Variance Estimation in Multi-phase Calibration

Noam Cohen*, Dan Ben-Hur and Luisa Burck

February 12, 2015

Abstract

The derivation of estimators in a multi-phase calibration process requires a sequential computation of calibrated weights of previous phases in order to obtain those of later ones. Already after two phases of calibration the estimators and their variances involve calibration factors and regression remainders from both phases and the formulae become cumbersome and uninformative. As a consequence the literature so far deals mainly with two phases while three phases or more are rarely being considered. The analysis in those cases is ad-hoc for the specific design and no comprehensive methodology for constructing calibrated estimators, and even more challengingly, estimating their variances in three or more phases was formed and thus in most cases does not exit. We provide a closed form formula for the variance of multi-phase calibrated estimators that holds for any number of phases of calibration. This new estimator of the variance is not only general for any number of phases but also has some favorable characteristics.

KEY WORDS: Multi-phase sampling, Generalized regression.

1 Introduction

Survey statistics makes use of available auxiliary information on known population totals in order to improve survey estimates. A calibration estimator uses calibrated weights which are as close as possible, according to a given distance measure, to the initial sampling design weights, while also satisfying a set of constraints induced by the auxiliary information. Arbitrary sampling designs are allowed at all phases of sampling and the auxiliary information can be used at any phase and is incorporated in the estimation process in each phase.

Multi-phase sampling along with calibration to known auxiliary information is a powerful and cost effective technique. Rao (1973) and Cochran (1977, ch. 12) provided the basic results for stratification and non-response in two-phase sampling. The process of calibration has been extensively studied and a detailed framework of the linear weighting approach in two-phase sampling appears in Särndal et al. (1992) chapter 9. More related to our work, Breidt and Fuller (1993) gave efficient estimation procedures for three phase

*Noam Cohen, Dan Ben-Hur and Luisa Burck, Statistical Methodology Department, The Central Bureau of Statistics, 95464 Jerusalem, Israel.

sampling in the presence of auxiliary information and Hidiroglou and Särndal (1998) studied the use of auxiliary information for two-phase sampling while allowing a minor modification in the distance function that results with additive *g-factors* rather than multiplicative ones. A common characteristic of all these results is the presentation of last phase calibrated weights via calibrated weights of previous phases. This is a major drawback as it requires computation of weights of all former phases in order to obtain those of later ones and as a consequence makes it difficult to provide a well established methodology of how to estimate the variance of the calibrated estimators in designs with more than two phases. The special case of two phases is an exception that was elaborately investigated.

To address this problem we use the modification of the generalized least squares (GLS) distance function, introduced by Hidiroglou and Särndal (1998), to provide a new presentation of the vector of multi-phase calibrated weights which are presented solely through the initial weights based on the sampling design and does not include calibrating factors (also known as g-factors). From this presentation we derive a consistent estimator for the variance of multi-phase calibrated estimators that holds for any number of phases of calibration.

2 Notation

Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$. A first phase probability sample $s_1 (s_1 \subseteq U)$ is drawn from the population U, using a sampling design that generates the selection probability π_{1k} for the k'th unit in the population. Given that s_{i-1} has been drawn, the i'th phase sample $s_i (s_i \subseteq s_{i-1})$, is selected from s_{i-1}, with sampling design with the selection probabilities $\pi_{ik|s_{i-1}} \equiv Pr(k \in s_i | k \in s_{i-1})$. Note the conditional nature of the consequent phase selection probabilities. From this point on, we work only with weights in the estimation process. The conditioned i'th phase sampling weight of unit $k \in s_i$ and its overall sampling weight will be denoted by $w_{ik} = 1/\pi_{ik|s_{i-1}}$ and $w^*_{ik} = \prod_{j=1}^{i} w_{jk}$ respectively.

Let y_k be the value of the target variable for the k'th population unit with which an auxiliary vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{pk})$ is associated. Denote by y the vector of elements of the target variable obtained at the last phase of sampling, p. The population total of $\mathbf{x}$, $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$ is assumed to be unknown. However, some totals may be known from relatively accurate sources such as census data or other types of administrative files. Without loss of generality let $\mathbf{x_1}$ be the vector of variables known for all units in the population U. Let $\mathbf{x_2}$ be the vector of variables obtained in the first phase sample s_1, and so on. For elements in $s_r, r \leq p$ the complete information is thus summarized in the vector $\mathbf{x} = (\mathbf{x_1}', \mathbf{x_2}', \ldots, \mathbf{x_r}')'$. Denote also $t_i = t_{\mathbf{x_i}}$.

The auxiliary information available at each phase of sampling can be used to obtain improved weights by the process of calibration which will produce calibration factors to be used in the estimation process. We use the superscript "*" to denote overall weights, *i.e.*, weights taking all phases into account. The super-imposed symbol "$\sim$" denotes calibrated weights. The i'th phase calibration factors are denoted as g_{ik}, resulting with i'th phase

calibrated weights $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ for $k \in s_i$, where $\tilde{w}_{i-1,k}$ are the calibrated weights of the *i-1*'th phase, and $\tilde{w}_{0k} = 1$. For $k \in s_i$ the calibration with respect to all phases produces overall calibration factors denoted as g^*_{ik}. As a result we will have overall calibrated weights $\tilde{w}_{ik} = w^*_{ik} g^*_{ik}$ where w^*_{ik} is the overall sampling weight.

3 Calibration with GLS distance

Calibration requires the specification of a distance function measuring the distance between the initial weights and the new calibrated weights. Several distance functions have been studied, see a selected summary in Deville and Särndal (1992). We will use the GLS distance function, introduced by Hidiroglou and Särndal (1998). This distance finds the values $\tilde{w}_{ik}$ for the set $k \in s_i$, that minimize the expression

$$\sum_{k \in s_i} \frac{(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik})^2}{w^*_{i-1,k} w_{ik}} \tag{1}$$

subject to

$$\sum_{k \in s_i} \tilde{w}_{ik} x_{ik} = \sum_{k \in s_{i-1}} \tilde{w}_{i-1,k} x_{ik} \tag{2}$$

where $\{\tilde{w}_{i-1,k} : k \in s_i\}$ are the initial weights at the beginning of phase i, *i.e.*, the calibrated weights obtained at phase $i-1$ and $\{\tilde{w}_{ik} : k \in s_i\}$ are the calibrated weights for phase i that we want to obtain. The weights resulting from this calibration scheme are $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ where $g_{ik} = 1 + (\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il})' T_i^{-1} x_{ik}$ with $T_i = \sum_{l \in s_i} w^*_{il} x_{il} x'_{il}$. The special characteristics of this distance is that the calibration factors in this process operate additively so the overall calibrated weights resulting from minimizing (1) subject to (2) are

$$\tilde{w}_{pk} = w^*_{pk}(g_{1k} + \ldots + g_{ik} + \ldots + g_{pk} - (p-1)) \tag{3}$$

for $k \in s_p$. Denote w_i the vector with components $w_{ik}; k = 1 \ldots n_i$, and D_i a diagonal matrix of size n_i with w_i on its diagonal. The same notation will be used with the vectors w^*_i and $\tilde{w}_i$. Let $\hat{B}^+_{ij} = \left(\sum_{k \in s_i} w^*_{ik} x_{ik} x'_{ik}\right)^{-1} \sum_{k \in s_j} w^*_{jk} x_{ik} x'_{jk}$ and $\hat{B}^-_{ij} = \left(\sum_{k \in s_i} w^*_{ik} x_{ik} x'_{ik}\right)^{-1} \sum_{k \in s_{j-1}} w^*_{j-1,k} x_{ik} x'_{jk}$ be estimators for $B_{ij} = (\sum_{k \in U} x_{ik} x'_{ik})^{-1} \sum_{k \in U} x_{ik} x'_{jk}$ the regression coefficient of $\mathbf{x_j}$ on $\mathbf{x_i}$. The difference between the two estimators is that while $\hat{B}^-_{ij}$ uses the entire set of units known for $\mathbf{x_j}$ which is obtained in s_{j-1}, $\hat{B}^+_{ij}$ uses only the subset $s_j \subseteq s_{j-1}$ and thus more variable than $\hat{B}^-_{ij}$. Let $\hat{Z}_{ij} = \hat{B}^+_{ij} - \hat{B}^-_{ij}$ the difference between the two regression estimates which is consistent to zero. Denote also $\hat{Z}_{i_1 i_2 \ldots i_k} = \prod_{j=2}^{k} \hat{Z}_{i_{j-1} i_j}$ for $k \geq 2$ and $\hat{Z}_{i_1} = 1$ for $k = 1$. Let $\hat{t}^-_i = \sum_{k \in s_{i-1}} w^*_{i-1k} x_{ik}$ and $\hat{t}^+_i = \sum_{k \in s_i} w^*_{ik} x_{ik}$ be two Horvitz-Thompson estimators for t_i, based on the units obtained in samples s_{i-1} and s_i respectively. Note that all the estimators defined in this paragraph use overall design weights w^* and not calibrated weights.

3.1 Estimation

The motivation to our next analysis comes from the recursive nature of $\tilde{w}_{ik}$ in (3), where calibrated weights of previous phases $1,\ldots,i-1$ are nested in each g_{ik}. In the following lemma we provide a presentation of $\tilde{w}_p$, the vector of calibrated weights after p phases of calibration, that relies solely on the pre-known sampling design weights $\{w_i^*\}_{i=1}^p$.

Lemma 3.1 *Consider a multi-phase sampling design with a calibration scheme that produces additive g-factors as in (1). A presentation of the calibrated weights at phase p that is based entirely on the design weights is*

$$\begin{aligned}\tilde{w}_p &= D_p^{*'}1_{n_p} + \sum_{i_1=1}^{p} A_{i_1} - \sum_{i_1<i_2}^{p} A_{i_1 i_2} \\ &+ \ldots + (-1)^{k+1} \sum_{i_1<\ldots<i_k}^{p} A_{i_1 i_2 \ldots i_k} + \ldots + (-1)^{p+1} A_{i_1 i_2 \ldots i_p} \end{aligned} \tag{4}$$

where

$$A'_{i_1 i_2 \ldots i_k} = (\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+})' \hat{Z}_{i_1 i_2 \ldots i_k} (X'_{i_k} D^*_{i_k} X_{i_k})^{-1} X'_{i_k} D^*_p.$$

Proof. We use brute force to substitute each g-factor with its expression and repeat this process until no calibrated weights are left. The proof is omitted. □

The calibrated weight $\tilde{w}_p$ therefore equals to $D_p^{*'}1_{n_p}$, the overall design weight, plus correction terms of lower orders of magnitude, and maintains the familiar characteristic of calibrated weights. Let y be some variable of interest for which we want to estimate the population total Y. Denote $\hat{\beta}_j = \left(X_j D_j^* X_j\right)^{-1} X_j D_p^* y$, the regression coefficient of y on $\mathbf{x_j}$, and $\hat{Y}_{HT_p} = 1'_{n_p} D_p^* y$ the non-calibrated Horvitz-Thompson estimator computed over the elements in s_p. Rearranging the terms in (4) produces a more conventional presentation of the multi-phased calibrated estimator $\tilde{w}'_p y$ as a "one-phase" multi-variate regression estimator

$$\tilde{w}'_p y = \hat{Y}_{HT_p} + \sum\nolimits_{i_1=1}^{p} (\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+})' \hat{\gamma}_{i_1} \tag{5}$$

where

$$\begin{aligned}\hat{\gamma}_{i_1} &= \hat{\beta}_{i_1} - \sum\nolimits_{i_1<i_2}^{p} \hat{Z}_{i_1 i_2} \hat{\beta}_{i_2} + \\ &\ldots + (-1)^{k+1} \sum\nolimits_{i_1<\ldots<i_k}^{p} \hat{Z}_{i_1 i_2 \ldots i_k} \hat{\beta}_{i_k} + \ldots + (-1)^{p-(i_1-1)+1} \hat{Z}_{i_1 \ldots p} \hat{\beta}_p.\end{aligned}$$

This presentation in equation (5) now enables a computation to produce an innovative consistent estimator for the variance of multi-phase calibrated estimators.

Theorem 3.1 *Let* $\hat{e}_{rk} = x'_{rk}\hat{\gamma}_r - x'_{r+1,k}\hat{\gamma}_{r+1}$ *for* $r < p$ *and* $\hat{e}_{pk} = x'_{pk}\hat{\gamma}_p - y_k$. *A consistent estimator for the variance of* $\tilde{w}'_p y$ *is*

$$\sum_{1\leq r_1,r_2\leq p}\sum_{k\in s_{r_m},l\in s_{r_M}}\frac{w^*_{r_Ml}}{w^*_{r_ml}}\left(w^*_{r_mk}w^*_{r_ml}-w^*_{r_mkl}\right)\hat{e}_{r_mk}\hat{e}_{r_Ml} \tag{6}$$

where $r_m=\min(r_1,r_2)$ *and* $r_M=\max(r_1,r_2)$.

Proof. The proof repeats the steps used in the case of one-phase multi-variate calibration. It involves evaluation of the highest orders of magnitude and the estimation of their variance. Special care is given to the evaluation of the joint probability of events $\{k\in s_i, l\in s_j\}$ and estimation of the covariance between units from different phases of sampling. See appendix. □

3.2 Example: Two-phase calibration

We will use the special case of two-phase calibration ($p=2$) to demonstrate our methodology and the modifications from other estimators commonly used in literature. The calibrated estimator is given according to (5) by

$$\tilde{w}_2'y=\hat{Y}_{HT_2}+(\hat{t}_1^-\hat{t}_1^+)'\hat{\gamma}_1+(\hat{t}_2^-\hat{t}_2^+)'\hat{\gamma}_2$$

where $\hat{\gamma}_1=\hat{\beta}_1-\hat{Z}_{12}\hat{\beta}_2$ and $\hat{\gamma}_2=\hat{\beta}_2$. This estimator produces the same estimates as the two-phase calibrated estimator used in Hidiroglou and Särndal (1998) or in Särndal et al. (1992) section 9.7. But once one has computed the estimates to the parameters $\hat{\gamma}_i$, the presentation of $\tilde{w}_2'y$ becomes simple and informative, having the structure of a simple multi-variate regression estimator. $\hat{\gamma}_i$ expresses the overall impact of the calibration to the variable $\mathbf{x_i}$ on the estimation of Y. It takes into account the projection of y on $\mathbf{x_i}$, the projection of y on $\mathbf{x_{i+1}}$ multiplied by the projection of $\mathbf{x_{i+1}}$ on $\mathbf{x_i}$ and so on. Moreover, as we will now show, the variance estimators differ significantly both in value and presentation. The common estimator for the variance used up to now in literature was given according to

$$\begin{aligned}\hat{V}_C(\tilde{w}_2'y) &= \sum_{k,l\in s_2}w_{2kl}(w_{1k}w_{1l}-w_{1kl})(g_{1k}\check{e}_{1k})(g_{1l}\check{e}_{1l})\\ &\quad+\sum_{k,l\in s_2}w_{1k}w_{1l}(w_{2k}w_{2l}-w_{2kl})(g_{2k}\check{e}_{2k})(g_{2l}\check{e}_{2l})\end{aligned} \tag{7}$$

where the error factors are $\check{e}_{1k}=y_k-x_{1k}'\hat{\gamma}_1$ and $\check{e}_{2k}=y_k-x_{2k}'\hat{\gamma}_2$ both defined for $k\in s_2$ because y is observed only at s_2. The g-factors are defined the same as in our analysis. On the other hand, the variance estimator suggested in (6) for two-phase calibration is given by

$$\begin{aligned}\hat{V}_P(\tilde{w}_2'y) &= \sum_{k,l\in s_1}(w_{1k}w_{1l}-w_{1kl})\hat{e}_{1k}\hat{e}_{1l}+\sum_{k,l\in s_2}(w^*_{2k}w^*_{2l}-w^*_{2kl})\hat{e}_{2k}\hat{e}_{2l}\\ &\quad+2\sum_{k\in s_1,l\in s_2}\frac{w^*_{2l}}{w_{1l}}(w_{1k}w_{1l}-w_{1kl})\hat{e}_{1k}\hat{e}_{2l}.\end{aligned} \tag{8}$$

The difference in the variance estimator between the two methods represented by equations (7) and (8) is fundamental. It is expressed in a couple of layers. While the error factor of the second phase in both methods is the same, *i.e.*, $\hat{e}_{2k} = \check{e}_{2k}$, the error factor of the first phase differ. $\check{e}_{1k}$ is based on the difference between y_k and the regression predictor $x'_{1k}\hat{\gamma}_1$ while $\hat{e}_{1k}$ is based on the difference between the predictors of phases one and two $x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$. This modification enables the first summand in (8) to be computed over s_1 and not s_2 where the sample is larger. Moreover, the estimator (8) has a third summand which involves the product of the two error factors from both phases that has no parallel in (7). Although this product will often be close to zero whenever the error factors are not strongly correlated, it may still not be negligible whenever y has a very strong correlation with $\mathbf{x_1}$. An evident advantage is the absence of the g-factors which make the estimator much simpler to compute. That is, once we have computed the parameters estimates $\hat{\gamma}_i; i = 1 \dots p$, the estimator (8) can be computed using design parameters only. From an operational point of view, maybe most important is that (8) has the advantage that in the vast majority of designs the second summand constitutes the absolute majority of the variance while the summands in (7) are usually of the same order of magnitude. This characteristic stems from the fact that the term $(w^*_{2k}w^*_{2l} - w^*_{2kl})$ that involves the total sampling weights is very large in comparison with $(w_{1k}w_{1l} - w_{1kl})$ or $w_{1k}w_{1l}(w_{2k}w_{2l} - w_{2kl})$. The expression $w_k w_l - w_{kl}$ attains its maximum on the diagonal $k = l$ where it is proportional to w_k^2 which increases dramatically its value when it depends on total weights w^* instead of w. The second summand may therefore be a good estimator of the variance of the calibrated estimator practically on its own.

A typical pattern of the comparison between the two variance estimators in the special case of two-phase calibration is presented in Figure 1. It can be seen that in most realizations the difference between the two variance estimators is very small, although on a certain one it can reach up to 20%. The mean value of both estimators for the variance was very similar, namely, 54.17 and 54.65, while the true value of that specific simulation data was 54.46. This pattern repeated itself for all variables studied. We did not investigate this specific data or the special case of two-phase calibration any further as our objective was to provide a consistent estimator for the variance that holds for any number of phases of calibration. Another simulation study demonstrated an excellent estimation for the variance of a three-phase calibrated estimator for all variables examined. A comparison to other estimators in three or more phases was not preformed because alternative estimators to the variance in those cases do not exist.

Appendix

Proof of theorem 3.1 $\hat{B}^+_{ij}, \hat{B}^-_{ij}$ are both consistent to B_{ij}. Write $\hat{B}^+_{ij} = B_{ij} + (\hat{B}^+_{ij} - B_{ij})$ so $\hat{B}^+_{ij} = B_{ij} + O_p(n_j^{-1/2})$. Recall that $\hat{Z}_{ij} = \hat{B}^+_{ij} - \hat{B}^-_{ij}$ where $\hat{B}^-_{ij}$ is based over s_{j-1} while $\hat{B}^+_{ij}$ over its subsample s_j and thus also $\hat{Z}_{ij} = O_p(n_j^{-1/2})$ and $\hat{Z}_{i_1 i_2 \dots i_k}$ is bounded by $O_p(n_{i_k}^{-1/2})$. Likewise $\hat{\beta}_j$ is $\beta_j + O_p(n_p^{-1/2})$ because y is observed only at the last phase of sampling s_p. Hence $\hat{\gamma}_i$ is consistent for γ_i for all i, where $\hat{\beta}_i$ in $\hat{\gamma}_i$ are replaced by β_i in γ_i. Consistency does not necessarily imply the convergence of the moments and specifically

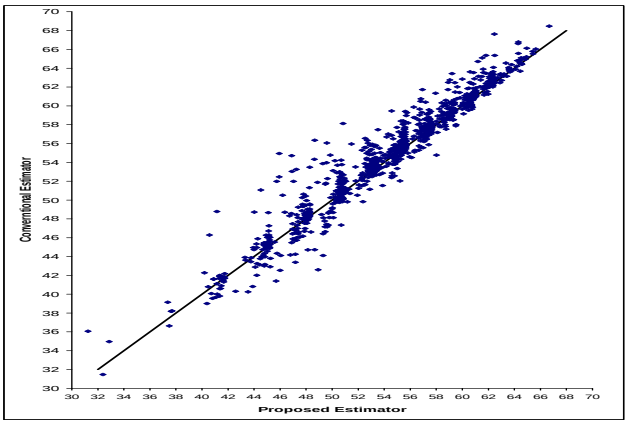

Figure 1: Variance estimates in two-phase calibration. 1000 realizations of the proposed estimator (eq. 8) Vs the conventional estimator (eq. 7) for the variance of the calibrated estimator of Y. The solid line is the main diagonal.

not of the variance. However, for a finite population, *i.e.*, a finite probability space, the concepts coincide. It follows that for n_p large enough $Var(\hat{Y}_{HT_p}+\sum_{i_1=1}^{p}(\hat{t}_{i_1}^{-}-\hat{t}_{i_1}^{+})'\hat{\gamma}_{i_1})$ and $Var(\hat{Y}_{HT_p}+\sum_{i_1=1}^{p}(\hat{t}_{i_1}^{-}-\hat{t}_{i_1}^{+})'\gamma_{i_1})$ are asymptotically equivalent and following the above discussion the difference can be quantified by

$$Var(\tilde{w}_p'y) = Var(\hat{Y}_{HT_p}+\sum\nolimits_{r=1}^{p}(\hat{t}_r^{-}-\hat{t}_r^{+})'\gamma_r) + N^2 o(n_p^{-1}).$$

The estimator $\hat{t}_r^{+}$ is a summation over units in s_r while $\hat{t}_r^{-}$ is over s_{r-1}. Rearranging the terms, the variance on the right hand side can be written as $Var(\sum_{r=1}^{p}\sum_{i\in S_r} w_{ri}^{*}e_{ri})$ which is equal to

$$\sum_{1\le r_1,r_2\le p}\sum_{k\in U}\sum_{l\in U} w_{r_1k}^{*}e_{r_1k}w_{r_2l}^{*}e_{r_2l}Cov(I_{k\in s_{r_1}},I_{l\in s_{r_2}})$$

so a sample based estimator would be

$$\sum_{1\le r_1,r_2\le p}\sum_{k\in s_{r_1},l\in s_{r_2}} w_{r_1k}^{*}\hat{e}_{r_1k}w_{r_2l}^{*}\hat{e}_{r_2l}[1-\frac{P(k\in s_{r_1})P(l\in s_{r_2})}{P(k\in s_{r_1},l\in s_{r_2})}]. \tag{9}$$

To compute the covariance between the indicators $I_{k\in s_{r_1}}$ and $I_{l\in s_{r_2}}$ we need to know the joint probability of events $\{k\in s_i, l\in s_j\}$. If $s_j\subset s_i$, then $P(k\in s_i, l\in s_j)$ equals the joint probability that both units k,l are in sample $s_i = s_{\min(i,j)}$, multiplied by the conditional probability that unit l is in sample s_j given that it belongs to s_i. Formally, if $s_j\subset s_i$ then $P(k\in s_i, l\in s_j) = \frac{w_{il}^{*}}{w_{jl}^{*}}w_{i,lk}^{*-1}$, hence eliminating the dependence on s_{r_2} in the brackets in (9) and the result follows. □

References

[1] Breidt, J. and Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyà*, 55, 297-309.

[2] Cochran, W.G., (1977). *Sampling Techniques,* 3rd edition. New-York, Wiley.

[3] Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, No. 418, 376-382.

[4] Farell, P.J., and Singh, S. (2002). Penalized chi-square distance function in survey sampling. *Proceedings of Joint Statistical Meeting, NY, USA*.

[5] Hidiroglou, M.A. and Särndal C.E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

[6] Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 6, 125-133.

[7] Särndal, C.E., Swensson B. and Wretman J., (1992). *Model Assisted Survey Sampling*. New-York, Springer-Verlag.