APPLICATION OF BALANCED SAMPLING, NON-RESPONSE AND CALIBRATED ESTIMATOR

Ieva Dirdaitė[1], Danutė Krapavickaitė[2]

[1]Pandaconnect, UAB. Address: Saulėtekio al. 15, Vilnius, 10224, Lithuania
[2]Vilnius Gediminas Technical University. Address: Saulėtekio al. 11, Vilnius, 10223, Lithuania
E-mail: [1]dirdaite.ieva@gmail.com, [2]danute.krapavickaite@vgtu.lt

Received: August 2016 Revised: October 2016 Published: November 2016

Abstract. The aim of this paper is to study the interplay between balanced sampling, non-response and calibrated estimator by simulation. The results of seven strategies, embracing a combination of balanced sampling via the cube method, simple random cluster sampling, adjustment for non-response, Horvitz–Thompson estimator of the total and calibration of design weights, are compared. Auxiliary information is used for all strategies at least at one of the stages (sampling or estimation). This auxiliary information consists of indicator variables for sex, age groups and urban/rural living area, and their totals. Real Labour Force Survey data of Statistics Lithuania are used for simulation. Bias, variance and relative mean squared error are measures of accuracy for the comparison of results.

Keywords: auxiliary information, adjustment for non-response, cube method, simulation.

1. Introduction

The idea of balanced sampling is very old and goes back to the beginning of survey sampling. In a way, it has already been used in the works of Kiaer [7]. Despite the fact that this concept evolved and was touched by many survey statisticians, it became known for a wide community of survey statisticians after the book [6] by Y. Tillé was published and its author has given a lot of talks at the conferences introducing sampling design, which he called "balanced sampling design".

The values of auxiliary variables are used in this method at the stage of sample selection. Another method, which uses the values of auxiliary variables at the estimation stage, is calibration of design weights. It was introduced by J.-C. Deville and C.-E. Särndal in 1992 [1]. This method became very popular at the statistical offices of many countries and is often used especially for estimation in social surveys.

The aim of the current paper is to study by simulation the use of balanced sampling and calibration together and the effect introduced by non-response into the process of sample selection and estimation.

2. Statistical methods

2.1. Probability sampling

Let us study a finite population $\mathcal{U} = \{1,2,...,N\}$, with the study variable y defined for elements of the population $y_1, y_2, ..., y_N$. The parameter of interest is a total $t_y = \sum_{k\in\mathcal{U}} y_k$.

Any subset $s = \{i_1, i_2, ..., i_n\} \subset \mathcal{U}$ is called a sample selected from a finite population. A random sample S from the finite population is called *probabilistic sample* if

a) the elements of the set of all possible samples $\mathcal{S} = \{s_1, s_2, ..., s_V\}$ (realizations of S) can be enumerated, and to any possible sample a probability of its selection $p(s_v) = P(S = s_v)$, $v = 1,2,...,V$, is attached, so that

$$\sum_{v=1}^{V} p(s_v) = 1;$$

b) any element of the population belongs to at least one possible sample;

c) technical possibility is available for the selection of indicated samples with the indicated probabilities.

Distribution $p(\cdot)$ is called a *sampling design*. The population $\mathcal{U}$ may be a set of clusters, primary sampling units, consisting of population elements – secondary sampling units. Probability $\pi_k = P(s \in \mathcal{S} : k \in s)$, $\pi_k > 0$, is called element *inclusion probability* to the sample, $d_k = 1/\pi_k$ is called a design weight.

The population total may be estimated from the probability sample by a Horvitz–Thompson estimator ([4]):

$$\hat{t}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k}, \tag{1}$$

which is unbiased, with the variance

$$Var(\hat{t}_{HT}) = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

The estimator of variance

$$\widehat{Var}(\hat{t}_{HT}) = \sum_{k \in S} \sum_{l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

is unbiased for $\pi_{kl} > 0$.

A fixed-size sampling design, assigning selection probability $p(s) = 1/C_N^n$ to any n size collection s of different elements and $p(s) = 0$ to any other collection of elements, is called simple random sampling. It is sampling design without replacement with inclusion probabilities $\pi_k = n/N$, $\pi_{kl} = n(n-1)/(N(N-1))$ for $k \neq l$, $k,l \in \mathcal{U}$.

2.2. Balanced sampling

Let us take a vector of auxiliary variables $\mathbf{x} = (x^{(1)}, x^{(2)}, ..., x^{(J)})'$ with the values $\mathbf{x}_k = (x_k^{(1)}, x_k^{(2)}, ..., x_k^{(J)})'$, $k \in \mathcal{U}$, known for the whole population before the sample selection. If this vector characterizes a study variable in the population, it is natural to seek for such a version of random sample S for which the Horvitz–Thompson estimator of the population total of auxiliary variables remains equal to the true total:

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = t_{\mathbf{x}}, \quad t_{\mathbf{x}} = \sum_{k=1}^{N} \mathbf{x}_k. \tag{2}$$

Let us suppose inclusion probabilities $\pi_1, \pi_2, ..., \pi_N$ are given. According to [6], the *sampling design* $p(\cdot)$ is said to be *balanced* with respect to auxiliary vector $\mathbf{x} = (x^{(1)}, x^{(2)}, ..., x^{(J)})'$ if it satisfies (2).

In the case of a social survey such kind of design can imply, for example, that Horvitz–Thompson estimates of the population size for some groups are equal to the true values if the components of the auxiliary vectors are chosen as indicators of these groups.

A question arises: is it always possible to select balanced samples, satisfying (2) for given variables $x^{(1)}, x^{(2)}, ..., x^{(J)}$? An exhaustive answer to this question is given in [6], [7] and many other papers by Deville and Tillé.

Tillé suggested an algorithm for the selection of a balanced sample, which is called a *cube method* because of the geometric representation of balanced sampling design by a random walk on an N-dimensional cube. First of all, let us note that for a fixed sample size n and only one auxiliary variable x with the values $x_k = \pi_k$, $k \in \mathcal{U}$, any probability sample will be balanced with respect to this variable, because the balancing equation is satisfied for any sample:

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in \mathcal{U}} I(k \in S) = \sum_{k \in \mathcal{U}} \pi_k = n.$$

Now, let us express a sample S as a vector of indicators $\mathbf{s} = (I_1, I_2, ..., I_N)'$ with

$$I_k = I(k \in S) = \begin{cases} 1, & k \in S, \\ 0, & k \notin S. \end{cases} \tag{3}$$

Such a sample can be represented as a vertex of an N-dimensional unit cube. Then balancing equations (2) may be rewritten as

$$\sum_{k\in\mathcal{U}} \frac{\mathbf{x}_k I_k}{\pi_k} = t_{\mathbf{x}}$$

or

$$\sum_{k\in\mathcal{U}} \frac{x_k^{(j)} I_k}{\pi_k} = t_{xj}, \quad j = 1,2,...,J.$$

These balancing equations should be satisfied.

In order to construct a balanced sample, we should come from the opposite direction and look for the vector $\mathbf{a} = (a_1, a_2, ..., a_N)'$, which satisfies the system of equations

$$\sum_{k\in\mathcal{U}} \frac{x_k^{(j)} a_k}{\pi_k} = t_{xj}, \quad j = 1,2,...,J. \tag{4}$$

This is a system of J linear equations with N unknowns $a_1, a_2, ..., a_N$, and it follows from linear algebra so that, without restriction of the generality, the equation system can be rewritten as

$$\sum_{k=1}^{J} \frac{x_k^{(j)} a_k}{\pi_k} = t_{xj} - \sum_{k=J+1}^{N} \frac{x_k^{(j)} a_k}{\pi_k},$$

and for any choice of $N-J$ components $a_{J+1}, ..., a_N$, a unique solution $a_1, ..., a_J$ of (4) exists, if $\det(\mathbf{A}) \neq 0$ with

$$\mathbf{A} = \begin{pmatrix} x_1^{(1)}/\pi_1 & ... & x_1^{(J)}/\pi_1 \\ ... & ... & ... \\ x_J^{(1)}/\pi_J & ... & x_J^{(J)}/\pi_J \end{pmatrix}.$$

Generally, a solution $\mathbf{a}$ of an equation system (4) may consist of any numbers. If we succeed to find a solution $\mathbf{a}$ consisting of components equal to 0 and 1, this solution will give us a balanced sample: $\mathbf{s} = \mathbf{a}$. Otherwise, a balanced sample is not selected, and a vector $\mathbf{I} = (I_1, I_2, ..., I_N)'$, $I_k = 0$ or $I_k = 1$, $k = 1,2,...,N$, $I_1 + ... + I_N = n$, close to the solution $\mathbf{a} = (a_1, ..., a_N)'$ in a way, should be found. The vector $\mathbf{I}$ will indicate a sample $\mathbf{s} = \mathbf{I}$ which is approximately balanced and its finding is named a *rounding problem*. Unfortunately, a problem to find the vector $\mathbf{I}$ is often faced in practice. The cube method is one of the solutions to this problem. The method consists of two phases: flight phase and landing phase. The flight phase means a solution of the linear equation system (4) by the random walk starting at the point $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_N)'$ and stoping at the point $\mathbf{a} = (a_1, a_2, ..., a_N)'$, which satisfies the equation system (4) and is on the ridge of the N dimensional cube. The landing phase – rounding the solution $\mathbf{a}$ obtained in the flight phase to the closest vertex of the cube $\mathbf{I} = (I_1, I_2, ..., I_N)'$ if the flight phase did not give it. The balanced sample cannot be reached exactly if one of the constraints in (4) is a fixed sample size and sum of the inclusion probabilities is not an integer: $\sum_{k\in\mathcal{U}} \pi_k \neq n$. A solution of the rounding problem by the cube method is presented in [6], [7] theoretically, and also implemented into a software R package *sampling* [8] practically. It has to be mentioned that the elements of balanced sample have predefined inclusion probabilities; therefore a Horvitz–Thompson estimator of the total can be used. The estimator of variance for this estimator without using joint inclusion probabilities is also given in [7].

2.3. Calibrated estimator

Let us suppose a probability sample s is selected and data from its elements are collected. Let us suppose we have a vector of auxiliary variables $\mathbf{x} = (x^{(1)}, x^{(2)}, ..., x^{(J)})$ with the values of the sampled elements and known population totals $t_{\mathbf{x}} = (t_{\mathbf{x}1}, ..., t_{\mathbf{x}J})'$. Let us fix a sample $s \in \mathcal{S}$. *A calibrated estimator* of the total t_y is such an estimator $\hat{t}_{yw} = \sum_{k\in s} w_k y_k$, whose weights w_k, $k \in s$, satisfy the requirements:

a) w_k differ as little as possible from the design weights $d_k = 1/\pi_k$ in the sense of the distance function

$$L(w_k, d_k, q_k, k \in s) = \sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k} \to \min,$$

q_k are freely chosen constants;

b) calibration equations are valid

$$\hat{t}_{\mathbf{x}w} = \sum_{k \in s} w_k \mathbf{x}_k = t_{\mathbf{x}}. \tag{5}$$

We see that the calibration equation (5) is similar to the balancing equation (2). The difference is in the weights; also all values of $\mathbf{x}$ are needed for balancing before sample selection; the values of $\mathbf{x}$ for the selected elements only and totals $t_{\mathbf{x}}$ are needed for the calibration of the design weights at the estimation stage. The connection between these two methods is widely discussed in [7].

2.4. Dealing with non-response

Non-response is unavoidable in any real survey. Probabilities for the population elements to respond to the survey questionnaire may be equal or non-equal. When non-response occurs, the bias of the estimator for a population parameter is almost unavoidable. Then the aim of the statistician is to select an estimator of the parameter with the bias which is not too large and variance which is not too high. Many estimators are known for the estimation of parameters in the case of non-response. Here some of them are applied.

Reweighting estimator. Let $s^{(r)}$ be a subsample of respondents $s^{(r)} \subset s$. The probability to get data from the population element can be expressed as

$$\pi_k^{(r)} = P(k \in s^{(r)}) = P(k \in s^{(r)}|s)P(k \in s) = \kappa_k \pi_k,$$

where κ_k is the response probability of the element k, $k \in \mathcal{U}$ [4]. The response probability will be considered as known in our study; therefore, the Horvitz–Thompson estimator will be used to estimate the total from the sample of respondents $s^{(r)}$.

Imputation by logistic regression. The probability of a population element to obtain a value 1 for a binary study variable y will be simulated by the logistic regression model [2]:

$$P(y = 1) = \frac{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'\boldsymbol{\beta}\}},$$

where $\mathbf{x} = (1, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(J)})'$ is a matrix of auxiliary variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$ is a vector of coefficients. Model coefficients are estimated from the observations available using the maximum likelihood method, and estimates $\widehat{P}(y_k = 1)$ are obtained. After that, the values y_k are simulated as values of Bernoulli random variables with probabilities of success $\widehat{P}(y_k = 1)$ and are denoted by $\hat{y}_k$.

Multiple imputation. The data of the sampled elements which would be available in the case of full response are called real data. If some elements are not responding, then their real data are not known. If the statistician makes certain assumptions about the non-response distribution and imputes the values of the variable for missed observations, all values of the sampled elements become available for that variable, but some of them are not real. The imputation of missing values means the input of additional uncertainty into the data set, in comparison with the real data of the sampled elements. The variance of the estimator for a population parameter based on the data with some imputed values cannot be considered as variance of the estimator for a population parameter obtained for real data because the variability of the imputed values should also be taken into account in the variance of the estimator.

Using the method of logistic regression, all values of the study variable for sampled elements are obtained, and a parameter of study $\theta = t_y = \sum_{k=1}^{N} y_k$ is estimated. Let us denote the estimator by $\hat{\theta} = \hat{t}_y$. The method used for imputation is random. It is repeated a C number of times, C complete data sets are obtained, and C estimators $\hat{\theta}_1, ..., \hat{\theta}_C$ become available for θ. The estimator

$$\overline{\hat{\theta}}_C = \frac{1}{C} \sum_{c=1}^{C} \hat{\theta}_c$$

is obtained by *multiple imputation*, and its variance is estimated by

$$\widehat{Var}(\bar{\hat{\theta}}_C) = \overline{W}_C + \frac{C+1}{C}\widehat{B}_C \tag{6}$$

with the component of variance within the complete samples

$$\overline{W}_C = \frac{1}{C}\sum_{c=1}^{C}\widehat{Var}(\hat{\theta}_c)$$

and the component of variance between the estimates for complete data sets

$$\widehat{B}_C = \frac{1}{C-1}\sum_{c=1}^{C}(\hat{\theta}_c - \bar{\hat{\theta}}_C)^2.$$

The term $\hat{B}_C$ estimates the increase in variance $Var(\hat{\theta})$ due to imputation [3].

3. Problem formulation

Let us suppose a balanced sample is available for a social survey. Auxiliary information is used at the stage of sample selection. Unfortunately, non-response occurs, and the set of respondents becomes unbalanced. If the data of the respondents are only used to estimate the parameter of the finite population, considering the set of the respondents to be selected for the sample, then a bias of the estimator for the population parameter may arise, and the variance of this estimator is increasing due to the lower size of the set of the respondents, in comparison with the selected sample size. The sample should be adjusted for non-response. If imputation for missing values of a study variable is used, then the sample balance is still preserved. If the reweighting of the respondent set is used, then sample balance is destroyed, and the respondent set is no longer balanced.

The method using auxiliary information at the estimation stage is calibration of the design weights. As it is mentioned in [7], the combination of balancing and calibration is a good strategy. Our aim is to study this strategy introducing non-response by simulation.

Sample balancing and calibration with the same auxiliary variables means the usage of auxiliary information twice. Our aim is to answer if it is worth doing.

4. Statistical simulation

4.1. Study population

Labour Force Survey data of Statistics Lithuania [5] are used for a simulation study. A fictitious population consists of $M = 21\,318$ individuals aged 16–69. 19 586 of them are employed and 1 732 are unemployed (inactive individuals are not included in the population). The parameter of interest is the number of the unemployed in the population, and it will be estimated in the study. The study variable y is binary with the value 1, if a person is unemployed, and 0 otherwise.

The population consists of $N = 11\,236$ households, with the average size of 1.9 persons. These households are considered as clusters in our study. The cluster size equals to the number of its members.

The same auxiliary variables will be used at the sampling design and estimation stage. The variables having influence on the unemployment of a person are selected as auxiliary. From the data analysis of the previous surveys, it is known that they are sex, age and urban/rural living area. Age is categorized into groups [16;22], [23;29], [30;39], [40;49], [50;59], [60;69]. Indicator variables are constructed for each of the groups mentioned above. Their population values and totals are considered to be known before sampling.

Simulation is carried out with the sample size $n = 100,\ 1000,\ 5000$ clusters in order to perceive dependency of the accuracy of the results on the sample size. Each strategy is repeated $K = 10$ times and simulation results are averaged. The number of repetitions K is small and diminishes the validity of the conclusions; however, computer resources available do not allow using more repetitions.

4.2. Simulation strategies

The strategy is a pair consisting of sampling design and estimator. The following seven strategies are studied:

1. Balanced cluster sampling and Horvitz–Thompson estimator. A cube method and auxiliary information is used at the sampling stage. Inclusion probabilities are considered to be proportional to the household size: $\pi_k = nm_k/M$, $k = 1,2,...,N$, m_k – household size, $m_1 + ... + m_N = M$, n – cluster sample size. We denote this strategy by BC+HT.

2. Simple random cluster sampling and calibrated estimator. The same auxiliary information as for the first strategy is used here at the estimation stage only (SRCS+CAL).

3. Balanced cluster sampling, non-response, and calibrated estimator of the total. Reweighting is used for non-response adjustment (BC+NR+Rew+CAL).

4. Balanced cluster sampling and calibrated estimator (BC+CAL).

5. Simple random cluster sampling, non-response and calibrated estimator (SRCS+NR+Rew+CAL).

6. Balanced cluster sampling, non-response and Horvitz–Thompson estimator (BC+NR+Rew+HT).

7. Balanced cluster sampling, non-response, logistic regression model for imputation of missing values for a study variable, and Horvitz–Thompson estimator (BC+NR+Imp+HT).

As auxiliary information, the same indicator vectors for sex, age and urban/rural living area are used for balanced sampling, calibration and logistic regression model.

Non-response has to be simulated. It is assumed that all household members are responding to the survey questionnaire or not, and response probabilities of the household (and its members) are assumed to be equal: $\kappa_k = 0.9$. The inclusion probability for a responding individual [4] is

$$\pi_k^{(r)} = P(k \in s^{(r)}) = \kappa_k \pi_k = 0.9\,\pi_k. \tag{7}$$

Here by $s^{(r)}$ is denoted a subsample of respondents. The inclusion probability $\pi_k^{(r)}$ is used in the third, fifth and sixth strategy. Because of equal and known response probabilities κ_k, the Horvitz–Thompson estimator becomes a reweighting estimator.

For the seventh strategy, the probability of a household member to be unemployed is simulated using the logistic regression model. Firstly, the logarithm of the odds ratio is estimated by the maximum likelihood method with the use of the function *glm* of the software R package *stats*:

$$\begin{aligned}\ln \frac{\hat{P}(y=1)}{1-\hat{P}(y=1)} &= -4{,}0579 + 0{,}1625 * x_{male} - 0{,}4259 * x_{urban} + 1{,}001 * x_{agegr.1} + 2{,}3502 * x_{agegr.2} \\ &\quad + 2{,}1715 * x_{agegr.3} + 2{,}0546 * x_{agegr.4} + 2{,}0058 * x_{agegr.5}.\end{aligned}$$

The values to be imputed instead of the missing values of the binary study variable y are simulated according to the Bernoulli distribution with the probability of success $\hat{P}(y_k = 1)$, and simulation results are considered as the estimates $\widehat{y}_k$. Consequently, the total of the study variable y is estimated as follows:

$$\hat{t}_y = \sum_{k \in s^{(r)}} d_k y_k + \sum_{k \in s \setminus s^{(r)}} d_k \hat{y}_k,$$

here $d_k = 1/\pi_k$ are design weights, $s \setminus s^{(r)}$ is a subsample of non-respondents, $\hat{y}_k$ are the values of the study variable y for individuals, simulated by Bernoulli distribution using the logistic regression model.

The R package *sampling* [8] is used to estimate parameters and their variances. In the case of a balanced sampling design, the estimates of variance for the estimator are computed by the function *varest* using the Deville's method for which only first-order inclusion probabilities are needed. In the case of simple random cluster sampling for variance estimation of the estimator of the total, the

function *calibev* is used for an unbiased estimator and for a calibrated estimator. To use it, second-order inclusion probabilities for individuals $\pi_{kl} = P(k \in s, l \in s)$ have to be indicated. For simple random cluster sampling, let us suppose two clusters and their elements are available: $u_k = \{e_{ki}, i = 1,...,m_k\}$ and $u_l = \{e_{li}, i = 1,...,m_l\}$. For two elements, we have

$$\pi_{ki,lj} = P(e_{ki} \in s, e_{lj} \in s) = P(e_{ki} \in s | e_{lj} \in s) P(e_{lj} \in s),$$

$$P(e_{ki} \in s | e_{ij} \in s) = \begin{cases} P(e_{ki} \in s | e_{kj} \in s) = 1, & \text{for } k = l,\ k = 1,...,M, \\ P(e_{ki} \in s | e_{lj} \in s) = \frac{n-1}{N-1}, & \text{for } k \neq l,\ k,l = 1,...,M, \end{cases}$$

$$P(e_{lj} \in s) = \frac{n}{N}.$$

From here, joint inclusion probabilities for two elements are

$$\pi_{ki,lj} = \begin{cases} \frac{n}{N}, & \text{for } k = l,\ k = 1,...,M \\ \frac{n-1}{N-1}\frac{n}{N}, & \text{for } k \neq l,\ k,l = 1,...,M, \end{cases}$$

with i and j being the elements of the clusters, $i = 1,...,m_k,\ j = 1,2,...,m_l$.

5. Main results

For any strategy a sample was drawn $K = 10$ times, and the parameter $\theta = t_y$ of a study variable y was estimated by $\hat{\theta}_k$, $k = 1,2,...,K$. The accuracy measures for estimates are used as follows:

empirical mean or average of the estimates

$$\bar{\hat{\theta}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k,$$

empirical bias

$$\widehat{Bias}(\hat{\theta}) = \bar{\hat{\theta}} - \theta,$$

relative empirical bias

$$\widehat{RBias}(\hat{\theta}) = \widehat{Bias}(\hat{\theta})/\theta,$$

average of the variance estimates

$$\overline{\widehat{Var}}(\hat{\theta}_k) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Var}(\hat{\theta}),$$

coefficient of variation

$$\widehat{cv}(\hat{\theta}) = \frac{\sqrt{\overline{\widehat{Var}}(\hat{\theta})}}{\bar{\hat{\theta}}},$$

relative mean squared error

$$\widehat{RMSE}(\hat{\theta}) = \frac{\sqrt{\widehat{Bias}^2(\hat{\theta}) + \overline{\widehat{Var}}(\hat{\theta})}}{\bar{\hat{\theta}}}.$$

The relative biases of the estimates of the total for balancing variables are presented in Table 1. It shows a possibility to achieve complete balance of auxiliary variables for a small, medium and large sample size. It is seen in Table 1 that for a small sample size ($n = 100$), the estimates of totals for balancing variables are far from the real values due to the rounding problem arising essentially. Large samples ($N = 5000$) do not encounter such a problem. It means that for a small sample size, it is difficult to achieve balance of auxiliary variables. Therefore, if the study variable is correlated with

auxiliary variables, the estimates of its total for small, not well-balanced sample sizes should not be very precise. Table 1 also show, that balance for indicator variables characterizing smaller groups, for example, age groups, is worse than balance for indicators characterizing large groups: sex and living area. In other words, it can be said that the relative empirical bias of the estimates for the total of auxiliary variable is higher for an indicator characterizing a small group than for an indicator characterizing a large group (value of t_{xji}) in the balanced sample. Means for ten estimates of the totals for auxiliary variables and their empirical biases are presented in Table 1.

Table 1. Empirical biases for the estimates of totals for balancing variables in balanced sampling design, sample size $n = 100, 1\,000, 5\,000$

j	i	x_{ji}	$t_{x_{ji}}$	$\widehat{\overline{RBias}}(\hat{t}_{x_{ji}})$		
				$n = 100$	$n = 1000$	$n = 5000$
Sex 1	1	Male	9 674	-0.0010	0.0118	0
	2	Female	11 644	0.0009	-0.0098	0
Age 2	1	1 age g.	3 096	0.1670	0.0959	0.0048
	2	2 age g.	1 700	-0.2582	-0.0282	-0.0118
	3	3 age g.	2 899	-0.1397	-0.1787	-0.0028
	4	4 age g.	4 978	-0.1006	0.1171	-0.0028
	5	5 age g.	5 107	0.1508	-0.1999	0.0127
	6	6 age g.	3 358	0.0164	0.1837	-0.0096
Area 3	1	Urban	14 635	-0.0095	0.0051	-0.0001
	2	Rural	6 683	0.0208	-0.0111	0.0011

Multiple imputation is used to adjust a sample for non-response in Strategy 7. Simulation results presented in Table 2 demonstrate an increase in the estimate for variance of the estimator of the total due to imputation using logistic regression and Bernoulli distribution. They show that variance due to imputation increases by about 6–15%. Results of estimation of the population the total t_y for seven strategies are presented in tables 3, 4 and 5.

Table 2. Results of multiple imputation, $C = 10$

n	$\bar{\hat{t}}_y$	$\widehat{Var}(\bar{\hat{t}}_y)$	$\overline{W}_{10}$	$\hat{B}_{10}$	$\hat{B}_{10} \cdot 100/\widehat{Var}(\hat{t}_y)(\%)$
100	2 258	241 469	226 576	13 539	5.6
1 000	1 758	19 779	16 411	3 062	15.5
5 000	1 725	2 365	2 118	225	9.5

Table 3. Estimates of accuracy measures for estimators of the total of a study variable for seven strategies, $n = 100$

Strategy	$\bar{\hat{t}}_y$	$\widehat{\overline{Var}}(\hat{t}_y)$	$\widehat{cv}(\hat{t}_y)$	$\widehat{Bias}(\hat{t}_y)$	$\widehat{RMSE}(\hat{t}_y)$
1. BC+HT	1 595	169 191	0.259	-137	0.272
2. SRCS+CAL	1 730	137 509	0.211	-2	0.214
3. BC+NR+Rew+CAL	1 740	178 696	0.242	8	0.243
4. BC+CAL	1 729	208 781	0.265	-3	0.264
5. SRCS+NR+Rew+CAL	1 692	208 339	0.265	-40	0.271
6. BC+NR+Rew+HT	1 992	228 248	0.244	260	0.273
7. BC+NR+Imp+HT	1 828	201 827	0.245	96	0.251

Table 4. Estimates of accuracy measures for estimators of the total of a study variable for seven strategies, $n = 1\,000$

Strategy	$\bar{\hat{t}}_y$	$\widehat{\overline{Var}}(\hat{t}_y)$	$\widehat{cv}(\hat{t}_y)$	$\widehat{Bias}(\hat{t}_y)$	$\widehat{RMSE}(\hat{t}_y)$
1. BC+HT	1 716	17 168	0.076	-16	0.077
2. SRCS+CAL	1 768	12 863	0.064	36	0.067
3. BC+NR+Rew+CAL	1 672	17 166	0.079	-60	0.086
4. BC+CAL	1 621	17 977	0.083	-111	0.107
5. SRCS+NR+Rew+CAL	1 768	18 508	0.077	36	0.080
6. BC+NR+Rew+HT	1 899	21 374	0.077	167	0.117
7. BC+NR+Imp+HT	1 800	18 013	0.074	68	0.084

Table 5. Estimates of accuracy measures for estimators of the total of a study variable for seven strategies, $n = 5\,000$

Strategy	$\bar{\hat{t}}_y$	$\widehat{\overline{Var}}(\hat{t}_y)$	$\widehat{cv}(\hat{t}_y)$	$\widehat{Bias}(\hat{t}_y)$	$\widehat{RMSE}(\hat{t}_y)$
1. BC+HT	1 725	2 186	0.027	-7	0.027
2. SRCS+CAL	1 744	1 530	0.022	12	0.023
3. BC+NR+Rew+CAL	1 727	2 157	0.027	-5	0.027
4. BC+CAL	1 727	2 417	0.028	-5	0.029
5. SRCS+NR+Rew+CAL	1 714	2 408	0.029	-18	0.030
6. BC+NR+Rew+HT	1 924	2 914	0.028	192	0.104
7. BC+NR+Imp+HT	1 750	2 210	0.026	18	0.028

When comparing the results of tables 3–5, one should have in mind that balanced sampling is applied with probabilities proportional to the cluster size, but the cluster size is not taken into account in simple random cluster sampling. These differences may slightly influence the accuracy of the estimates.

6. Discussion

Strategy 1 – balanced sampling and the Horvitz–Thompson estimator of the total – shows that the estimator has empirical bias, which decreases with an increasing sample size. Beside the common regularity property, the balance of samples for small sample sizes is not adequate, and it influences empirical biases of the estimates.

Strategy 6 is obtained, appending non-response to Strategy 1. Empirical bias is observed. It decreases with an increasing sample size, but still remains significant. The variance of the estimator due to non-response also increased, and it influences *RMSE* for large samples.

Strategy 4 means balanced sampling, as for Strategy 1, but the Horvitz–Thompson estimator is replaced by a calibrated estimator. Empirical bias is approximately the same as for Strategy 1, but variance increases and relative measures of accuracy are also higher than for Strategy 1.

Strategy 3 consists of the conditions for Strategy 4 appended with non-response. The estimates became closer to the estimates for Strategy 1. Relative measures of accuracy decreased for small sample sizes, but remain unchanged for large sample sizes.

Strategy 7. Balanced sampling and non-response. A logistic regression model is used for the imputation of the study variable values for non-responding elements, and the Horvitz–Thompson estimator of the total is used. It is reasonable to compare this estimator with Strategy 6 because of the same sampling design, the same estimator, but different adjustment for non-response. Biases are smaller for Strategy 7 than for Strategy 6, and they are decreasing with an increasing sample size.

In comparison with Strategy 3, the variance estimates for Strategy 7 are higher. With increasing sample sizes, the estimates of Strategy 7 approached the estimates obtained for the Strategy 3, remaining a little bit higher, and the variances are higher.

Strategy 2. This is a strategy giving the best accuracy for the estimator of the total. The variance estimates are lower than for Strategy 1 for any sample size. In the case of small sample sizes, there is

no bias for Strategy 2, which is significant for Strategy 1. In the simulation carried out, calibration does not improve the accuracy in the case of balanced sampling without non-response (Strategy 4). When there is no non-response, the classical Strategy 2 is the best.

Strategy 5. Simple random sampling, non-response and calibration. We compare the estimates with the results of Strategy 3. Unfortunately, in the case of Strategy 5, the empirical biases are more significant, and variance estimates are higher.

The calibration estimator in Strategy 4, in comparison with the Horvitz–Thompson estimator in Strategy 1 for balanced sampling design, does not improve accuracy; all accuracy measures are higher for the former. But if non-response occurs for balanced sampling design (Strategy 3) the calibrated estimator shows more accurate results, in comparison with Strategy 4 without non-response. It should be mentioned that the calibrated weights are random, but their variability is not taken into account in the estimator of variance for the calibrated estimator.

There is no monotonicity in the change of bias due to an increasing sample size. It may occur because of a small number of repetitions $K = 10$.

Conclusion. Simulation results for Labour Force Survey data show that
if there is no non-response, a simple random sample of clusters and a calibrated estimator of the total (strategy 2) gives the highest accuracy;

if non-response occurs, then balanced sampling, adjustment for non-response by reweighting and calibration (Strategy 3) gives the highest accuracy.

Acknowledgement. The authors are thankful to the two reviewers for their careful reading of the paper and making comments. Taking them into account, the paper was improved essentially.

References

[1] Deville J. C., Särndal C. E., Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*. 1992, 87(418): 376–382.

[2] Lohr S. *Sampling: Design and Estimation.* Pacific Grove: Duxbury Press, 1999.

[3] Rubin N. B., *Multiple Imputation for Nonresponse in Surveys*. 1987, New York: John Wiley and Sons.

[4] Särndal C. E., Swenson B., Wretman J. H., *Model Assisted Survey Sampling*. 1992, New York: Springer-Verlag.

[5] Statisatics Lithuania. *Public data files. Labour Force Survey (carried out quarterly* [interactive]. Vilnius, online: http://osp.stat.gov.lt/en/viesos-duomenu-rinkmenos/. [Accessed: August 10, 2016].

[6] Tillé Y., *Sampling Algorithms*. 2006, New York: Springer Science&Business Media.

[7] Tillé Y., 10 Years of Balanced Sampling with the Cube Method: An Appraisal, *Survey Methodology*. 2011, 12(37): 215–226.

[8] Tillé Y., Matei, A. *Sampling: Survey Sampling. R package version* 2.7. 2015. Online: https://cran.r-project.org/web/packages/sampling/.

SUBALANSUOTO ĖMIMO TAIKYMAS, NEATSAKYMAI Į APKLAUSĄ IR KALIBRUOTASIS ĮVERTINYS

Ieva Dirdaitė, Danutė Krapavickaitė

Santrauka Šio darbo tikslas yra modeliavimo būdu ištirti subalansuoto ėmimo, neatsakymų į apklausą ir kalibruotojo įvertinio tarpusavio sąveiką. Lyginami septynių strategijų, apimančių subalansuotą ėmimą (naudojant kubo metodą), paprastąjį atsitiktinį lizdinį ėmimą, atsižvelgimą į neatsakymus, Horvico ir Tompsono įvertinio bei kalibruotojo įvertinio derinius, rezultatai. Visais atvejais bent viename iš etapų (imties ėmimo arba parametrų vertinimo) yra naudojama papildoma informacija. Ji išreikšta indikatoriais, apibūdinančiais asmens lytį, amžių, gyvenamąją vietą, ir šių indikatorių sumomis. Modeliuoti naudojami realūs Lietuvos statistikos departamento 2011 m. gyventojų užimtumo statistinio tyrimo duomenys. Įvertinio poslinkis, dispersija, santykinė vidutinė kvadratinė paklaida yra tikslumo matai, taikomi įverčiams palyginti.

Reikšminiai žodžiai: papildoma informacija, atsižvelgimas į neatsakymus, kubo metodas, modeliavimas.