?*, Congress, Madrid.

BUROS, K., BLACKALL, S. (1992), *The Responsibility for Service Delivery: Using Research to Generate Effective Ownership and Implementation: the US and UK Experience*, Esomar Congress, Madrid.

BURTON, M. L. (1975), *Dissimilarity Measures for Unconstrained Sorting Data*, "Multivariate Behavioral Research", No. 10, pp. 409—24.

GREEN, PAUL E., SRINIVASAN, V. (1990), *Conjoint Analysis in Marketing Research: New Developments and Directions*, "Journal of Marketing", Vol. 54, No. 4, pp. 3—19.

KING, S. (1991), *Brand Building in the 1990's*, "Journal of Marketing Management", Vol. 7, No. 3—13.

MILLER, G. (1956), *The magic number seven — plus or minus two: some limits on our capacity for processing information*, "Psychological Review", No. 63, pp. 81—97.

PARASURAMAN, A., ZEITHAML Z., BERRY L. (1984), *A Conceptual Model of Service Quality and its Implications for Further Research*, Cambridge, Marketing Science Institute.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 327—335

CONSTRUCTING SOME A-OPTIMAL WEIGHING DESIGNS WITH 23 WEIGHINGS: A NEW METHOD

Farmakis Nicolas[1]

ABSTRACT

A-Optimal Weighing Designs $R^{*}_{23\times 15}$ with n=23 observations and k=2,3,4,…,17 parameters are constructed. The information matrix M^{*}_{k} of such a design is a block matrix with s_{opt} blocks. The construction of an A-optimum design $(n,k,s_{\text{opt}})=(23,k,k)$, k=2,3,…,10 is obvious, based on Hadamard Matrices $\mathbf{H}_{24}$. The construction of the A-optimum design $(n,k,s_{\text{opt}})=(23,k,s_{\text{opt}})$, k=11,12,13,14, with $s_{\text{opt}}<k$ is based on the method in Kounias and Farmakis (1984). The construction of (n,k,s_{opt}) with $s_{\text{opt}}<k$ and k=15,16,17 is based on a new method similar to the method used in Farmakis (1991).

Key words: Experimental, Design, Optimal, Trace, Weighing.

1. Introduction

Suppose that we have to estimate the weights of k objects with $n \geq k$ weighings (observations), using a chemical balance. The problem of constructing matrices – *experimental designs* — of the type $R_{n\times k} = \{r_{ij}\}$, with $r_{ij} = \pm 1$, i=1,2,...,n and j=1,2,…,k, then arises. All the matrices $R_{n\times k}$ are elements of the set denoted by $\mathbf{D}(n,k)$. The matrix $M_k = R'_{n\times k} \cdot R_{n\times k}$ is the so-called *information matrix* related to the design $R_{n\times k}$. We define the function Φ as follows:

$$\mathbf{D}(n,k) \xrightarrow{\Phi} \mathbf{R} = \{\text{real numbers}\} \qquad (1.1)$$

The form of Φ in (1.1) is

[1] Aristotle University of Thessaloniki, Dept. of Mathematics, GR-54124 Thessaloniki – GREECE, farmakis@math.auth.gr.

$$\Phi(R_{n\times k})=\frac{1}{k}\cdot\sum_{i=1}^{k}\mu_i^{-1}=\frac{1}{k}\cdot\text{trace}\left(M_k^{-1}\right)=\frac{1}{k}\cdot\text{trace}\left(\left(R'_{n\times k}\cdot R_{n\times k}\right)^{-1}\right) \qquad (1.2)$$

where the quantities μ_i (i=1,2,…,k) are the eigenvalues of the information matrix M_k. The function Φ in (1.1) is called the A-criterion and the design $R^* \in \mathbf{D}(n,k)$ for which

$$\Phi\left(R^*\right)=\min(\Phi(R)) \qquad (1.3)$$

is called the "A-optimal Design".

2. Blocks and Block Matrices

For the purposes of this paper we need the concepts of *block* and of *block matrix*.

Definition 2.1: A block of size r is an $r\times r$ matrix with all its diagonal elements n and off diagonal elements 3, is as

$$B_r=(n\text{-}3)\cdot I_r+3\cdot J_r \qquad (2.1)$$

where I_r is the identity matrix of order r and J_r is the $r\times r$ matrix with all its elements equal to 1.

Definition 2.2: A block matrix $\mathbf{M}_k$, with block sizes $r_1, r_2, r_3, \ldots, r_s$ satisfying $\sum_{i=1}^{s} r_i = k$, is a $k\times k$ matrix denoted also by

$$\text{BM}(n,k,s;r_1,r_2,\ldots,r_s)=\mathbf{M}_k$$

with its s diagonal blocks of the respective sizes and all its other elements equal to −1.

We can write this as

$$\text{BM}(n,k,s;r_1,r_2,r_3,\ldots,r_s)=\text{diag}\left\{\left(B_{r_1}+J_{r_1}\right)+\left(B_{r_2}+J_{r_2}\right)+\ldots+\left(B_{r_s}+J_{r_s}\right)\right\}-J_k \qquad (2.2)$$

Also we denote by M_k^* the block matrix, which minimizes traceT^{-1} over all block matrices T with blocks of only one size r or of two contiguous sizes r and r+1. Sathe and Shenoy (1989) proved that if a design $R^*_{n\times k}$ has M_k^* as information matrix, then $R^*_{n\times k}$ is an A-*optimal block design*. They also constructed a table with the block forms M_k^* of the A-optimal block designs

R^*_{nxk} =(n,k,s_{opt}), for n<100 and $n\equiv 3 \pmod 4$. They proved that for every design $R^*_{n\times k}$ there are s_{opt} blocks; u of size r and v of size r+1, i.e. $k=u\cdot r+v\cdot(r+1)$.

It is easy to see that we have

$$r=\left[\frac{k}{s_{\text{opt}}}\right] \tag{2.3}$$

where [x] is the integer part of x, i.e. the biggest integer less than or equal to x, and

$$v=k\text{-}r\cdot s_{\text{opt}} \text{ and } u=s_{\text{opt}}\text{-}v. \tag{2.4}$$

Table 2.1. gives, for n=23, the forms of all the A-optimal designs (23,k,s_{opt}), k=2,3,4,…,23 with the values of all the suitable parameters r, u, v, trace$\left(M_k^{*-1}\right)$.

Table 2.1.

k	s_{opt}	r	u	v	trace$\left(M_k^{*-1}\right)$
2	2	1	2	0	0.0871
3	3	1	3	0	0.1310
4	4	1	4	0	0.1750
5	5	1	5	0	0.2193
6	6	1	6	0	0.2639
7	7	1	7	0	0.3088
8	8	1	8	0	0.3542
9	9	1	9	0	0.4000
10	10	1	10	0	0.4464
11	10	1	9	1	0.4935
12	8	1	4	4	0.5408
13	7	1	1	6	0.5881
14	7	2	7	0	0.6357
15	6	2	3	3	0.6835
16	6	2	2	4	0.7313
17	6	2	1	5	0.7794
18	5	3	2	3	0.8274
19	5	3	1	4	0.8754
20	5	4	5	0	0.9236
21	5	4	4	1	0.9720
22	5	4	3	2	1.0206
23	5	4	2	3	1.0692

The traces in the last column of Table 2.1 are given in general by the following formula for block matrices with *s* blocks of sizes $r_1, r_2, r_3, \ldots, r_s$, as in Sathe and Shenoy (1989) trace

$$\left(M_k^{-1}\right) = \frac{k-s}{n-3} + \sum_{i=1}^{s} L_i^{-1} + \frac{\sum_{i=1}^{s} r_i \cdot L_i^{-2}}{1 - \sum_{i=1}^{s} r_i \cdot L_i^{-1}}, \; L_i = n - 3 + 4 \cdot r_i, \; i=1,2,\ldots,s. \qquad (2.5)$$

In the case of the A-optimal design the corresponding information matrix is a block matrix with *u* blocks of size *r* and *v* blocks of size *r*+1. So the formula becomes trace:

$$\left(M_k^{*-1}\right) = \frac{k-s}{n-3} + \frac{u}{n-3+4r} + \frac{v}{n+1+4r} + \frac{\frac{ur}{(n-3+4r)^2} + \frac{v(r+1)}{(n+1+4r)^2}}{1 - \frac{ur}{n-3+4r} - \frac{v(r+1)}{n+1+4r}}, \qquad (2.6)$$

where *u*+*v*=*s*.

More specifically for *n*=23 it is trace

$$\left(M_k^{*-1}\right) = \frac{k-s}{20} + \frac{u}{20+4r} + \frac{v}{24+4r} + \frac{\frac{ur}{(20+4r)^2} + \frac{v(r+1)}{(24+4r)^2}}{1 - \frac{ur}{20+4r} - \frac{v(r+1)}{24+4r}}, \qquad (2.7)$$

where *u*+*v*=*s*.

3. Constructing A-optimal Designs (23,k,k)

The case $s_{\text{opt}}=k$ occurs for *k*=2,3,4,…,10 as we see in Table 2.1. It is easy to construct these A-optimal designs (A-OWD) by using the Hadamard matrices $\mathbf{H}_{24}$. Recall that one important use of Hadamard matrices is to take them as a basis for getting optimal designs (not only A-optimal ones). For the case we are dealing with (A-optimal designs with $n \equiv 3 \pmod 4$) we use the suitable Hadamard matrix $\mathbf{H}_{n+1}$ and we use the following short algorithm:

1st step: Multiply by −1 all the columns of $\mathbf{H}_{n+1}$ which have −1 as their first element.

2nd step: Delete first row of the matrix

3rd step: In order to obtain the design (23,*k*,*k*), select any *k* columns of the remaining matrix, *k*=2,3,…,10.

4. Constructing A-optimal Designs (23,*k*,s_{opt}), *k*>s_{opt}

The first 4 cases of A-optimal designs with n=23 and k>s_{opt}, i.e. k=11,12,13,14, can be easily constructed using the method in Kounias and Farmakis (1984). The suitable initial matrix **G′** (transpose of **G**, the Goethals-Seidel Matrix) of size 24×24 is presented below in terms of the 2×2 cyclic matrices $A=\begin{pmatrix}1&1\\1&1\end{pmatrix}$ and B=$\begin{pmatrix}1&-1\\-1&1\end{pmatrix}$:

$$\mathbf{G'}=\begin{bmatrix}
A & -B & B & B & B & A & B & A & A & A & A & -A\\
B & A & -B & B & A & B & A & A & B & A & -A & A\\
-B & B & A & A & B & B & A & B & A & -A & A & A\\
-B & -B & -A & A & -B & B & -A & -A & A & A & B & A\\
-B & -A & -B & B & A & -B & -A & A & -A & B & A & A\\
-A & -B & -B & -B & B & A & A & -A & -A & A & A & B\\
-B & -A & -A & A & A & -A & A & -B & B & -B & -B & -A\\
-A & -A & -B & A & -A & A & B & A & -B & -B & -A & -B\\
-A & -B & -A & -A & A & A & -B & B & A & -A & -B & -B\\
-A & -A & A & -A & -B & -A & B & B & A & A & -B & B\\
-A & A & -A & -B & -A & -A & B & A & B & B & A & -B\\
A & -A & -A & -A & -A & -B & A & B & B & -B & B & A
\end{bmatrix}$$

We now give the constructions of (23,k,s_{opt}), k=11,22,13,14.

- **(23,11,10)**

1st step: Multiply by –1 all the block-rows of **G′**with first block –A or –B.
2nd step: Keep the first block-row.
3rd step: Delete the two last block-rows
4th step: Delete from the remaining matrix the next (simple) rows: 4th, 6th, 8th, 10th, 11th (or 12th), 14th, 15th (or 16th), 17th (or 18th) and finally 19th (or 20th).
5th step: Delete the 1st column of the remaining matrix. The result is a suitable A-OWD.

As we see this design can be constructed via many (equivalent) ways from **G′** by using the method of Kounias and Farmakis (1984). This is valid for the next constructions too.

- **(23,12,8)**

1st step: Multiply by –1 all the block-rows of **G′**with first block –A or –B.
2nd step: Delete the block-rows: 1st, 6th, 7th, 8th.
3rd step: Delete from the remaining matrix the next (simple) rows: 2nd, 4th, 6th, 8th.
4th step: Delete the 1st column of the remaining matrix and the result is suitable A-OWD.

- **(23,13,7)**

1st step: Multiply by –1 all the block-rows of $\mathbf{G}'$with first block $-A$ or $-B$.
2nd step: Delete its first five block-rows.
3rd step: Delete from the remaining matrix the 4th (simple) row.
4th step: Delete the 1st column of the remaining matrix and the result is suitable A-OWD.

- **(23,14,7)**

1st step: Multiply by –1 all the block-rows of $\mathbf{G}'$with first block $-A$ or $-B$.
2nd step: Delete all the block-rows with first block element B.
3rd step: Delete the 1st column of the remaining matrix and the result is suitable A-OWD.

- **(23,15,6), (23,16,6), (23,17,6)**

For the construction of $(23,k,6)$, k=15,16,17 a method other than the one in Kounias and Farmakis (1984) is needed. This new method is similar to the method used in Farmakis (1991) for the construction of some designs $(19,k,s_{\text{opt}})$ for k=10,11,…,16. For the new constructions of this paper, first of all, the following 14×23 matrix $\mathbf{C}'$ (obtained from $\mathbf{G}'$ as a submatrix) is needed:

$$\mathbf{C}'=\begin{bmatrix} \mathbf{1}_2 & -B & B & B & B & A & B & A & A & A & A & -A \\ \mathbf{1}_2 & B & B & B & -B & -A & -A & A & A & -A & -A & -B \\ \mathbf{1}_2 & A & B & -A & A & -A & -B & -A & B & B & A & B \\ \mathbf{1}_2 & B & A & A & -A & -A & B & -B & -A & A & B & B \\ \mathbf{1}_2 & A & -A & A & B & A & -B & -B & -A & -A & B & -B \\ \mathbf{1}_2 & -A & A & B & A & A & -B & -A & -B & -B & -A & B \\ \mathbf{1}_2 & -A & -A & -A & -A & -B & A & B & B & -B & B & A \end{bmatrix}$$

In matrix $\mathbf{C}'$ the entry $\mathbf{1}_2$ stands for the vector $\begin{pmatrix}1\\1\end{pmatrix}$.

After this:

1) We retain the first 6 block rows of $\mathbf{C}'$as they are, i.e. 12 (simple) rows.
2) We take the last block row of $\mathbf{C}'$and we multiply its last element A by –1.
3) We use only one of the two resulting last (simple) rows as the third row of the design (23,15,6) we try to construct.
4) We delete the remaining last row of $\mathbf{C}'$.
5) We fill in the resulting 13×23 matrix by 2 more new rows for (23,15,6), by 3 new rows for (23,16,6) and by 4 rows for (23,27,6). The suitable rows are taken from the output of a computer running, as will be explained.
6) The program uses the resulting 13×23 matrix as initial nucleus of the target designs. In Farmakis (1991) the nucleus matrix has only 5 rows. Since for our goal we need some blocks of size 3 and some of size 2 the

problem is now to find suitable rows of length 23 with which the 2nd block (6th row), the 3rd block (9th row), the 4th block (12th row) and the 5th block (15th row) will be filled. The last block will be of size 2 anyway (see Table 2.1).

The program yielded:

6 rows as candidates for 6th row		(I)
4 rows as candidates for 9th row	(23,15,6)	(II)
4 rows as candidates for 12th row	(23,16,6)	(III)
4 rows as candidates for 15th row	(23,17,6)	(IV)

From the above 18 rows we use the 5th one from (I) as the 6th row of the design and the 4th one in (II) as the 9th row of the (23,15,6) A-optimal design, which is:

$$R'_{23\times 15} = \left[\begin{array}{rrrrrrrrrrrrrrrrrrrrrrr}
1&-1&1&1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&-1&-1\\
1&1&-1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&1&-1&-1\\
1&-1&-1&-1&-1&-1&-1&-1&-1&-1&1&1&1&1&-1&1&-1&-1&1&1&-1&-1&-1\\
1&1&-1&1&-1&1&-1&-1&1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&-1&1\\
1&-1&1&-1&1&-1&1&1&-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1\\
-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&1&-1&1&-1&-1&-1\\
1&1&1&1&-1&-1&-1&1&1&-1&-1&-1&1&-1&-1&1&-1&1&-1&1&1&1&-1\\
1&1&1&-1&1&-1&-1&1&1&-1&-1&1&-1&-1&-1&-1&1&-1&1&1&1&-1&1\\
-1&-1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&1&-1&-1&-1&-1&1&1&-1&-1\\
1&1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&-1&-1&1&1&1&-1&1&-1\\
1&-1&1&1&1&1&1&-1&-1&-1&-1&-1&1&1&-1&-1&-1&1&1&-1&1&-1&1\\
1&1&1&-1&-1&1&1&1&-1&1&1&-1&1&-1&1&-1&-1&-1&-1&1&-1&-1&1\\
1&1&1&-1&-1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&-1&-1&-1&1&1&-1\\
1&-1&-1&1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&1&-1&1&-1&-1&1&-1\\
1&-1&-1&1&1&-1&1&1&1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&-1&1
\end{array}\right]$$

The above (23,15,6) A-OWD has $\mathbf{M}_{15}$=BM(23,15,6;3,3,3,2,2,2,2)=(23,15,6) as information matrix. In this (23,15,6) A-optimal design we insert as the 12th row the 2nd row in case (III) of the previous outputs and we obtain the (23,16,6) A-optimal design, which is:

$$R'_{23\times16} = \begin{bmatrix}
1&-1&1&1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&-1&-1\\
1&1&-1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&1&-1&-1\\
1&-1&-1&-1&-1&-1&-1&-1&-1&-1&1&1&1&1&-1&1&-1&-1&1&1&-1&-1&-1\\
1&1&-1&1&-1&1&-1&-1&1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&-1&1\\
1&-1&1&-1&1&-1&1&1&-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1\\
-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&1&-1&1&-1&-1&-1\\
1&1&1&1&-1&-1&-1&1&1&-1&-1&-1&1&-1&-1&1&-1&1&-1&1&1&1&-1\\
1&1&1&-1&1&-1&-1&1&1&-1&-1&1&-1&-1&-1&-1&1&-1&1&1&1&-1&1\\
-1&-1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&1&-1&-1&-1&-1&1&1&-1&-1\\
1&1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&-1&-1&1&1&1&-1&1&-1\\
1&-1&1&1&1&1&1&-1&-1&-1&-1&-1&1&1&-1&-1&-1&1&1&-1&1&-1&1\\
-1&1&1&-1&1&1&-1&-1&-1&1&-1&-1&-1&-1&-1&1&-1&1&1&-1&-1&-1&-1\\
1&1&1&-1&-1&1&1&1&-1&1&1&-1&1&-1&1&-1&-1&-1&-1&1&-1&-1&1\\
1&1&1&-1&-1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&-1&-1&-1&1&1&-1\\
1&-1&-1&1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&1&-1&1&-1&-1&1&-1\\
1&-1&-1&1&1&-1&1&1&1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&-1&1
\end{bmatrix}$$

The above (23,16,6) A-OWD has $\mathbf{M}_{16}$=BM(23,16,6;3,3,3,3,2,2)=(23,16,6) as information matrix. If in this (23,16,6) A-optimal design we insert as the 15th row the 1st row in case (IV) of the previous outputs, we will obtain the (23,17,6) A-optimal design, which is:

$$R'_{23\times17} = \begin{bmatrix}
1&-1&1&1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&-1&-1\\
1&1&-1&-1&1&-1&1&-1&1&1&1&-1&1&1&1&1&1&1&1&1&1&-1&-1\\
1&-1&-1&-1&-1&-1&-1&-1&-1&-1&1&1&1&1&-1&1&-1&-1&1&1&-1&-1&-1\\
1&1&-1&1&-1&1&-1&-1&1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&-1&1\\
1&-1&1&-1&1&-1&1&1&-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1\\
-1&-1&-1&-1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&1&-1&1&-1&-1&-1\\
1&1&1&1&-1&-1&-1&1&1&-1&-1&-1&1&-1&-1&1&-1&1&-1&1&1&1&-1\\
1&1&1&-1&1&-1&-1&1&1&-1&-1&1&-1&-1&-1&-1&1&-1&1&1&1&-1&1\\
-1&-1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&1&-1&-1&-1&-1&1&1&-1&-1\\
1&1&-1&1&1&1&1&-1&-1&-1&-1&1&-1&-1&1&-1&-1&1&1&1&-1&1&-1\\
1&-1&1&1&1&1&1&-1&-1&-1&-1&-1&1&1&-1&-1&-1&1&1&-1&1&-1&1\\
-1&1&1&-1&1&1&-1&-1&-1&1&-1&-1&-1&-1&-1&1&-1&1&1&-1&-1&-1&-1\\
1&1&1&-1&-1&1&1&1&-1&1&1&-1&1&-1&1&-1&-1&-1&-1&1&-1&-1&1\\
1&1&1&-1&-1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&-1&-1&-1&1&1&-1\\
-1&1&1&1&-1&-1&1&1&1&1&-1&1&1&1&1&-1&-1&-1&1&-1&-1&-1&-1\\
1&-1&-1&1&1&1&-1&1&1&1&1&-1&1&-1&-1&-1&1&-1&1&-1&-1&1&-1\\
1&-1&-1&1&1&-1&1&1&1&1&1&1&-1&-1&-1&1&-1&1&-1&-1&-1&-1&1
\end{bmatrix}$$

The above (23,17,6) A-OWD has $\mathbf{M}_{17}$=BM(23,17,6;3,3,3,3,3,2)=(23,17,6) as information matrix. This method gave us three more A-optimal designs $(23,k,s_{\text{opt}})$, k=15,16,17 and its ability to generate additional A-optimal designs seems to be exhausted. As it has been proved that only (23,23,5) and (23,22,5) are unconstructible, see Farmakis (1992) and Sathe and Shenoy (1991), an open problem is constructing suitable A-optimal designs $(23,k,5)$, k=18,19,20,21. It seems to be difficult to determine whether the designs (23,*k*,5), *k*=18,19,20,21, are unconstructible or not by the above method.

REFERENCES

FARMAKIS, N. (1991) “Constructions of A-optimal weighing designs when n=19”, *J. Statist. Plann. Inference* 27, 249—261.

FARMAKIS, N. (1992) “On constructibility of A-optimal weighing designs (*n,k,*5), when $n \equiv 3$ (mod 4) and $k = n-1,\ n$”, *J. Statist. Plann. Inference* 33, 275—283.

KOUNIAS, S. and FARMAKIS, N. (1984) “A construction for D-optimal weighing designs when $n \equiv 3$ (mod 4) and $k = n-1,\ n$”, *J. Statist. Plann. Inference* 10, 177—187.

SATHE, Y.S. and SHENOY, R.G. (1989) “A-optimal weighing designs when $n \equiv 3$ (mod 4)”, *Ann. Statist.* 17, 1906—1915.

SATHE, Y.S. and SHENOY, R.G. (1990) “Construction method for some A and D-optimal weighing designs when $n \equiv 3$ (mod 4)”, *J. Statist. Plann. Inference* 24, 369—375.

SATHE, Y.S. and SHENOY, R.G. (1991) “Futher results on construction methods for some A and D-optimal weighing designs when $n \equiv 3$ (mod 4)”, *J. Statist. Plann. Inference* 28, 339—352.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 337—357

SOME MODIFICATION OF THE SIMPLE COMPONENT ANALYSIS

Andrzej Młodak[1]

ABSTRACT

We analyse effectiveness of the Simple Component Analysis using the socio–economical example of spatial differentiation of labour market in Polish towns with substantially methodologically diversified statistical data instead of yet investigated biomedical information, which are rather uniform in this context. We formulate also some proposals aimed at improvement of interpretability of the components without any significant loss of optimality and quality.

1. Introduction

Among various problems investigated by statisticians, a reduction in number of variables in the multivariate data analysis plays an important part. Composite phenomena are usually described by a system of many variables measuring different aspects of the subject of research. The most effective approach leading to decrease in data dimension (without any significant loss of information resource) consists in definition of components as linear combinations of the original variables.

The most popular technique used in realization of this postulate is the Principal Component Analysis (PCA), introduced by H. Hotelling (1933). Principal components, determined uniquely by vectors of coefficients of above–mentioned linear combinations (so–called "loadings") defined as successive eigenvectors of correlation matrix of the original variables are versatile optimal: they are uncorrelated, extract a maximum of variability of the original variables and vectors of their loadings are orthonormal. Unfortunately, principal components have also an important disadvantage. They define rather abstract indices and on this account there exists a serious problem concerning their practical interpretation. V. Rousson and T. Gasser (2003) have pointed to two

[1] Central Statistical Office, Statistical Office in Poznań, Branch in Kalisz, pl. J. Kilińskiego 13, 62—800 Kalisz, Poland, e–mail: a.mlodak@stat.gov.pl .

main premises being its reason. On the one hand, each component can be determined by small, medium and large loadings with some random element, on the other hand, PCA produces most often only one block–component (i.e. such that all its loadings have the same sign) which is much better interpretable (as a weighted sum of the original variables establishing some natural ordering) than other components, called difference–components (having some positive and some negative loadings).

After the Second World War, several attempts at construction of better interpretable components with minimal loss of the level of variability explanation were undertaken. A good example of those methods is the "varimax" rotation of principal components (H. F. Kaiser (1958) and I. T. Jolliffe (1995)). S. K. Vines (2000) has proposed another system of components whose loadings are proportional to integers. But the incline of interpretability resulting from application of those results is again and again not satisfactory.

An interesting suggestion was formulated recently by Swiss biostatisticians. V. Rousson and T. Gasser (2004) have introduced two–stage procedure of the Simple Component Analysis (SCA), which combines a clustering of variables according to their correlation and a shrinkage of principal components. Their approach is aimed at receiving much better interpretable components than the loadings constructed by PCA with relatively low loss of variability. Notwithstanding, one can observe also some inconveniences occurring in this case.

First, the empirical research concerning effectiveness of the SCA (with optimistic results) was conducted using some biomedical data, being rather methodologically uniform. Therefore, it is worth verifying the utility of this construction taking into account also information regarding various aspects of some composite socio–economic phenomena. The statistics related to this field are usually gathered during many surveys significantly diversified on the account of scope, methodology and measure of resulting data. Such a much sophisticated basis of assessment of the SCA will be assumed and considered in this article. More precisely, a set of 18 statistical variables describing situation of labour market in 66 key Polish towns in 2002 will be analysed.

Secondly, despite intention of the SCA constructors, this method often tends to create only one block–component. This observation is connected with slightly unfavourable choice of threshold value of optimal distance between clusters of variables. Moreover, from the informative point of view, the difference component could not be clearly distinguished. Thus, more than one simple difference component can have non–zero loadings referred to the same block, what makes difficult a perception of inter–group relations (the loadings of various difference–components having opposed signs for the same variable can "neutralize" one another). We propose some modification of construction of difference–components towards an optimisation of transformation of principal components according to clusters of variables.

2. Analysed data concerning the labour market in Polish towns

As we have noted in the introduction, the empirical database that will be analysed, contains 18 statistical variables (having — for a better comparability – a character of indices) describing various aspects of situation on the labour market in 66 Polish towns with the powiat status. Thus, in consequence of new territorial division of the country introduced in 1999, all these towns (within their administrative borders) are established as territorial units at the fourth level of the NUTS classification (the Nomenclature of Territorial Units for Statistical Purposes) standing in the European Union. Therefore, much more statistical information is available for them than for other towns being only capitals of urban gminas or urban–rural gminas (i.e. Polish NUTS 5 regions).

Table 1. presents a detailed description of the collected statistical variables. All the data concern year 2002. A significant argument for such choice of reference date is the fact that that year the National Population and Housing Census as well as the National Agricultural Census were conducted in Poland. They have contributed to gather many additional data not collected traditionally, in the midst of standard surveys[1].

Table 1. Variables describing situation of the labour market in 2002 in Polish towns established as NUTS 4 units

Symbol	Description	Comments
x_1	Population at working age per 100 population at non–working age.	As of 31 XII.
x_2	Number of persons, whose main source of maintenance is an income from work per 1000 population.	As of 20 V. Data from the National Population and Housing Census.
x_3	Number of persons, whose main source of maintenance is an income from benefit for unemployed persons per 1000 population.	As of 20 V. Data from the National Population and Housing Census.
x_4	Index of economic activity (in %).	As of 20 V. Data from the National Population and Housing Census.
x_5	Unemployment rate (in %).	As of 20 V. Data from the National Population and Housing Census.
x_6	Employed persons per 1000 population.	As of 31 XII. According to actual workplace; excluding economic entities employing up to 9 persons; including persons employed in private farms in agriculture.

[1] Because of limited size of this paper, it is not possible to present the whole data set, but the author is ready to make it available for everybody, who would be interested in them.

Symbol	Description	Comments
x_7	Employed persons in industry and construction (in % of total employed persons).	As of 31 XII.
x_8	Employed persons in market services (in % of total employed persons).	As of 31 XII.
x_9	Registered unemployed school – leavers (in % of total registered unemployed persons).	As of 31 XII.
x_{10}	Registered unemployed persons not entitled to benefit (in % of total registered unemployed persons).	As of 31 XII.
x_{11}	Registered unemployed persons terminated for company reasons (in % of total registered unemployed persons).	As of 31 XII.
x_{12}	Registered unemployed persons with tertiary educational level (in % of total registered unemployed persons).	As of 31 XII.
x_{13}	Registered unemployed persons at age up to 24 years (in % of total registered unemployed persons).	As of 31 XII.
x_{14}	Registered long–term unemployed persons (in % of total registered unemployed persons).	As of 31 XII. Persons unemployed longer than 1 year.
x_{15}	Persons working in hazardous conditions per 1000 employed persons.	As of 31 XII. Listed only by predominant factor, i.e. having the most harmful importance at a given workplace; data concern entities employing more than 9 persons.
x_{16}	Persons injured in occupational accidents per 1000 employed persons.	Registered in a given year; excluding accidents in private farms in agriculture.
x_{17}	The number of days incapacity to work per one person injured in occupational accidents.	Excluding persons injured in fatal accidents.
x_{18}	Average monthly wage and salary (in PLN).	According to actual workplace; excluding economic entities employing up to 9 persons.

Source: Author's elaboration on the basis of publications by CSO (2003 a, 2003 b).

The diversification of variables expressed by values of their classical coefficient of variation (i.e. a standard deviation divided by arithmetical mean) vary from 4.53% (for $\boldsymbol{x}_4$) to 68.88% (for $\boldsymbol{x}_{15}$). The value of this index for prevailing majority of variables is greater than 10%, so one can say that the data set seems to be an effective basis for further investigation.

An important question connected with multivariate statistical model concerns a determination of character of each variable from the point of view of its influence on general situation of researched phenomenon (for some variables it is "good" to have a high value whereas for some other variables it is "good" to have a low value). More formally and systematically, taxonomicians distinguish three types of statistical variables (cf. A. Zeliaś (2002)):

1) stimulants — variables, whose high values testify to better situation of a given object on the account of the aggregated field of research,
2) destimulants — variables, whose high values characterize worse situation of a given object,
3) nominants — variables having so–called nominal level of value; their values increasing to this level have a positive impact on the assessment of a phenomenon, whilst further increase above the nominal level leads to the negative impact on the investigated subject matter (for example share of investments in the Gross Domestic Product). Of course, the nominant can have also opposed property — i.e. it can have a character of a destimulant below the nominal level and a stimulant — above it.

Among 18 analysed variables, the following indices can be clearly perceived as stimulants: $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_4$, $\boldsymbol{x}_6$, and $\boldsymbol{x}_{18}$. Without a doubt, destimulants are: $\boldsymbol{x}_3$, $\boldsymbol{x}_5$, $\boldsymbol{x}_9$, $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{11}$, $\boldsymbol{x}_{12}$, $\boldsymbol{x}_{13}$, $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{16}$, and $\boldsymbol{x}_{17}$. A more serious problem concerns the variables $\boldsymbol{x}_7$ and $\boldsymbol{x}_8$. According to their definitions, one can suppose that they are stimulants. On the other hand, there exists a strong statistical and logical contradiction between them — expressed by negative correlation coefficient with high absolute value (amounting to -0.82891). It is fully understandable — the increase of employees in the industry and construction should generally cause a decrease of the number of persons employed in market services (and conversely). So, directions of influence of both variables on the situation on the labour market are quite opposed. Therefore we can conclude that at least one of them is a nominant. From the rational point of view, it is rather the variable $\boldsymbol{x}_8$ than $\boldsymbol{x}_7$, because after an overdraft of some optimal threshold, further increase of employment in market services is not favourable (it may be caused by outflow of employed persons from other domains of the economy, e.g. from industry and construction). The variable $\boldsymbol{x}_7$ is still regarded as a stimulant. It is worth noting that for several other pairs of analysed variables similar contradictions occur, but their intensity is much lower.

To obtain uniformity of character of variables it is necessary to convert destimulants and nominants into stimulants. In the further part of this paper we will realize this postulate by inversion of signs of values of destimulants and of “destimulating” part of the nominant $\boldsymbol{x}_8$. As a nominal level of the latter variable we can assume its average value, i.e. 35.

3. Theoretical model of the PCA and SCA methods

Let m and n be natural numbers greater than 1. Assume, that some socio–economic composite phenomenon has to be investigated, which is described by m statistical variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ observed for n objects. Therefore, each variable can be represented as a vector from the $\mathbf{R}^{\mathrm{n}}$ space: $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj}), j = 1, 2, \ldots, m$, and whole data set is reflected by a matrix of $n\times m$ dimension:

$$\mathbf{X} = [x_{ij}],$$

where x_{ij} is a value of observation of variable $\boldsymbol{x}_j$ for i – th object $i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$.

Assume that all the analysed variables were standardized, i.e. arithmetical mean of each of them amounts to 0 and standard deviation is equal to 1. By the same token, it is sufficient and necessary to satisfy the two following conditions:

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^{2} = 1$$

for any $j = 1, 2, \ldots, m$. Denote by $\mathbf{S}$ covariance (and simultaneously — correlation) matrix of investigated variables. Thus, $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/(n–1)$.

Let $1<p<m$ be a natural number. We would like to determine vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p$, (called *factors* or *components*), such that $\boldsymbol{y}_k = (y_{1k}, y_{2k}, \ldots, y_{nk})\in\mathbf{R}^n$, $k = 1, 2, \ldots, p$, and

$$\mathbf{Y} = \mathbf{XW}, \tag{1}$$

where $\mathbf{Y}$ is a $n\times p$ matrix, which elements are coordinates of vectors $\boldsymbol{y}_k$, i.e. $\mathbf{Y} = [y_{ik}]$, $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, p$, and $\mathbf{W} = [w_{jk}], j = 1, 2, \ldots, m, k = 1, 2, \ldots, p$ denotes a $m\times p$ matrix of proceeding coefficients of such transformation called *factor loadings*. Hence, we have $\boldsymbol{y}_k^{\mathrm{T}} = \mathbf{X}\boldsymbol{w}_k^{\mathrm{T}}$, where $\boldsymbol{w}_k = (w_{1k}, w_{2k}, \ldots, w_{mk})$, $k = 1, 2, \ldots, p$. The loadings uniquely determine the components.

Covariance matrix of factors has the following form:

$$\mathbf{V} = \mathrm{cov}(\mathbf{Y}) = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}/(n–1) = (\mathbf{XW})^{\mathrm{T}}(\mathbf{XW})/(n–1) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XW}/(n–1) = \mathbf{W}^{\mathrm{T}}\mathbf{SW}.$$

So, each coefficient v_{kk} lying on the diagonal of the matrix $\mathbf{V}$ denotes variance of factor $\boldsymbol{y}_k$, (called also *reserve of common variation*), and a value $v_{kk}/\mathrm{tr}(\mathbf{S}) = v_{kk}/m$ is said to be the percentage of total variability accounted by this factor, $k = 1, 2, \ldots, p$. If the matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ is diagonal, then the matrix $\mathbf{W}$ is called *orthogonal*. If $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$ (where $\mathbf{I}$ is an identity matrix of respective dimension — here $p\times p$), then we say that loadings of factors are *orthonormal*. Orthonormality of loadings guarantees their linear independency as coefficients of linear combination (1) — and a geometrical significance of the factors — as well as square normalization, i.e.

$$\sum_{j=1}^{m} w_{jk}^2 = 1 \text{ for any } k = 1, 2, \ldots, p. \qquad (2)$$

From practical point of view, we expect that optimally determined factors will be satisfied — besides orthonormality of their loadings — also two another key conditions:

1) **maximization of variation**, what guarantees minimization of loss of variability generated due to summarizing the information into a smaller number of components,
2) **uncorrelation**, guaranteeing the best possible exploitation of information contained in the model. So, owing to this property, it will be possible to analyze each factor separately, as a source of knowledge about other aspect of investigated phenomenon. Factors are *uncorrelated*, if the matrix **V** is diagonal.

The problem amounts then to maximization of trace of covariance matrix of factors, i.e. tr(**V**) = tr($\mathbf{W}^{\mathrm{T}}\mathbf{SW}$), satisfying conditions of: diagonality of this matrix, i.e. $v_{kr} = \mathrm{cov}(\boldsymbol{y}_k, \boldsymbol{y}_r) = \boldsymbol{w}_k\mathbf{S}\boldsymbol{w}_r^{\mathrm{T}} = 0$ for all k, $r = 1, 2, \ldots, p$, $k \neq r$, and orthonormality $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$. The optimal solution in this case is a matrix $\mathbf{W} = \Theta_p$ of first p eigenvectors of the matrix **S**, i.e. related to p successively the greatest eigenvalues $\lambda_1>\lambda_2>\ldots>\lambda_p$ of this matrix. Then the covariance matrix of factors is diagonal: $\mathbf{V} = \mathrm{diag}(\boldsymbol{w}_1\mathbf{S}\boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{w}_2\mathbf{S}\boldsymbol{w}_2^{\mathrm{T}}, \ldots, \boldsymbol{w}_p\mathbf{S}\boldsymbol{w}_p^{\mathrm{T}}) = \Lambda_p = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$. We can obtain these results using, for example, the sequential method of Lagrange multipliers, i.e. by maximization of each element of the diagonal of matrix **V** and adoption of the given conditions. By the way, we can conclude that the postulate of orthogonality of the matrix **W** is redundant. It is sufficient to assume the normalization (2) — and resulting matrix of loadings Θ_p is orthonormal. On the other hand, Θ_p determines the only possible system of factors, which are uncorrelated and have orthonormal loadings. So, the matrix $\mathbf{Y} = \mathbf{X}\Theta_p$ seems to be generally the best. The described procedure is called the *Principal Component Analysis* — PCA.

Unfortunately, the PCA method has an important disadvantage. It concerns a weak interpretability of the principal components. V. Rousson and T. Gasser (2003) think, that there are two main reasons for this fact:

1) for a given component, loadings may be small, medium and large, often with a substantial random element,
2) in most practical cases, PCA produces only one *block–component* (i.e. a component which loadings have the same sign), which is rather easy to interpret as it defines some natural ordering (the situation of subject

matter domain may be "good" or "bad" with respect to such a component). This creation of single block–component is caused by a fact that block–components are generally correlated with each other, whereas principal components are not. The remaining components (called *difference–components*) having some positive and some negative loadings are much more difficult to provide an effective interpretation.

These premises were a basis of concept of the *Simple Component Analysis – SCA*, created by the Swiss biostatisticians from Zurich. They have used some interesting property of the original PCA procedure. Let the correlation matrix **S** be *block – diagonal*, i.e. consists of b square blocks (where b is a natural number lower than m, and each variable belongs exactly to one block) $\mathbf{S}_1$, $\mathbf{S}_2$, …, $\mathbf{S}_b$, and other its elements are equal to zero. (formally, $\mathbf{S} = \mathrm{diag}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_b)$)[1]. V. Rousson i T. Gasser (2004) have noted and proved that for such matrix a system of principal components obtained using the PCA procedure consists of b block – components and m–b difference — components (if k is a natural number, whereas $1<k<m$, then for a $k\times k$ block we have k–1 difference components). We can approximate each correlation matrix by a block — diagonal matrix using grouping of variables according their similarity (measured by the correlation coefficient) and next replacing the smallest coefficients (i.e. lower than the arbitrarily assumed some threshold value) with zero. This way, we decide which variables significantly contribute to an informative worth of a given component and which aspect of analysed problem they represent.

The Swiss scientists have formulated the three following conditions of simple components (assume, that $b<p$):

- **Condition 1.** Each of the first b columns of the matrix **W** contains only values proportional to 0 or 1. These b columns determine a division of the set of variables and b block–components. A nonzero loading corresponds to any variable only for one block–component.
- **Condition 2.** Other $p - b$ columns of the matrix **W** are vectors of natural numbers c, $l<m$: c times the value l, l times the value $-c$ and $m - c - l$ times the value 0. Such columns determine difference–components. Their nonzero loadings are called *contrasts* of variables (because the sum of loadings in each column amounts to zero).
- **Condition 3.** For each of $p - b$ columns of the matrix **W**, the nonzero elements correspond to the same block.

The numerical algorithm leading to obtaining such described loadings (normalized with squares – see (2)) is given as follows:

Stage 1. In order to obtain approximatively block form of the matrix **S** we make a division of the set of variables into disjoint nonempty subsets appropriately to their similarity expressed by a function of the correlation coefficient. In this way, we obtain clusters of similar variables, which create

[1] Then $\mathbf{S}_i$ is a correlation matrix of respective subset of the set of variables, $i = 1, 2, \ldots, b$.

blocks. Here standard procedures of cluster analysis are applied. At the beginning (step 1.), as single clusters, we assume the one–element subsets of the set of variables, i.e. $B_{(1)j} = \{\boldsymbol{x}_{\mathrm{j}}\}$, $j = 1, 2, \ldots, m$. In the t–th step ($t = 2, 3, \ldots$) we join blocks $B_{(t-1)h}$ and $B_{(t-1)g}$, for which a inter–cluster distance given by the formula

$$d_{hg} = 1 - f_{j:\boldsymbol{x}_j \in B_{(t-1)h},\, r:\boldsymbol{x}_r \in B_{(t-1)g}}(s_{jr}) \qquad (5)$$

is the smallest. The function f is generally defined as minimum (single linkage method), median or maximum (complete linkage method). The procedure is conducted until $t>b$ (if b was earlier arbitrarily established) or a correlation level of at least one pair of block – components is greater than arbitrarily assumed threshold value (most often 0.3 or 0.4). The b cluster $B_1, B_2, \ldots, B_b$ received after $t = m- b-1$ steps is the searched optimal division. The loadings of those b simple block–components we can determine as $w_{jk} = |B_k|^{-1/2}$ (where $|B_k|$ is a number of elements of the set B_k) if $\boldsymbol{x}_j \in B_k$ and $w_{jk} = 0$ if $\boldsymbol{x}_j \notin B_k$, $k = 1, 2, \ldots, b$, for every $j \in \{1, 2, \ldots, m\}$.

Stage 2. The following property of principal components will be used: for every $k = 1, 2, \ldots, p$ k–th principal component determined for the set of variables $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ is equal to first principal component obtained using covariance matrix of residuals in regression of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ on $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{k-1}$ (i.e. determined by first eigenvector of the matrix $\mathbf{S} - \hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}}/(n-1)$, where $\hat{\mathbf{X}}$ is a matrix of dependent variables in this regression). Because the best unbiased estimator in this regression (obtained using the classical least square method) is of the form $\hat{\mathbf{X}} = \mathbf{Y}_{k-1}\left(\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{Y}_{k-1}\right)^{-1}\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{X}$, (where $\mathbf{Y}_{k-1}$ is a $n\times(k-1)$ matrix consisting of the first $k-1$ columns of the matrix $\mathbf{Y}$), then covariance matrix of regresants in this model is given as follows:

$$\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}}/(n-1) = \left(\left(\mathbf{Y}_{k-1}\left(\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{Y}_{k-1}\right)^{-1}\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{X}\right)^{\mathrm{T}}\left(\mathbf{Y}_{k-1}\left(\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{Y}_{k-1}\right)^{-1}\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{X}\right)\right)\Big/(n-1) =$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{Y}_{k-1}\left(\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{Y}_{k-1}\right)^{-1}\mathbf{Y}_{k-1}^{\mathrm{T}}\mathbf{X}\right)/(n-1) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Theta}_{k-1}\left(\boldsymbol{\Theta}_{k-1}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Theta}_{k-1}\right)^{-1}\boldsymbol{\Theta}_{k-1}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)/(n-1) =$$

$$= \mathbf{S}\boldsymbol{\Theta}_{k-1}\left(\boldsymbol{\Theta}_{k-1}^{\mathrm{T}}\mathbf{S}\boldsymbol{\Theta}_{k-1}\right)^{-1}\boldsymbol{\Theta}_{k-1}^{\mathrm{T}}\mathbf{S}.$$

Thus, in order to minimize a deviation of simple components from principal components we will apply the following definition: $\mathbf{S}_{res} = \mathbf{S} - \mathbf{S}\mathbf{W}_b\left(\mathbf{W}_b^{\mathrm{T}}\mathbf{S}\mathbf{W}_b\right)^{-1}\mathbf{W}_b^{\mathrm{T}}\mathbf{S}$, where $\mathbf{W}_b$ is a $m\times b$ matrix containing the first b columns of the matrix $\mathbf{W}$. Let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_m)$ be a first eigenvector of this matrix.

Stage 3. For successive values $u = 0$, $|\varphi_1|$, $|\varphi_2|$, …, $|\varphi_m|$ we define so–called *shrunken vectors* $\tilde{\boldsymbol{\varphi}}(u) = (\tilde{\varphi}_1(u), \tilde{\varphi}_2(u), \ldots, \tilde{\varphi}_m(u))$ such that $\tilde{\varphi}_j(u) = \mathrm{sgn}(\varphi_j(u))$, if $|\varphi_j|>u$ and $\tilde{\varphi}_j(u) = 0$ otherwise, $j = 1, 2, \ldots, m$.

Stage 4. We omit these values u, for which $\tilde{\boldsymbol{\varphi}}(u)$ doesn't represent a difference–component. For other shrunken vectors, let $c(u)$ and $l(u)$ be numbers of coordinates of the vector $\tilde{\boldsymbol{\varphi}}(u)$ which are equal to 1 or -1, respectively. Denote by $D_{c,l}(u)$ the largest common divisor of the numbers $c(u)$ and $l(u)$ and define a vector $\tilde{\tilde{\boldsymbol{\varphi}}}(u) = (\tilde{\tilde{\varphi}}_1(u), \tilde{\tilde{\varphi}}_2(u), \ldots, \tilde{\tilde{\varphi}}_m(u))$, such that $\tilde{\tilde{\varphi}}_j(u) = l(u)/D_{c,l}(u)$ if $\tilde{\varphi}_j(u) = 1$, $\tilde{\tilde{\varphi}}_j(u) = -c(u)/D_{c,l}(u)$ if $\tilde{\varphi}_j(u) = -1$ and $\tilde{\tilde{\varphi}}_j(u) = 0$ if $\tilde{\varphi}_j(u) = 0$, $j = 1, 2, \ldots, m$.

Stage 5. We perform a normalization of the vectors $\tilde{\tilde{\boldsymbol{\varphi}}}(u)$, i.e. we determine m – dimensional vectors $\hat{\boldsymbol{\varphi}}(u) = (\hat{\varphi}_1(u), \hat{\varphi}_2(u), \ldots, \hat{\varphi}_m(u))$, such that $\hat{\boldsymbol{\varphi}}(u) = \tilde{\tilde{\boldsymbol{\varphi}}}(u) \Big/ \sqrt{\sum_{j=1}^{m} \tilde{\tilde{\varphi}}_j^2(u)}$.

Stage 6. To the matrix $\mathbf{W}_b$ we add such a vector $\hat{\boldsymbol{\varphi}}(u)$, which maximizes the variance $\hat{\boldsymbol{\varphi}}(u)\mathbf{S}_{res}\hat{\boldsymbol{\varphi}}^{\mathrm{T}}(u)$. It determines a simple difference–component.

Stage 7. If obtained matrix contains p columns, then the algorithm stops. If the number of columns is smaller than p – we return to the Stage 2, taking into account current matrix of loadings.

Optimality of the system of simple components have been assessed by D. Gervini and V. Rousson (2004) by means of the following index

$$\mathrm{CSV}(\mathbf{W}) = \frac{\sum_{k=1}^{p} \left(\boldsymbol{w}_k \mathbf{S} \boldsymbol{w}_k^{\mathrm{T}} - \boldsymbol{w}_k \mathbf{S} \mathbf{W}_{k-1} \left(\mathbf{W}_{k-1}^{\mathrm{T}} \mathbf{S} \mathbf{W}_{k-1} \right)^{-1} \mathbf{W}_{k-1}^{\mathrm{T}} \mathbf{S} \boldsymbol{w}_k^{\mathrm{T}} \right)}{\sum_{k=1}^{p} \lambda_k} \tag{6}$$

This formula is applicable to a broad range of components. Values of the CSV index belong to the interval (0, 1] and their maximum (which amounts to 1) is reached if and only if $\mathbf{W} = \boldsymbol{\Theta}_p$ (because of the fact that the principal components are uncorrelated and hence $\mathbf{W}_{k-1}^{\mathrm{T}} \mathbf{S} \boldsymbol{w}_k^{\mathrm{T}} = \mathbf{0}$).

4. Results of PCA and SCA for labour market data

Now we will apply both of the above presented theoretical models of component analysis to a research of the data set described in the paragraph 2 of this article. It is worth noting that V. Rousson and T. Gasser (2004) have suggested inverting signs of values of some variables. This concept aims at increasing the number of positive correlation coefficients and allows some improvement of quality comparison of the both methods. This proposal (which application results also in a fact that the first principal component is a block–component) seems to be reasonable. But it generates some other inconveniences – one of them is slightly artificial choice of variables to inversion which may disturb interpretability of the results of PCA.

Instead of this approach, we would like to propose a conversion of destimulants and nominants into stimulants done by inversion of signs of their respective values described at the end of the paragraph 2. It reflects a premise that the choice of variables for inversion should depend on direction of their influence on general position of an object. The resulting first principal component may be not a block–component but can have better practical interpretation than in the case of the Swiss way.

Results of computation of loadings of the PCA and SCA components collected in the Table 2 contain first 14 principal components and the same number of simple components. The establishment of this number will have an importance during further modified analysis. The threshold value of correlation between components using in the Stage 1 of the SCA procedure was assumed as 0.3. Clustering was done according to the median method (formula (5) with f established as median). Our computation is based directly on a correlation matrix of the variables, so any additional transformation leading to the standardization of them is not necessary. The Figure 1 presents a comparison of variability extracted by particular components.

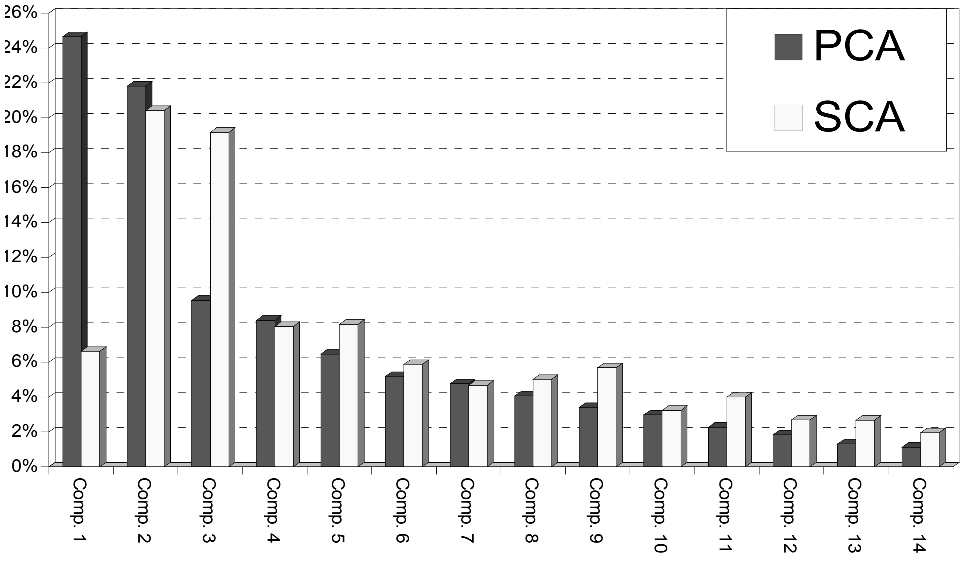
Table 2. Loadings of Principal and simple components for situation of the labor market in 2002 in Polish towns established as NUTS 4 units

Varia-bles	Loadings of components													
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}
PCA														
x_1	-0.138	0.119	0.380	0.088	0.538	0.040	-0.176	0.320	0.208	-0.256	0.368	0.327	-0.001	0.160
x_2	-0.408	0.003	0.219	-0.151	0.153	0.040	0.136	-0.169	-0.057	0.204	-0.072	-0.203	-0.083	-0.024
x_3	-0.226	0.363	-0.122	0.120	0.016	-0.114	-0.058	-0.302	-0.235	-0.168	0.093	0.120	0.268	-0.150
x_4	-0.239	-0.156	0.409	-0.223	0.290	0.171	0.154	-0.285	0.142	0.185	-0.155	-0.145	-0.034	-0.230
x_5	-0.367	0.266	-0.032	0.082	-0.006	-0.036	0.047	-0.071	-0.262	0.070	0.023	0.024	0.168	-0.033
x_6	-0.262	0.151	-0.031	-0.365	-0.358	0.193	0.030	0.073	0.097	0.318	-0.067	0.652	-0.053	0.186
x_7	0.283	0.292	0.159	-0.081	-0.087	0.196	0.167	0.058	-0.153	0.166	0.310	-0.151	-0.523	0.171
x_8	0.316	0.068	0.191	-0.353	-0.004	0.002	0.100	-0.226	-0.372	0.058	0.488	-0.006	0.362	-0.022
x_9	-0.011	-0.340	-0.178	0.273	0.075	0.308	0.311	-0.378	0.009	-0.201	0.187	0.351	0.024	0.078
x_{10}	0.119	-0.328	0.178	0.223	-0.049	-0.171	0.266	0.372	-0.018	0.475	0.101	0.145	0.364	-0.217
x_{11}	0.003	-0.020	0.447	-0.098	-0.510	-0.336	0.173	-0.163	0.361	-0.427	0.059	0.029	0.019	-0.068
x_{12}	0.355	0.033	0.232	0.231	0.115	0.161	0.232	-0.248	-0.153	-0.033	-0.410	0.249	-0.105	0.095
x_{13}	-0.220	-0.339	-0.263	0.009	-0.118	0.167	0.046	-0.122	0.200	0.049	0.475	-0.176	-0.149	0.015
x_{14}	-0.239	0.111	0.091	0.466	-0.061	-0.444	0.268	-0.045	-0.134	0.184	0.143	0.005	-0.304	0.173
x_{15}	-0.132	-0.375	0.020	-0.269	0.071	-0.232	0.084	0.077	-0.320	-0.163	-0.139	-0.082	0.107	0.684
x_{16}	0.075	0.023	-0.361	-0.395	0.341	-0.402	0.431	0.083	0.020	-0.133	-0.007	0.207	-0.257	-0.316
x_{17}	-0.191	-0.291	0.182	0.001	-0.201	0.166	-0.123	0.276	-0.567	-0.304	-0.008	0.117	-0.292	-0.407
x_{18}	-0.136	0.260	-0.060	0.033	-0.082	0.383	0.591	0.395	0.054	-0.267	-0.072	-0.258	0.250	0.054

Varia-bles	Loadings of components													
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}
SCA														
$\boldsymbol{x}_1$	0.236	0.244	0.000	0.298	0.274	0.373	0.000	0.436	0.298	0.354	0.316	0.330	0.000	0.316
$\boldsymbol{x}_2$	0.236	0.000	0.244	0.000	0.274	0.373	0.000	0.000	0.000	0.000	-0.316	-0.275	0.000	0.000
$\boldsymbol{x}_3$	0.236	0.244	0.244	0.000	0.000	0.000	-0.309	-0.327	0.000	0.000	0.000	0.000	0.316	-0.316
$\boldsymbol{x}_4$	0.236	0.000	0.000	0.000	0.274	0.373	0.231	0.000	-0.373	0.000	-0.316	0.000	0.000	0.000
$\boldsymbol{x}_5$	0.236	0.244	0.244	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.316	-0.316
$\boldsymbol{x}_6$	0.236	0.244	0.244	-0.373	0.000	-0.298	0.231	0.000	-0.373	-0.354	0.000	0.330	0.316	0.316
$\boldsymbol{x}_7$	0.236	0.244	-0.342	0.000	0.000	0.000	0.231	0.000	0.298	-0.354	0.316	-0.275	-0.316	0.000
$\boldsymbol{x}_8$	0.236	0.000	-0.342	-0.373	0.000	0.000	0.000	-0.327	0.298	0.000	0.316	-0.275	0.316	0.000
$\boldsymbol{x}_9$	0.236	-0.342	0.000	0.298	-0.456	0.000	0.231	-0.327	0.000	0.354	0.316	0.000	0.000	0.316
$\boldsymbol{x}_{10}$	0.236	-0.342	-0.342	0.298	0.000	0.000	-0.309	0.436	0.000	-0.354	0.000	0.000	0.316	-0.316
$\boldsymbol{x}_{11}$	0.236	0.000	-0.342	0.000	0.274	-0.298	-0.309	0.000	-0.373	0.354	0.000	0.000	-0.316	0.000
$\boldsymbol{x}_{12}$	0.236	0.000	-0.342	0.298	0.000	0.000	0.231	-0.327	0.000	0.000	-0.316	0.330	0.000	0.000
$\boldsymbol{x}_{13}$	0.236	-0.342	0.244	0.000	0.000	0.000	0.231	0.000	-0.373	0.000	0.316	-0.275	0.000	0.000
$\boldsymbol{x}_{14}$	0.236	0.244	0.244	0.298	0.000	-0.298	-0.309	0.000	0.000	-0.354	0.000	-0.275	-0.316	0.316
$\boldsymbol{x}_{15}$	0.236	-0.342	0.000	-0.373	0.000	0.000	-0.309	0.000	0.298	0.000	-0.316	0.000	0.000	0.316
$\boldsymbol{x}_{16}$	0.236	0.000	0.000	-0.373	-0.456	0.373	-0.309	0.000	0.000	0.000	0.000	0.330	-0.316	-0.316
$\boldsymbol{x}_{17}$	0.236	-0.342	0.244	0.000	0.274	-0.298	0.231	0.000	0.298	0.000	0.000	0.330	-0.316	-0.316
$\boldsymbol{x}_{18}$	0.236	0.244	0.000	0.000	-0.456	-0.298	0.231	0.436	0.000	0.354	-0.316	-0.275	0.000	0.000

Source: Author's calculation on the basis of the data described in Table 1 using the SCA package written by V. Rousson and M. Maechler in 2004 as an application of the ***R*** *2.0.1 program, the R Foundation of Statistical Computing, 2004.*

Figure 1. Comparison of variability of PCA and SCA components for labor market in polish towns



*Source: Author's calculation on the basis of the data described in Table 1 using the SCA package written by V. Rousson and M. Maechler in 2004 as an application of the **R** 2.0.1 program, the R Foundation of Statistical Computing, 2004.*

Grey colour marks the values of loadings which absolute values are not smaller than $1/\sqrt{m}$ = 0.236. They reflect variables having the greatest share in information resources provided by a given component.

According to the above results we can formulate several interesting conclusions. The first principal component represented by $\boldsymbol{w}_1$ is not the block–component. For any component there exist loadings both with positive and negative values, also in the case of the "key" level. The resulting principal components are generally not easy to practical interpretation (although they explain 97.9% of total variability of the model). Actually, we observe no serious problem concerning formulation of rational empirical description only for the principal component $\boldsymbol{w}_{14}$. It can be perceived as a contrast between people working in hazardous conditions and a level of possible negative effects of such a job. Some pairs of other components are carriers of very similar information (for example $\boldsymbol{w}_3$ and $\boldsymbol{w}_5$ as well as $\boldsymbol{w}_6$ and $\boldsymbol{w}_7$).

So, the simple components (extracting 83.7% of total variability) are rather much more useful. Their optimality assessed in relation to PCA results and expressed by the CSV index (according to formula (6)) amounts to 85.5%. It is rather difficult to obtain here over 90% optimality (as V. Rousson and T. Gasser (2003)) — mainly due to the fact that in contrast to the biomedical data analysed by them, the socio–economic information is much more methodologically diversified. But the clustering procedure within the first stage of the SCA procedure generates in this case only one cluster (i.e. the full set of all the variables). It is a significant disadvantage from the point of view of practical interpretation of the results. This block–component $\boldsymbol{w}_1$ is then a weighted average of values of variables. Of course, it is also much lower effective than the first and even several next principal components (cf. Fig. 1). The remaining simple components are difference–components and reflect contrasts between some variables. For example, the component $\boldsymbol{w}_4$ describes some aspects of employment in contrast with condition of mobile resources within the population of unemployed persons. It is also worth noting, that the simple component $\boldsymbol{w}_3$ compensates great amount of the loss generated by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.

5. Proposals of improvement of the SCA procedure

As we can conclude on the basis of the results presented in the previous paragraph, the procedure of computation of simple components has several important unfavourable aspects. We will try now to propose some repair method aimed at reduction of such disadvantages, which should not cause any substantial loss of optimality level of the components.

The clustering algorithm used in the first stage of the SCA procedure tends to create trivial clusters of variables. In the case of our labour market data, we have received the full set of them as an optimal cluster. We can find similar observations in many analyses provided by V. Rousson and T. Gasser (2003). The main reason of those problems is the method of establishment of threshold value of inter–group dissimilarity determining optimal clustering. Both arbitrarily fixed maximal number of clusters and allowable level of correlation of components are rather "external" indices, not depending on the actual distribution of distances between clustered variables. It seems to be quite better to use in this context a value of positional statistics of minimal correlation coefficients. Thus, we assume

$$\delta = \operatorname{med}(\boldsymbol{\xi}) + 2.5 \cdot \operatorname{mad}(\boldsymbol{\xi}) \tag{7}$$

where

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m), \quad \xi_i = \min_{\substack{k=1,2,\ldots,m \\ k \neq i}} s_{ik}, \quad i = 1, 2, \ldots, m,$$

and

$$\operatorname{mad}(\boldsymbol{\xi}) = \operatorname*{med}_{i=1,2,\ldots,m} \left| \xi_i - \operatorname{med}(\boldsymbol{\xi}) \right|$$

is the absolute median deviation of the vector $\boldsymbol{\xi}$. The formula (7) is perceived by many statisticians as a positional counterpart of the "two sigma" rule used often in the numerical taxonomy (cf. Z. Hellwig (1968) as well as P. J. Rousseeuw and A. M. Leroy (1987) for example). A key argument for such choice is also a weak sensibility of the statistics (7) to an influence of outliers.

Correctness of clustering can be assessed by the coefficient

$$q = \frac{\mu}{\eta} \tag{8}$$

where

$$\mu = \underset{\substack{k=1,2,\ldots,b \\ |B_k|>1}}{\operatorname{med}} \mu^*(B_k) \quad \text{with} \quad \mu^*(B_k) = \underset{g:\boldsymbol{x}_g \in B_k}{\operatorname{med}} \ \underset{\substack{j:\boldsymbol{x}_j \in B_k \\ j>g}}{\operatorname{med}} \left(1 - s_{gj}\right), \ k = 1, 2, \ldots, b,$$

is a measure of homogeneity of clusters and an index

$$\eta = \underset{k=1,2,\ldots,b}{\operatorname{med}} \eta^*(B_k) \quad \text{with} \quad \eta^*(B_k) = \underset{g:\boldsymbol{x}_g \in B_k}{\operatorname{med}} \ \underset{j:\boldsymbol{x}_j \notin B_k}{\operatorname{med}} \left(1 - s_{gj}\right), \ k = 1, 2, \ldots, b,$$

reflects their heterogeneity. The nearer to zero is the value of this index (8), the better is the quality of grouping.

Loadings of simple block–components related to those clusters are determined according to principal rules of the SCA procedure.

The second main problem connected with the SCA results concerns the difference–components. As we have noted in the introduction, more than one such component may represent the same group of variables. This fact might contribute to make a “legibility” of intra–group relations between variables substantially difficult (loadings of various components reflecting the same cluster for the same variable can “neutralize” one another due to possible opposite sings of values). In our proposal the difference components will uniquely reflect intra–group contrasts. A characteristic feature of this original method is a fact, that nonzero loadings correspond only to the entire variables contained exactly in one group with cardinality greater than one and simultaneously each such group is reflected in exactly one component of this type. The loadings of those components are determined in the following way.

Denote by $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, …, $\boldsymbol{w}_m$ vectors of loadings of all the principal components. Let Ω be a subset of those vectors representing block–components, and B_1, B_2, …, B_{b^*} are clusters of variables generated during the first stage of the SCA procedure containing more than one element (hence $b^* \leq b$). These loadings determine respective simple block — components. Put $Z_0 = \varnothing$. In the k–th step of our procedure we look for an index $r_0 \in \{1, 2, \ldots, m\}$ such that $\boldsymbol{w}_{r_0} \notin \Omega$ and loadings of this component represent the group B_k in a most “visible” way, i.e. the following optimisation equality is satisfied:

$$\left|\sum_{\substack{j=1,2,\ldots,m\\ \boldsymbol{x}_j\in B_k}}\left|w_{jr_0}\right| - \sum_{\substack{j=1,2,\ldots,m\\ \boldsymbol{x}_j\notin B_k}}\left|w_{jr_0}\right|\right| = \max_{\substack{r=1,2,\ldots,m\\ \boldsymbol{w}_r\notin\Omega\cup Z_{k-1}}}\left|\sum_{\substack{j=1,2,\ldots,m\\ \boldsymbol{x}_j\in B_k}}\left|w_{jr}\right| - \sum_{\substack{j=1,2,\ldots,m\\ \boldsymbol{x}_j\notin B_k}}\left|w_{jr}\right|\right| \tag{9}$$

after solution of the equation (9) we assume

$$Z_k := Z_{k-1} \cup \{\boldsymbol{w}_{r_0}\}$$

If all the loadings in $\boldsymbol{w}_{r_0}$ corresponding to the elements of B_k have the same sign we return to (9) and modify Z_k replacing $\boldsymbol{w}_{r_0}$ by next successively nearest vector (according to (9)) satisfying this request or (if such a vector doesn't exist) we omit this step (i.e. we put $Z_k = Z_{k-1}$), $k = 1, 2, \ldots, b^*$.

Next, we make a transformation of vectors belonging to the set Z_{b^*} into loadings of simple components. Let $\boldsymbol{w}_\text{k}$ be an element of this set corresponding to the cluster B_k. We minimize the absolute value of coordinates of this vector. Thus, we find $j_0 \in \{1, 2, \ldots, m\}$ such that

$$w_{j_0k} := \min_{j=1,2,\ldots,m}\left|w_{jk}\right|.$$

So, the respective simple difference — component is defined as $\widetilde{\boldsymbol{w}}_k = (\widetilde{w}_{1k}, \widetilde{w}_{2k}, \ldots, \widetilde{w}_{mk})$, where

$$\widetilde{w}_{jk} = \begin{cases} \left[\dfrac{w_{jk}}{w_{j_0k}}\right] & \text{if } \boldsymbol{x}_j \in B_k, \\ 0 & \text{if } \boldsymbol{x}_j \notin B_k, \end{cases}$$

$k = 1, 2, \ldots, b^*$; $[a]$ denotes an integral part of a real number a, i.e. the greatest integer not greater than a. Finally, coordinates of this vector will be normalized according to the formula described in the Stage 5 of the SCA algorithm.

An application of proposed modification to the analysed aggregated situation of the urban labour market leads to receiving the following optimal clusters of variables (the threshold value computed according to (7) amounts to 0.6331):

- Cluster 1: $\{\boldsymbol{x}_2, \boldsymbol{x}_4\}$;
- Cluster 2: $\{\boldsymbol{x}_3, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_{14}, \boldsymbol{x}_{18}\}$;
- Cluster 3: $\{\boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_{12}\}$;
- Cluster 4: $\{\boldsymbol{x}_9, \boldsymbol{x}_{13}\}$;
- Cluster 5: $\{\boldsymbol{x}_{15}, \boldsymbol{x}_{17}\}$;
- Cluster 6: $\{\boldsymbol{x}_1\}$;
- Cluster 7: $\{\boldsymbol{x}_{10}\}$;
- Cluster 8: $\{\boldsymbol{x}_{11}\}$;

- Cluster 9: $\{\boldsymbol{x}_{16}\}$.

The coefficient of homogeneity of clusters (0.2151) is much smaller than the index of their heterogeneity (1.0100). This fact results in good level of correctness (0.2130). The division is then very effective.

Table 3. contains loadings of simple components computed using our whole proposed procedure. The first 9 vectors represent the block — components as well as presentation of level of variability of particular components.

Similarly as in Table 2., values of loadings, which absolute values are not smaller than 0.236 were marked with grey colour. Variability of most of the modified components is greater than in the PCA and "classical" SCA components (cf. Fig 1.). Precise analysis of diversification of components enable to make an interesting observation that in both the "classical" and modified SCA components extreme loss of variability is accumulated in the first 2—3 block–components. In our proposal, the loss in relation to the standard PCA procedure is slightly greater as in the case of the SCA algorithm (the model explains 86.7% of total variability). The level of the CSV index (70.2%) seems to be rather satisfactory. A main advantage of the results collected in Table 3. is their clarity. Loadings of block–components reflect composition of particular groups and generate components being weighted average of variables within clusters. The loadings of the difference–components give a picture of intra–group relations between variables and show a type of dominating information carried by them. For example, loadings of component represented by the vector $\boldsymbol{w}_{10}$ signal that the greatest influence of the values of it has the variable $\boldsymbol{x}_2$. Is importance is in about 75% balanced by $\boldsymbol{x}_4$. So, interpretation of other difference — components may lead to formulation of also many similar conclusions concerning interactions between variables.

Table 3. Loadings of modified simple components for situation of the labor market in 2002 in Polish towns established at NUTS 4 level

Varia-bles	Loadings of components													
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}
x_1	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
x_2	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.807	0.000	0.000	0.000	0.000
x_3	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.460	0.000	0.000	0.000
x_4	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.591	0.000	0.000	0.000	0.000
x_5	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.106	0.000	0.000	0.000
x_6	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.354	0.000	0.000	0.000
x_7	0.000	0.000	0.577	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.962	0.000	0.000
x_8	0.000	0.000	0.577	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.192	0.000	0.000
x_9	0.000	0.000	0.000	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.054	0.000
x_{10}	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
x_{11}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
x_{12}	0.000	0.000	0.577	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.192	0.000	0.000
x_{13}	0.000	0.000	0.000	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.999	0.000
x_{14}	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.708	0.000	0.000	0.000
x_{15}	0.000	0.000	0.000	0.000	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.594
x_{16}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
x_{17}	0.000	0.000	0.000	0.000	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.804
x_{18}	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.389	0.000	0.000	0.000
Varia-bility	***9.7%***	***14.9%***	***11.4%***	***9.0%***	***8.7%***	***5.6%***	***5.6%***	***5.6%***	***5.6%***	***1.6%***	***6.1%***	***5.5%***	***5.2%***	***2.5%***

Source: Author's calculation.

6. Final conclusions

The SCA method proposed by Swiss biostatisticians is a good theoretical tool, which can be used to conduction of effective component analysis. The optimality in relation to the PCA results in the case of composite socio–economic phenomenon may be not as high as obtained for methodologically more uniform biomedical data. Despite of this fact, the loss of quality is not so significant and seems to be quite satisfactory. Our modification of the SCA procedure can, maybe, generate some additional loss of optimality in relation to the PCA, but an application of it enables obtaining results with much more clear practical interpretation as in the "classical" SCA method. Of course, it is possible that there exist also another methods aimed at the same purpose. Therefore, it seems reasonable to expect that this type of approach might be worth of further exploration. This direction of research may by especially important given the growing demand of data users for complete and versatile information useful for elaborating new economic strategies.

REFERENCES

CSO (2003 a) *Statistical Yearbook of the Labor*, Central Statistical Office of Poland, Warsaw.

CSO (2003 b), *Powiats in Poland*, Central Statistical Office of Poland, Warsaw.

GERVINI D., ROUSSON V. (2004) *Criteria for Evaluating Dimension — Reduction Components for Multivariate Data*, The American Statistician, vol. 58, No. 1, pp. 72—76.

HELLWIG Z. (1968) *Procedure to Evaluating High Level Manpower Data and Typology of Countries by Means of the Taxonomic Method*, Statistical Review, vol. XV, No. 4, pp. 307—327 (in Polish).

HOTELLING H. (1933). *Analysis of a Complex of Statistical Variables into Principal Components*, Journal of Educational Psychology, vol. 24, pp 417—441 and 498—520.

JOLLIFFE I. T. (1995) *Rotation of Principal Components: Choice of Normalization Constraints*, Journal of Applied Statistics, vol. 22, pp. 29—35.

KAISER H. F. (1958) *The Varimax Criterion for Analytic Rotation in Factor Analysis*, Psychometrika, vol. 23, pp. 187—200.

ROUSSEEUW P. J., LEROY A. M. (1987) *Robust Regression and Outlier Detection*, ed. by John Wiley and Sons, New York.

ROUSSON V., GASSER T.(2003) ROUSSON V, GASSER T. (2003) *Some Case Studies of Simple Component Analysis,* Institute for Social and Preventive Medicine, Department of Biostatistics, University of Zürich, Switzerland, typescript (available at http://www.unizh.ch/biostat/Manuscripts/).

ROUSSON V, GASSER T. (2004) *Simple Component Analysis*, Applied Statistics, vol. 53, pp. 539—555.

VINES S. K. (2000) *Simple Principal Components*, Applied Statistics, vol. 49, pp. 441—451.

ZELIAŚ A. (2002) *Some Notes on the Selection of Normalization of Diagnostic Variables,* Statistics in Transition, vol. 5, No. 5, pp. 787—802.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 359—360

Editor's note:

SOME REMARKS ON CHALLENGES FOR PUBLIC STATISTICS FROM INTERNATIONALLY DISPERSED STUDY OBJECT

Most of the newly emerging sources of challenges for public statistics call, first and foremost, for the appropriate methodological framework to be established at the outset of any attempts to overcome them. The case of phenomena that are international in their scope and nature — such as cross-border movements, including economic cooperation (embracing also informal activities), to which we would like to turn attention of the reader of this section — conforms that view markedly. Namely, all the key elements of the statistical research process — from building the frame population through data collection modes to assembling the data in a system of robust indicators — need to be designed in an innovative way, yet assuring international comparison of the quality data sets.

However, despite some of the issues being specific to the area of study focused on transnational movements, many of those issues are either directly embraced by some broader type of problems or have a lot in common with certain problems being already identified or under explicit consideration in other research contexts. It seems worthwhile, therefore, identifying some methodological commonalities appearing in different contexts that might prove useful for dealing with specific ones, like the above cross-border issue. For example, the one of primary importance seems to be the dynamic fuzziness of the universe to be observed, including difficulties with definition of the unit of analysis (Okrasa, 2007)[1]. Structurally, this issue seems to have a lot in common with the problem of fuzzy boundaries encountered in statistics of *new economy* environment, that are often paralleled by difficulties with delineation of geographical boundaries between national and international scopes of business operations, including transnational flows of capital and people, such as job- or income-related migration.

Another example consists in the problem of integrating 'measurable" and non-measurable" within an official account system. It can be approached from different angles — including non-market (e.g., voluntary) activities, next to the market — and needs to be solved in a way. For instance, one may attempt to use

[1] Okrasa, W., 2007, *Possible Research Activities* — presentation at the meeting of Eurostat-ERPROS Working Group „Research Activities in Statistics", December 13 (2007), Luxembourg.

satellite accounts framework as a promising tool to do it also in the international context. Yet another class of approaches turns to analysis of statistical languages, its semiotics, to reduce the deficiencies and gaps in the language of official statistics in different countries, and to promote the language-based advancement toward their consistency and referential validity.

One can consider such an underlying strategy — i.e., searching for structurally identical (from a methodological standpoint) issues through different research contexts and cross-national systems of statistics - a kind of the methodological isomorphism. It seems promising for studying flows of migration (in the perspective of growing massive migration) which are to be far-reaching in terms of both measurable and non-measurable consequences, including such occurring concomitantly with economic cooperation realms as collective action and cultural diversity (Okrasa, 2002)[1].

Despite that only some of the descriptive aspects of the international movements are touched in this section, it sufficiently shows that a larger project seems necessary to be undertaken as an international research venture. Papers presenting cumulated experience or new ideas on those topics are more than welcome. We will publish them in one of the nearest issues of the journal.

Włodzimierz Okrasa

[1] Cultural Diversity, Collective Identity and Collective Action: Towards a Joint Science and Policy Endeavour to Deal with Consequences of the Opening up of National Borders in Europe- Okrasa, W., 2002, Follow-up Report on the ESF Forward Look in the Social Sciences) http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0& file=fileadmin /be_user/research_ areas/social_sciences/areas/social_sciences/documents

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 361—366

CROSS-BORDER SURVEYS — SOME METHODOLOGICAL ASPECTS

Marek Cierpiał-Wolan

ABSTRACT

A growing interest in regional statistics, especially in cross-border cooperation, can be observed in the last years. The reason for this is mainly changes of functions of the borders caused by globalization process. The need for using the results of surveys of cross-border areas on the micro-mezo-macroeconomic level gives rise to establishing a consistent research system for these areas. Despite efforts of several international institutions, there still exist problems with lack of information on particular levels of aggregation as well as with data comparability level in individual countries. As a consequence, there is still a need for identification of major research areas and discussion on important methodological aspects relating to cross-border areas.

1. Introduction

Investigation of processes occurring in cross-border areas is of supranational and multidimensional nature and covers various socio-economic aspects. Regional statistics often comes across different problems of, among other things, limited availability of data for areas located on both sides of the national border, lack of information on a certain level of aggregation in individual countries and low level of data comparability, especially those pertaining to economic issues.

It is essential then, to create a uniform information structure, understood as a set of comparable data and tools for making them available. It is also important to create effective research methods for these areas. The existence of a coherent research system for cross-border areas would allow to effectively use information on local, regional, nationwide and international level.

The aim of the article is to characterize some methodological issues connected with cross-border surveys.

2. Delimitation of cross-border areas

There exist, of course, numerous ideas for setting cross-border area. Thus, its delimitation is pre-arranged to a large extent and depends on the purpose it serves. The delimitation can be carried out based, for instance, on morphological criterion, with geographical features taken mainly into consideration (in particular the lie of the land), or on functional criterion, viewed as a commonly dependent production and consumption actions, and those related to exchange and administration. Delimitation in the institutional sense does not denote a compact area — it creates a kind of a network because it is determined by locations of units cooperating within the cross-border area (units of territorial division on NUTS 4—5 level, entities etc.). As part of these criteria, common problems to solve, e.g. areas of ecological threat or common chances for development, for instance, creation of cross-border touristic area, are also taken into consideration. Groups of indicators reflecting various processes of socio-economic development can be also used for delimitation. It should be stressed, that, as a result of numerous criteria the boundaries of cross-border area are fuzzy since spatial scope of individual features does not coincide with each other.

In practice, preliminary delimitation is carried out first. It is based on limited, not finally coherent list of variables that include law, administrative, political and factual aspects. Systematic analysis of socio-economic phenomena in preliminary defined cross-border area, which mainly focuses on labour market, entrepreneurship, tourism, environmental protection or institutional infrastructure, usually leads to changes of the outlined area, what can be named as dynamic delimitation.

An instance of defining cross-border area is a preliminary delimitation based on three criteria: Regulation No. 1931/2006 of the European Community, dated 20 December 2006, according to which the border zone covers an area of 30—50 km from the border; the rule stating that the smallest administrative unit is a unit of territorial division on NUTS 4 level; and the results of surveys on journeys to work. The first reason for initial corrections of the preliminary delimitation can be analysis of results of sample surveys carried out on borders concerning, for example, goods turnover, in which we ask the respondent about distances from the border to the place of residence and shopping.

3. Monitoring of socio-economic phenomena

In order to maintain comparability, the first step seems to be creation of a uniform set of variables concerning individual socio-economic fields (e.g. demography, entrepreneurship, environmental protection), which will be based on joint glossary of terms, often relating to various classifications.

Only a set of variables prepared in this way allows to use a great number of data analysis methods, such as neuron networks, genetic algorithms, taxonomic methods, classification and regression trees, supporting vector methods or association and sequence methods. In regional statistics, taxonomic methods — hierarchical and non-hierarchical — appears to be particularly useful.

Aiming at comparability, one should not forget about a unique character of individual regions. What seems to be vital in this context, is distinguishing the most important exogenous factors which influence economy and region's development. Combining both approaches enables to create a „ portrait of cross-border region".

4. Data sources

Administrative and statistical databases can be the primary data sources used in monitoring of cross-border areas. Data of customs service and border guards, which are exceptionally important in generalizing results of sample surveys, plays a special role here.

A potential source of information is also outcomes of automatic points of road traffic measurement, which can estimate its intensity (by selected days of a week, month. etc., including seasonal fluctuation), and also cover categories of vehicles. In some countries, the automatic measurement system can precisely identify a vehicle (e.g. registration number, number of persons travelling), and its localization at any point of time.

Collecting of data used in cross-border surveys may be carried out also through bank system. Information on the usage of credit cards might be particularly useful. It should be emphasized, though, that in most countries reports on this matter made by central banks for the purpose of payment balance are of virtually no use for cross-border areas (e.g. in accordance with the EU regulations, transactions not higher than 12.5 thousand euro are not registered). Due to confidentiality of bank data, functioning of an independent information system, powered by commercial banks for regional purposes, appears, however, difficult to realize.

Another and extremely important source of information is certainly sample surveys.

5. Sample surveys

Analysis of processes observed in cross-border areas requires creation of such a system of sample surveys that covers the possibly broadest scope of socio-economic phenomena. Among the most important areas, the following should be mentioned: surveys of household, enterprises, tourist accommodation facilities and questionnaire surveys on borders.

In the household survey, the module concerning changes which occur in the labour market, with focus on non-registered employment (reasons, kind and frequency of starting non-registered work, socio-demographic characteristics of persons performing such work, incomes from non-registered work, etc.) is very important. Another important module is the survey of non-registered shopping level in households (characteristics of households buying in non-registered zone, shopping frequency, amount of expenses on goods from non-registered zone made by households, structure of selected purchased goods, etc.).

The module which is also important, is the one connected with migration of population, of which foreign tourism in particular (aim, time and directions of migration, amount of expenses made. etc.).

In surveys of enterprises located in cross-border areas, a greater attention should be paid to the module of non-registered transactions (size and costs of employment, basic balance data and financial indicators concerning grey area, etc.).

On the other hand, the survey of tourist accommodation facilities should be first and foremost addressed to foreign visitors (number and structure of visitors by country of permanent stay, place of crossing the border and means of transport, aim of travel, kind of accommodation, expenses on goods and services, frequency of crossing the border, duration of stay, etc.)

In most of the countries such surveys are being carried out. Their disadvantage is, though, the size of the sample because it does not allow to generalize the results for cross-border areas delimited on the NUTS 4—5 level. A good example is a survey of economic situation in enterprises and households, economic activity of population or household budgets. Moreover, in different countries they are conducted under different methodology, what results in the lack of comparability of results.

A separate group is questionnaire surveys on borders. One should note, that on land borders, it pertains mostly to countries not covered by liberalization of the rules on requirements as to crossing the border.

The nature of questionnaire surveys on borders concerns both the process of preparing those surveys (it concerns the sample design in particular), and their realization.

In the sample design for cross-border surveys, a two-step approach, namely choosing a unit prior to structuring the frame, seems to be natural. In questionnaire surveys on borders, the total population is often divided into homogenous layers, which make up observation of a given day, from a given crossing, and also, if it concerns land crossings, in accordance with a given way of crossing the border. An important element in preparing a survey is setting the days for carrying it out. The reason for this is the traffic intensity which significantly changes during a year, and the value of observed variables. For sampling the week days, the ones which are non-representative (e.g. national, religious holidays) should not be taken into it. The selection of elements for a sample from each layer is usually made, for practical reasons, by means of systematic sampling. Theoretically, a proper sample design should include sampling of time segments during a day, unless previous analyses of border traffic profiles allow to assume that the population structure is not substantially diverse at that time. For individual border crossings, sampling intervals are usually set. They include expected intensity of visitors traffic on particular crossings and chances for a pollster to carry out the questionnaire survey at a given time.

In order to maintain the quality of collected information and low level of "non response", border surveys require experienced and professional pollsters. The most challenging part of such a survey is to recruit a respondent since they are in the course if journey, often annoyed by awaiting customs and passport clearance for a long time. In such conditions, the pollster usually works under time pressure. Therefore, a selective recruitment of pollsters, intense training courses, and also effective monitoring of their work play significant role. The questionnaire itself is an element of a great importance. To achieve a proper effect, the questionnaire should be designed in the simplest and clear way, namely, it must be relatively short and comprehensible, contain clearly formulated questions, as well as easy to fill in (respondent-friendly).

Due to the nature of questionnaire surveys on borders, it is desirable that pilot test and surveys are prepared with care. They will allow, for the most part, to assess to feasibility of the survey's objectives, sample design, estimation methods, precision, data collection methods and their analysis. At this stage, it is particularly important to prepare a programme for maintaining a proper quality level of each stage of the survey. Sets of indicators for assessing sampling errors and minimizing non-sampling errors are usually used for this purpose. Practice proves that indices of "non response" and procedures for dealing with respondents who refuse to answer are especially important.

6. Summary

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

The functioning of a coherent research system for cross-border areas will provide opportunity to use econometric models, as well as employ the results of analyses on micro-mezo-macroeconomic level.

The establishing of such a system requires factual and organizational cooperation not only within one country, but also on the international level. With the aim of building such a system, one should point at the significance of cooperation between countries in setting survey areas that will include the key socio-economic phenomena present in border areas, taking into consideration the uniqueness of a given region. One should mention also standardization of methods and forms of monitoring of phenomena, as well as conducting joint surveys in these areas.

The above mentioned cooperation could be largely coordinated by public statistics, which fulfils its mission in most of the countries by using and improving modern research methods, actively participating in cooperation in statistics on the international level, as well as coordinating and supporting actions in statistics performed by other state and public administration bodies.

It seems that the first step towards creation of a coherent research system of cross-border areas should be making an inventory of information sources of public statistics, and also other sources of data in individual countries. A detailed analysis of these information, as well as, slight corrections in programmes of already existing surveys, might prove very useful. What should be considered next are chances of using already existing survey sampling by extending the sample so as to make the results for cross-border areas possible to generalize. The basis for this system is, obviously, the establishment of a uniform information infrastructure of cross-border areas (metainformation, databases, methodological reports, etc.). It is worth stressing, that this process of its creation, coordinated by public statistics, can be carried out simultaneously by many international partners on the regional level.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 367—377

REGIONAL STATISTICS IN A TRANSNATIONAL RESEARCH PERSPECTIVE

Andrzej Miszczuk[1]

ABSTRACT

Changing role of state borders impacts on functional and structural transformation of border regions. European integration process causes that borderlands more frequently become a subject of regional statistics. Transborder information exchange is an incentive of overcoming "hostility" of state borders. Regional statistics of borderlands is the information basis of more advanced transborder cooperation forms like preparing and realizing common projects or development strategies. Besides, it creates favourable conditions of spreading and unifying nomenclatures and classifications used in statistical systems of different countries.

Key words: regional statistics, border, border region, transborder cooperation

1. Introduction

Increasingly, regional studies tend to focus on issues concerning border and border zones. Cross-border cooperation is not only an indication of but also an agent that stimulates changes in the functioning of borders and their adjacent areas. Such cross-border cooperation can get off the ground and continue to develop when reliable information on neighbouring regions is provided not only to overcome stereotypes and prejudices that exist on both sides of the borders but also to indicate development opportunities. In several cases, such opportunities give rise to the implementation of joint projects.

In view of the above, structural and functional changes in state borders and border zones as well as cross-border cooperation appear to be a new and very promising area for statistical research, primarily regional statistics, as a tool with which to recognise these changes and apply them in practice. (J. Oleński 1996).

[1] Warsaw University, Warsaw and Statistical Office of Lublin, e-mail: a.miszczuk@uw.edu.pl.

This article aims to present possible structural and functional transformations of the borders and border zones in the light of the European experience and to point out the role of regional statistics in this sphere.

2. The border as a prerequisite to regional development

The traditional concept of borders, largely influenced by F. Ratzel, is understood as clearly defined demarcation lines which determine the space occupied by a state/nation. At the core of this approach, also known as *boundary studies* and dominant in the 1960s, lies an in-depth analysis of factors that determine the shape of such demarcation lines and their evolutionary changes. A contemporary view, known as *border studies,* is a multifaceted phenomenon which defines the border as a social and spatial construct which highlights the existing differences (H.van Houtum 2005) Border studies focus, among other things, on the disappearance of state borders (D.Newman, A.Passi 2001), development of social and physical identity, borders of accessibility and exclusion, and border in various spatial orders.

A contemporary debate on state borders is dominated by a post-modernist approach manifested by the optimistic perception of the disappearance of state borders as spatial barriers. This tendency is an offshoot of globalisation and the free information flow which has effectively replaced *a space of places* with *a space of flows.* This, in turn, means that a nation-state is no longer the subject of business relations in a “borderless world” and a world market devoid of any trade barriers. However, this is the experience of Western Europe, the USA and Canada, since at the beginning of the 1990s post-communist countries saw the rebirth of nation-states and delineation of their clearly defined borders. It is worth remembering that the disappearance of economic borders is not tantamount to the disappearance of borders in other spheres of human activity (D.Newman, A.Passi 2001).

On a regional scale, the specific characteristics of border zones depends on the political systems of the neighbouring countries on the one hand and on the relations between them on the other. This is manifested in the type of border in place. According to O.Martinez (1994), the evolution of state borders comprises four stages: hostility, coexistence, cooperation, and inter-dependence.

In principle, hostility results from violent political events where the existence of the state and the inviolability of its territory and borders are jeopardised. The state border becomes a dividing or disintegrating element, and all inter-state contacts, including cross-border relations are suspended. Such a condition may also result from international sanctions imposed upon a given country (A.Moraczewska 2008).

The shift from hostility to coexistence is a timely process. In the opinion of R.J.Bennett (1997), this process can be facilitated when cooperation triggered by administrative units, including border regions, goes beyond the functional space (economic, social, cultural, etc.) delineated by administrative or national borders.

It is also relatively easier to move from hostility to coexistence where the regional across the border enjoy a great deal of autonomy and self-government. Regional cooperation is difficult to get off the ground where at least one of the states is centrally governed, and the involvement of the central government is required.

The coexistence stage can be defined as the information exchange stage (C.Ricq 1995) which is effected on various levels and among various entities. Exchanging information helps partners across the border get acquainted with each other better, see how their respective public administrations function, what customs and taxation provisions are in place, how business activity is regulated as well as what these partners have on offer in terms of tourism, cultural heritage and the natural environment.

During the cooperation stage cross-border contacts become more intense. They focus primarily on public security, counteracting the effects of natural disasters, crime, education, scientific research, culture, and sport. This stage also gives rise to commerce, including grey economy commerce prompted by price differences across the state border.

During the coexistence and cooperation stages of border evolutionary changes, the function of the border can be described, to use the word coined by J.Rosenau, as fragmengration, i.e. openness to selected areas of activity or diversification of this openness towards individual countries (A.Moraczewska 2008).

During the inter-dependence stage of the border evolution bilateral relations across the border become even closer through the development of technological and capital links, movement of workforce, and joint undertakings of business partners. With its integrating function, the border gradually becomes invisible, which, in fact, is the desired objective. Its full realisation is possible when economic integration of countries has been reached within the framework of an economic or customs union and the common market.[1].

The shift from a closed border which functions as a dividing line through the filtering border to the open border that joins the countries is a long and not necessarily one-way process. Reverting to the previous stages in an abrupt and violent manner is not so uncommon.

Interestingly, the opening of the state border is not always associated with benefits to the regions located in the border zone. The cohesion, brain drain, insularity and exclusion effects[2] are some of the adversary consequences of such a situation. The cohesion effect consists in economic integration and economic growth stimulation in an area extending some 40 to 60 km into the given country. The prerequisites of cohesion include:

[1] For more information on integration stages see F.Machlup (1986).

[2] Cf.: The impact...(1996).

- favourable topographical conditions which facilitate easy traffic across the border,
- a relative density of the population on both sides of the state border,
- absence of towns within the proximity of at least 100 km from the border which would attract migration. Integration on both sides of the border is fostered by the development of the service sector.

The brain drain effect appears when the border zone suffers from inadequate development potential including enterprising people and capital resources. An unfavourable topography along the border can also stimulate this effect, especially where the opening of the border gives a boost to the development of urban agglomerations located at a distance. This urban sprawl is usually effected at the cost of the border zone.

The third effect, known as "insularity" refers to the border zone where the development of an urban agglomeration due to historical and geographical factors occurred despite the national border that separated its parts. Such an agglomeration, located on the territory of two or three states, will continue to develop autonomously and independently of other parts of these states.

The exclusion effect occurs when the opening of the national border becomes a development incentive for one part of the border zone only. Such a situation results from the fact that individual parts of the border zone do not constitute a complementary structure and social and economic potential.

3. Cross-border cooperation as a border and border zone evolution agent

Under the European Outline Convention on Transfrontier Cooperation between Territorial Communities or Authorities Done at Madrid on 21 May 1980 (known as the Madrid Convention), cross-border cooperation is defined as any concerted action designed to reinforce and foster neighbourly relations between territorial communities or authorities within the jurisdiction of two or more Contracting Parties and the conclusion of any agreement and arrangement necessary for this purpose. Transfrontier co-operation shall take place in the framework of territorial communities' or authorities' powers as defined in domestic law.

Cross-border cooperation is a manifestation of international regional cooperation, and where this cooperation is carried out within the terms of reference of an institutionalised cross-border region, it can be described as Euroregional cooperation.

The objectives of cross-border cooperation, as defined in the European Charter for Border and Cross-Border Regions include:

- ensuring a new quality of borders: meeting spaces,
- smoothing out the interfaces of European spatial development policy,

- overcoming border-related disadvantages and exploiting opportunities by improving infrastructure, and promoting locational quality and common economic development,
- improving cross-border protection of the environment and nature,
- promoting of cross-border cultural cooperation,
- making realities of subsidiarity and partnerships.

Cross-border cooperation is most frequently manifested in the following areas (C.Ricq 1995): the natural environment, physical development planning, transport and communication, public security and protection against the effects of natural disasters, economy, and employment, tourism, education and culture, border infrastructure and cross-border traffic.

However, there are some obstacles to the establishment and further development of cross-border cooperation. These include, among others, topographical barriers (rivers, mountain chains, etc.), geo-political issues (memberships in various political groups, contradictory aims of foreign policies, etc.) institutional and organisational distance between the regions and sub-regional units, which translates into incompatibility of respective competencies of local and regional authorities, as well as ethnic and demographic problems (e.g. depopulation of border zones, ethnic conflicts which may give rise to border skirmishes or even territorial disputes). Differences in economic development levels may, with the opening of national borders, bring unilateral benefits. Equally important are prevalent national stereotypes created when the border was closed as well as a language barrier.

Cross-border cooperation brings numerous tangible and intangible benefits defined as its added value. This comprises the following:

- *The European added value,* which arises from the fact that people who are living together in neighbouring border regions want to cooperate and thereby make a valuable contribution to the promotion of peace, freedom, security and the observance of human rights,
- *The political added value* involves making a substantial contribution towards the development of Europe and European integration, the implementation of subsidiarity and partnership, and increased economic and social cohesion and cooperation,
- *The institutional added value* entails active involvement by the citizens, authorities, political and social groups on both sides of the border, secure knowledge about one's neighbour, long-term cross-border cooperation in structures that are capable of working efficiently, joint drafting, implementation and financing of cross-border programmes and projects,
- *The socio-economic added value* becomes apparent in the respective regions through the mobilisation of endogenous potential by strengthening the regional and local levels as partners for and initiators of cross-border cooperation, the participation of actors from the economic and social

sectors, and the opening up of the labour market and harmonisation of professional qualifications,

- *The Socio-cultural added value* is reflected in lasting, repeated dissemination of knowledge about the geographical, structural, economic, socio-cultural and historical situation of a cross-border region (with the assistance of the media), the overview of a cross-border region afforded in maps, publications, teaching material, the development of a circle of committed experts, and equal opportunities and extensive knowledge of the language of the neighbouring country.

4. Scope and functions of regional statistics concerning cross-border areas

It is generally agreed that regional statistics is a part of public statistics which presents social and economic phenomena across administrative and/or statistical units of the administrative division (S.Godowski 2003). The even growing importance of regional statistics in EU member states is attributed to regional decentralisation processes, which classify states into federations and unitary states and which determine the position of regions in these countries. Taking into consideration the effects of decentralisation of public authorities on the legal and territorial structure of the state, J.Loughlin (1999) studied EU member states and suggested the following classification of states: federal, regional unitary, decentralised unitary, and centralised unitary states. The factors which determine the position of administrative regions within the framework of these types of states include election of regional authorities directly, the right to participate in designing the national policy, the right to conclude international treaties, and the right oversee political and legal dealings of subregional authorities (Table 1).

Table 1. Features of administrative regions in various types of states

Type of state	Features of administrative regions:			
	election of the authorities	participation in designing national policy	conclusion of international treaties	political/legal oversight over subregional authorities
Federal	+ +	+ +	+	+
Regional unitary	+ +	+	–	–
Decentralised unitary	+ + / +	–	–	–
Centralised unitary	–	–	–	–

Legend: + + high intensity, + moderate intensity, – no feature present.
Source: Author's own study based on J.Loughlin (1999).

It follows that in political and legal terms four types of administrative regions can be distinguished[1],i.e. autonomous regions in federal states, autonomous regions in unitary states, self-governing regions in unitary states, and administrative and functional regions in unitary states.

The determination of the position of the administrative region is a prerequisite to determining the objectives and tools of inter- and intra-regional policies adopted for such a region. The former issue boils down to the traditional dilemma: egalitarianism or efficacy (G.Gorzelak 1989), which, to use a more balanced language (G.Gorzelak 1998), regional improvement or support of the weaker or improvement nation-wide (balancing the spatial order). The latter reflects two approaches to the regional policy (J.Hausner 2001), i.e.

- centralised and balanced, or
- decentralised and competitive.

The latter approach is also more convergent with the aim of the intra-regional policy, i.e. improving the competitive potential of the region depending on its endogenic resources, and most notable, its human resources.

The adoption of the first model of the regional policy means that public statistics is dominated by centralised and sector-based approach, and regional statistics is nothing more than the result of the breakdown of national information selected in terms of the requirements of inter-regional policy.

The adoption of the other model in which regions enjoy a wide scope of competencies, including legislative powers, and are key players of the regional policy, indicates that regional statistics becomes an important tool with which development processes are regulated. The scope of regional statistics is primarily

[1] Cf.: J.Sługocki (1996).

geared towards the needs of the local authorities rather than the central government.

It is worth mentioning that material changes in the country's administrative division occur less frequently in decentralised states. In centralised states such phenomena are far more common, which hinders or makes it literally impossible to conduct retrospective statistical analyses where information is collected each time in compliance with the binding administrative division and not on the basis of the topographic paradigms which allow more freedom in grouping such information.

Experience gained so far clearly points out that regional statistics should fulfil three basic functions. Firstly, regional statistics should fulfil cognitive (informative) function by providing relevant information to domestic and international databases as well end-users (government and local administration, domestic and foreign entrepreneurs, universities, higher education facilities, etc.). This information comes primarily from public censuses, and comprehensive statistical research, representative studies and official registers which all allow spatial desaggregation.

Secondly, regional statistics should perform a methodological function relating to the defining and harmonising statistical categories and classifications applied in various territorial cross-sections (regional or local).

The third function — i.e. applicability — consists in processing basic information and developing a system of descriptive, analytical, and effective indicators and in conducting regional studies for the needs of business entities and public administration in the area of creating and upgrading development plans and in implementing projects, including those co-financed by the EU, stipulated in these plans.

The role of these functions is of paramount importance in the case of statistics for border zones, since at least two systems of public statistics are juxtaposed with each other. The experience gained by EU member states in the area of cross-border cooperation shows that the first stage of border evolution, i.e. the coexistence stage, hinges heavily on the exchange of information between border zones. It follows that the cognitive (informative) function of regional statistics is by far the most important one during the first stage of border evolution. Its efficacy largely determines the time required for the application function to come to the fore (the cooperation and inter-dependence stages). This, in turn, is prompted by joint business undertakings and the preparation of development projects and strategies.

The methodological function determines the efficacy of cognitive and application functions of border zone statistics. During the course of information exchange between border zone regions, a lot of problems occur (T.Borys 1999) in connection with:

- the quality of data and their relevance, their reliability and verifiability, spatial representativeness (in studies based on random sampling), their completeness and topicality,
- accessibility of data with respect to the entire border zone and its individual parts, their possible breakdown into various spatial cross-sections and the need of statistical confidentiality,
- comparability of data in terms of data collection methods, terminology applied and time periods which these data refer to.

It seems that the third problem listed above is most acute indicating a pressing need for a comparative analysis of concepts and classifications employed. Such an analysis will help distinguish three groups of concepts and classifications, i.e. (T.Borys 1999):

- full compatibility of methods employed,
- incomplete compatibility of methods employed and possible comparability of concepts and classifications,
- total incompatibility of methods employed and inability to make any viable comparisons.

The biggest disparities in this respect usually occur in such areas as environmental protection (classification criteria of purity and contamination of natural resources, and legal forms of environmental protection), economy (labour market, unemployment, classification of business activity) or social infrastructure (education and health care systems). However, the presence of these disparities serves as an incentive to developing international cooperation in harmonising various areas of statistics (J.Witkowski 2001). Apparently, significant progress has been made in overcoming these disparities in regions located within the EU borders, and in regions bordering with associated states or countries which aspire to become member states, though to a lesser degree. The disparities are most acute between those countries which do not make efforts to develop a uniform European system of public statistics.

5. Final remarks

An important attribute of statehood, the national border may not necessarily constitute a barrier to the movement of people or goods. Cross-border cooperation aims to gradually dismantle such a barrier offering numerous benefits to border zone regions located on both sides of the national border. The ultimate aim of cross-border cooperation is to make the national border invisible.

Given integration processes well underway across Europe, border zones become increasingly the objects of regional statistics. Cross-border exchange of statistical information contributes to overcoming the “hostility” of national borders. Regional statistics is also an important source of information for more advanced forms of cross-border cooperation schemes such as the preparation and execution of joint projects or the creation of common development strategies.

In terms of methods used and terminology applied, border zones are a challenge for regional statistics and an important step in popularizing and harmonizing domestic statistical classifications and nomenclature.

REFERENCES

BENNETT R.J., 1997, Administrative Studies and Economic Space, „Regional Studies" 31.3, p.323—336.

BORYS T., 1999, Obszary transgraniczne w statystyce regionalnej, „Statystyka w praktyce" t.6, GUS, Warszawa.

GODOWSKI S., 2003, Statystyka regionalna w Polsce. Harmonizacja z wymaganiami Unii Europejskiej [w:] Polityka regionalna Unii Europejskiej a statystyka regionalna — zadania polskiej statystyki urzędowej w tym zakresie, GUS, Warszawa (maszynopis powielony).

GORZELAK G., 1989, Rozwój regionalny Polski w warunkach kryzysu i reformy, UW. Warszawa.

GORZELAK G., 1998, Podstawowe pojęcia polityki regionalnej, „Reforma Administracji Publicznej", z.21, s. 15—30.

HAUSNER J., 2001, Modele polityki regionalnej w Polsce, „Studia Regionalne i Lokalne" nr 1, s. 5—24.

HOUTUM van H. 2005, The Geopolitics of Border and Boundaries, „Geopolitics", 10, p. 672—679,

The impact of the development of the countries of Central and Eastern Europe on the Community territory, 1996, „Regional Development Studies", 16.

LOUGHLIN J., 1999, Regional and local democracy in the European Union, Committee of Regions , Brussels.

MACHLUP F., 1986, Integracja gospodarcza — narodziny i rozwój idei, Warszawa, PWN.

MORACZEWSKA A., 2008, Transformacja funkcji granic Polski, Wydawnictwo UMCS, Lublin 2008.

MARTINEZ O., 1994, The dynamics of border interaction: new approaches to border analysis [in:] C.H.Schofield (ed.): Global Boundaries, World Boundaries, vol. I, Routledge, London, p. 1—15,

NEWMAN D., PASSI A., 2001, Rethinking Boundaries in Political Geography [in:] M.ANTONISCH, V.KOLOSSOV, M.P.PAGNINI (eds): Europe

Between Political Geography and Geopolitics, Societa Geografica Italiana, Roma, Vol. I, p. 301—316.

OLEŃSKI J., 1996, Statystyka transgraniczna — zarys problemów [w:] A.MISZCZUK, R.WIŚNIEWSKI (red.): Informacyjno-infrastrukturalne uwarunkowania współpracy transgranicznej, seria wydawnicza: „Euroregion Bug", t. 2, Norbertinum, Lublin, s. 21—22.

RICQ C., 1995,Handbook on transfrontier co-operation for local and regional authorities in Europe, Council of Europe, Strasbourg,.

SŁUGOCKI J., 1996, Pozycja prawnoustrojowa regionu: regiony w Europie Zachodniej, WSP, Olsztyn.

WITKOWSKI J., 2001, W nowy wiek z nową statystyką. Refleksje z 53 Sesji Międzynarodowego Instytutu Statystycznego, „Wiadomości Statystyczne" nr 11, s. 1—12.

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 379—380

THE 5TH INTERNATIONAL CONFERENCE ON SURVEY SAMPLING IN ECONOMIC AND SOCIAL RESEARCH

8—10 September 2008, Katowice, Poland

The conference was held at the Faculty of Management of the University of Economics in Katowice (Poland). It was organized by the Department of Statistics at the University of Economics in Katowice in co-operation with Department of Statistical Methods of Łódź University and Polish Statistical Association.

The Scientific Committee consisted of: Andrzej Barczak, Czesław Bracha, Czesław Domański (chair), Lorenzo Fattorini, Zdzisław Hellwig, Jan Kordos, Walenty Ostasiewicz, Jan Paradysz, Jan Steczkowski, Jacek Wesołowski, Janusz Wywiał.

Conference participants, representing universities, statistical agencies and opinion poll companies, came from eight countries. At the conference, three invited lectures and 18 papers were presented. The conference was organized to give an opportunity to present latest developments in survey sampling and related fields and to exchange experience on practical applications of survey sampling.

Topics discussed during the conference included:

- Estimation of population parameters based on complex samples
- Statistical inference based on incomplete data
- Small area estimation
- Sample size and cost optimization in survey sampling
- Sampling designs
- Statistical inference using auxiliary information
- Model-based estimation
- Longitudinal surveys
- Practical applications of survey sampling

Abstracts of the papers are available at http://web.ae.katowice.pl/metoda.

The invited lectures were presented by:

a. **Malay Ghosh** (University of Florida, USA), *Bayes and Empirical Bayes Benchmarking with Applications to Small Area Estimation,*
b. **Nicholas T. Longford** (Pompeu Fabra University in Barcelona), on *Small-area Estimation with Spatial Similarity*, and
c. **Yves Tillé** (The University of Neuchâtel), *Balanced Bampling, Principles, Algorithms and Applications*.

The list of authors and titles of contributed paper are given below:

1. Wojciech Gamrot, *On sampling from overlapping strata with pre-determined minimal sampling fractions*
2. Krzysztof Jakóbik, Grzegorz Ruta, *The effectiveness of CATI method in the survey "Structure of agricultural holdings"*
3. Jan Kordos, *Neglected stages of survey design, implementation and analysis*
4. Barbara Kowalczyk, Emilia Tomczyk, *Survey non-response and properties of expectations of industrial enterprises — analysis based on business tendency surveys*
5. Jan Kowalski, *Stationary model in rotation schemes*
6. Danutė Krapavickaitė, *Estimation of some proportion*
7. Jan Kubacki' Bartosz Grancow, Alina Jędrzejczak, *An example of empirical best linear unbiased predictor (EBLUP) application for small area estimation in Polish household budget survey*
8. Olha Lysa, *Approaches to improving of monthly LFS estimates reliability*
9. Andrzej Mantaj, Wiesław Wagner, *Remarks on formula of second order inclusion probability in the simple random sampling with replacement.*
10. *Salah Merad, Evaluating methodological changes in the ONS business register and employment survey using a pilot survey*
11. Aleksandras Plikusas, *On some ratio type estimators of the finite population total*
12. Renato Salvatore, *Selection of covariates in small area estimation with multilevel models*
13. Marcin Szymkowiak, *Calibration estimators in surveys with nonresponse*
14. Wiesław Wagner, *The representative method in research of the domestic tourism according to UN WTO*
15. Janusz L. Wywiał, *Assessing total in small area sampling by means of test-predictor*
16. Jacek Wesołowski, *Linear estimation and prediction under model-design approach with small area effects*
17. Agnieszka Zięba, *Poverty indicators — estimation at NUTS 4 level*
18. Tomasz Żądło, *On prediction of domain total based on unbalanced longitudinal data*

The conference was sponsored by the Polish Statistical Association and SPSS Poland. The sixth Conference on Survey Sampling in Economic and Social Research will take place in 2010.

Prepared by Tomasz Żądło (Department of Statistics, University
of Economics in Katowice).

STATISTICS IN TRANSITION-new series, September 2008
Vol. 9, No. 2, pp. 381—383

PUBLIC STATISTICS IN THE PROCESS OF EUROPEAN INTEGRATION IN VIEW OF THE TRANSNATIONAL CO-OPERATION

Lublin, 22—24 September, Lublin

The International Scientific Conference, entitled: *„Public Statistics in the Process of European Integration in View of the Transnational Co-operation"* was held this year in Lublin, on 22—24 September, as part of the celebration of the 90th Anniversary of the Central Statistical Office.

The honorary patronage over the conference organised by the Statistical Office in Lublin, College of Enterprise and Administration in Lublin (WSPA), and Maria Curie-Skłodowska University in Lublin was assumed by Józef Oleński, Prof. PhD hab. — President of the CSO, Genowefa Tokarska — Voivode of Lublin, and Krzysztof Grabczuk, PhD — Marshal of Lubelskie Voivodship. The scientific committee was chaired by the Vice-Chancellor of College of Enterprise and Administration in Lublin — Andrzej Miszczuk, PhD hab., Prof. of WSPA.

The conference was attended by representatives of central offices, local government and voivodship government, consular services of Ukraine and Euroregions, as well as by scientists and statisticians from central and regional statistical offices from Belarus, Russia, Slovakia and Ukraine. In total, more than 120 persons attended the conference.

Scientific discussions were conducted in the following thematic blocks, in the course of which as much as 24 scientific lectures were given:

- Polish Public Statistics within the European Statistical System,
- Cross-Border Regions — Methodological and Application-Related Issues of European Regional Statistics,
- The Eastern Borderland of the European Union as the Issue in Focus of Public Statistics.

Three conference sessions comprised lectures given by: Jerzy Kłoczowski, Prof. PhD hab. (*The Issues of Integration and Border Zone in the European Union*), Teresa Śmiłowska, Ph.D. (*Harmonization of Statistical Classifications as an Element of Quality and Comparability of Data*), Beata Bal-Domańska, Ph.D. (*Regional Data Bank as an Element of National and European System of Statistical Information*), Krzysztof Kałamucki, Ph.D. (*The Chief Sanitary Inspectorate as a Regional Statistics Tool*), Marian Żukowski, Prof. PhD hab. (*Public Statistics as the Data Source for Information Systems in Banks*), Wojciech Janicki, Ph.D. (*Allocation of International Migrants to the Regions of Immigration Countries — an Attempt to Estimate the Missing Data*), Andrzej Miszczuk, PhD hab., Prof. of WSPA (*Evolution of Border and Borderland Functions and Regional Statistics — an Attempt of Model Approach*), Sławomir Banaszak, MA (*Statistics of Cross-Border Areas — Methodological Dilemmas*), Marek Cierpiał-Wolan, Ph.D. (*Methodological Aspects Cross-Border Research*), Georgij Chwalko (*Cross-Border Information-Oriented Cooperation: Objectives, Tasks, and Development Prospects*), Wacław Wierzbieniec, Ph.D hab. (*General Censuses as a Source of Information on Cross-Border Areas*), Tomasz Komornicki, Ph.D. (*Trans-border Interactions Studies as a Challenge for Public Statistics*), Waldemar A. Gorzym-Wilkowski, Ph.D. (*Trans-Border Aspects of Spatial Planning — the Role of Public Statistics*), Władysław Wiesław Łagodziński, MA (*International Exchange and Access to Information in Practice*), Jacek Szlachta, Prof. PhD hab. and Bogdan Kawałko, MSc. Engineer (*Challenges Related to the Cross-Border Land Development of Poland, Belarus and Ukraine, in the Context of the European Neighbourhood Policy 2007—2013*), Krzysztof Hetman, Ph.D., Ryszard Boguszewski, MA (*Information Needs in the Process of Creating the Strategy of Polish — Ukrainian — Belarusian Borderland Development*), Semen Matkowski, Prof. PhD hab. (*Cross-Border Land vs. Methodological Consistency of the Ukrainian, Belarusian and Polish Regional Statistics*), Mieczysław Kowerski, Ph.D. (*The Influence of the Schengen Acquis Implementation on Economic Trends Among Inhabitants and Entrepreneurs in Lubelskie Voivodship*), Małgorzata Flaga, Ph.D. (*Contemporary Demographic Processes in the Polish-Ukrainian Borderland*), Piotr Witkowski, Ph.D. (*Cross-Border Trade in Commodities in the Context of the Polish Membership in European Union Illustrated with an Example of the Polish-Belarusian and Polish-Ukrainian Borderland*), Krzysztof Jakóbik, Ph.D. (*Estimation of Foreign Trade Turnover Scale at Regional Level — Voivodship's*), Janusz Gajda, Prof. PhD hab. Engineer and Ewa Szkic-Czech, Ph.D. (*Parameters of Measurement Systems of the Road Traffic as the Source of Statistical Information Essential for the Development of Transgenic Areas*).

Publication of the reviewed collection of these lectures in the special edition of *Statistical Information* will serve as the summary of the conference.

Poland's accession to the European Union has considerably changed the situation in the eastern border zone. On the one hand, the accession created a wide array of opportunities but, on the other hand, it gave rise to several challenges related to the previous forms of cooperation and to previously established contacts. Taking certain steps aimed at stimulating and promoting integration between Polish and foreign statistical services, as well as at harmonising cooperation between these services and scientific, local government and economic environments was thus indispensable.

To this end, the representatives of both domestic and foreign statistical services attending the conference signed the Letter of Intent concerning the cooperation related to conduction of common statistical surveys at the cross-border, supranational, and multinational level. Collaboration in the scope of the information-based support of regional development and cross-border cooperation in the regions located at the eastern border of the European Union constituted the objective of the aforementioned Letter.

Signatures on the document were affixed by representatives of the Central Statistical Office, statistical offices in Białystok, Kraków, Lublin, Olsztyn and Rzeszów, as well as foreign statistical offices in Brest and Hrodna (Belarus), Košice, Prešov and Žilina (Slovakia), Kaliningrad (Russia) and Lviv (Ukraine).

Study by: Elżbieta Łoś