STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 7–30

THE COST EFFICIENCY OF SAMPLING DESIGNS

MĀRTIŅŠ LIBERTS[1]

ABSTRACT

The aim of a sample survey is to obtain high quality estimates of population parameters with low cost. The expected precision of estimates and the expected data collection cost are usually unknown making the choice of sampling design a complicated task. Analytical methods can not be used often because of the complexity of the sampling design or data collection process. The aim of this paper is to develop a mathematical framework to compare chosen sampling designs with respect to the expected precision of estimates and the data collection cost. As a result a framework is developed which employs artificial population data generation, survey sampling techniques, survey cost modelling, Monte Carlo simulation experiments and other techniques. The framework is applied to analyse the cost efficiency of the sampling design currently used for the Latvian Labour Force Survey.

Key words: cost efficiency, simulation study, survey cost estimation, survey methodology, variance of estimators.

1. Introduction

The inspiration for this paper comes from pure practical necessity. National Statistical Institutes (NSIs) are the main providers of official statistics in most countries. A large proportion of official statistics produced by NSIs are done so using data collected via sample surveys, with the main customer of official statistics being the general public (or tax payers, in other words). These days, cost efficiency is an essential consideration in all government spending; the question is, *are NSI sample surveys cost efficient?*

There is not a simple answer to the question posed. A sample survey can possess one of many different sampling designs. The simplest sampling designs do not necessarily provide the lowest data collection cost. More complex sampling designs are considered in theory and applied in practice to obtain statistical information with an acceptable precision at a lower cost. In designing a sample survey, the following considerations should be decided upon: *What is the expected precision of the estimates of population parameters? What is the expected data collection cost? Which sampling design should be chosen in order to minimise sampling errors under a*

[1] University of Latvia, Raiņa bulvāris 19, Rīga, LV-1586, Latvia, *martins.liberts@gmail.com.*

fixed data collection cost? These are commonly asked questions during the planning stage of a sample survey. In most cases, the answers to the questions posed cannot be gained through analytical means and NSIs are usually reliant on expert judgement to some extent.

The relation between the precision of estimates and survey cost has been discussed in literature for at least 70 years, though the topic has not been comprehensively addressed. Different aspects of the relationship have been analysed and different goals of analysis have been set by authors but it is possible to observe the lack of common foundations for the topic. One of the first papers devoted to the topic are by Mahalanobis (1940) and Jessen (1942). The topic is extensively discussed by Hansen, Hurwitz, and Madow (1953) and Kish (1965). Significant book regarding the topic is by Groves (1989). The author advocates simulation studies to be the best-suited for a sample design analysis because of usual complexity of cost and precision functions.

Several events have been organised recently, in the United States of America, devoted to the topics of survey cost estimation and simulation models for survey fieldwork operations. For example "Survey Cost Workshop" (2006, Washington, D.C.) and "Workshop on Microsimulation Models for Surveys" (2011, Washington, D.C.). The research of survey field operations is a brand new topic in the scope of statistical research. Several research activities have been devoted to the topic only recently (Chen, 2008; Cox, 2012).

The Latvian Labour Force Survey (LFS) is the main object of the study in the paper. It was organised for the first time in November 1995 (Lapiņš, 1997) and ran biannually. The first redesign of the LFS sampling design was done after the 2000 Latvian Population Census with the new sampling design launched in 2002 (Lapiņš, Vaskis, Priede, & Bāliņa, 2002). It become a continuous survey after the redesign. The second redesign of the survey occurred in 2006. The re-launch of the LFS with the new sampling design and a much larger sample size took place in 2007. Finally, the latest redesign of the LFS sampling design was done by the author in 2009 (Liberts, 2010). The main reason for redesigning the LFS sampling design for the third time was the necessity to update the population frame used for the first-stage sampling units. The redesign resulted in a new sample drawn which was used to run the LFS since 2010. More information regarding the history of the LFS is given by Central Statistical Bureau of Latvia (2012) and European Commission (2012a, 2012b).

The target population and the parameters of interest in the case of the LFS are described in the second section. Artificial population data reflecting the target population of the LFS are necessary to do simulation experiments. A methodology to develop artificial population data is presented in the third section. Artificial population data with characteristics similar to the target population of the LFS has been produced with this methodology. The fourth section of the paper is devoted to the development and the application of the framework for the cost efficiency analysis of sampling designs.

2. Target population and parameters of interest

The target population of the LFS is defined as all residents permanently living in private households. Residents at working-age (15–74 years) compose the main domain of interest. The target population is continuously changing over time, for example some individuals are losing or gaining employment every day. The target population is observed on a weekly basis by the methodology of the LFS (European Commission, 2012b, p. 5).

An individual is called **unit** and denoted by v_i (there are cases when households are used as units). The set of all units is denoted by V. The size of V is M. The units are labelled with an index i where $i \in \overline{1, M}$, $V = \{v_1, v_2, \ldots, v_M\}$. The observation of unit v_i in week w is called **element** and denoted by $u_{i,w}$. The set of all elements in week w is denoted by U_w. There are M elements in U_w. The elements of U_w are labelled with a double index (i, w) where i refers to a unit and w refers to a week, $U_w = \{u_{1,w}, u_{2,w}, \ldots, u_{M,w}\}$. Values $y_{i,w}$ are associated to elements $u_{i,w}$ from U_w. The total of a variable y in week w is defined as

$$Y_w = \sum_{i=1}^{M} y_{i,w}.$$

The total number of weeks observed is denoted by W and w is the week index, $w \in \overline{1, W}$. The set of elements over W weeks is denoted by U, $U = \cup_{w=1}^{W} U_w$. Each U_w consists of the observation of units from V observed in different weeks. The size of U_w is constant over time, $|U_w| = M$ for all w. The size of U is denoted by N, $|U| = \sum_{w=1}^{W} M = MW = N$. An index k is used to label elements over W weeks, $k \in \overline{1, N}$. The elements of each U_w are ordered according to the order of the units of V. The indices

$$\{k : ((k-1) \bmod M) + 1 = i\}$$

correspond to the unit v_i. The example of the set U is given in Table 1. The M rows of the table represent units. The W columns of the table represent weeks observed. The cells of the table represent elements. The dimension of the table is $M \times W$.

The total of the variable y over W weeks is defined as

$$Y = \sum_{w=1}^{W} Y_w = \sum_{w=1}^{W} \sum_{i=1}^{M} y_{i,w} = \sum_{k=1}^{N} y_k.$$

Two types of parameter are considered in the further analysis – the average of weekly totals and the quarterly ratio of two totals. The average of weekly totals is defined

Table 1. Example of set U

i	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$	$\cdots$	$w=W$
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$\cdots$	$y_{1,W}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$\cdots$	$y_{2,W}$
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$\cdots$	$y_{3,W}$
				$\cdots$			
M	$y_{M,1}$	$y_{M,2}$	$y_{M,3}$	$y_{M,4}$	$y_{M,5}$	$\cdots$	$y_{M,W}$

by

$$Y_q = \frac{1}{13}\sum_{w=1}^{13} Y_w = \frac{1}{13}\sum_{w=1}^{13}\sum_{i=1}^{M} y_{i,w} = \frac{1}{13}\sum_{k=1}^{N} y_k = \frac{1}{13}Y,$$

and the quarterly ratio of two totals is defined by

$$R_q = \frac{Y_q}{Z_q} = \frac{\sum_{w=1}^{13} Y_w}{\sum_{w=1}^{13} Z_w} = \frac{\sum_{w=1}^{13}\sum_{i=1}^{M} y_{i,w}}{\sum_{w=1}^{13}\sum_{i=1}^{M} z_{i,w}} = \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{N} z_k}.$$

The estimators of Y_q and R_q are constructed using the π estimator (Särndal, Swensson, & Wretman, 1992, p.42, 176) as

$$\hat{Y}_q = \frac{1}{13}\sum_{(i,w)\in s}\frac{y_{i,w}}{\pi_{i,w}} = \frac{1}{13}\sum_{k\in s}\frac{y_k}{\pi_k}, \tag{1}$$

$$\hat{R}_q = \frac{\sum_{(i,w)\in s}\frac{y_{i,w}}{\pi_{i,w}}}{\sum_{(i,w)\in s}\frac{z_{i,w}}{\pi_{i,w}}} = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{z_k}{\pi_k}} \tag{2}$$

where s is a probability sample of elements and $\pi_{i,w}$ is an inclusion probability of element $u_{i,w}$ in a sample.

3. Artificial population data

Artificial population data are necessary to carry out simulation experiments. Artificial population data are created from the data of the Statistical Household Register (a statistical register owned and maintained by the Central Statistical Bureau of Latvia) and the survey data of the LFS. The artificial population data are represented by two files – one for a static population (the population of units) and other for a dynamic population (the population of elements). There are several assumptions incorporated in the artificial population model:

- the set of units V is fixed over W weeks,
- background variables such as age and place of residence are fixed during W weeks, while study variables (for example, employment status) can change from week to week,

- the membership of individuals to households is fixed over W weeks.

3.1. Static population data

Two data sources are used to construct the static population data. The list of individuals aged 15–74 on 30th January 2011 is extracted from the Statistical Household Register. There are 1 705 048 records (individuals) in the list. The list of individuals forms the frame for the static population. Demographic information (age and gender) and residence information (region, dwelling ID and geographical coordinates) is attached to the list. Dwelling ID allows individuals to be grouped by households (assume a single household per dwelling).

The LFS data are used to create study variables for the static population. The LFS data from 2007–2010 are used. The variables describing demographic information (age and gender), residence information (region and dwelling ID) and economic activity status are extracted from the survey data.

The data from both sources are merged using an imputation technique where recipients are the units in the register data and donors are the units in the survey data. Random donor imputation within classes is used (United Nations, 2010, p.162). However, this is not the classical application of random donor imputation because non-response is not the cause of data missingness here. The cause of data missingness is the fact that the register data do not contain the variable describing economic activity. Imputation classes are built in both data sets according to the same specification using demographic and residence information as auxiliary information.

The imputation is done at seven levels where imputation units are households at the first five levels and imputation units are individuals at the last two levels. Different specification of classes is used at each level. Donors and recipients are grouped in very detailed classes at the first level. As it is not possible to impute all households at the first level (there are not enough donors in each class at the first level), the imputation process is repeated for the not-imputed households at the succeeding levels by merging the imputation classes. There are 26 variables used to define household classes at the first level, 16 at the second level, 12 at the third level, 11 at the forth level and 10 at the fifth level (see Table 2 for more details, where strata is a variable with four values: “Riga”, “Cities”, “Towns”, and “Rural areas”; region is a variable with six values). Strata, region, gender and age are variables used to create imputation classes at the sixth level, and strata, region, gender and age group (12 age groups) are variables used to create imputation classes at the seventh level when imputation units are individuals.

The description of imputation procedure done at each level is given here. The imputation is done in each class c independently. A donor $d_k \in D_c$ is assigned to a recipient $r_i \in R_c$ with a probability $\frac{1}{|D_c|}$ if $|D_c| \geq 10$ where D_c is the set of donors in a class c, R_c is the set of recipients in a class c, and $|D_c|$ is the total number of donors in a class c. A donor $d_k \in D_c$ can be assigned to several recipients from R_c.

Table 2. Household imputation classes at the first five levels

Variable	Level 1	Level 2	Level 3	Level 4	Level 5
Males 15–19	1	1	1	1	1
Males 20–24	2	2	2	2	2
Males 25–29	3	3	3	3	3
Males 30–34	4	3	3	3	3
Males 35–39	5	4	3	3	3
Males 40–44	6	4	3	3	3
Males 45–49	7	5	4	4	4
Males 50–54	8	5	4	4	4
Males 55–59	9	6	4	4	4
Males 60–64	10	6	4	4	4
Males 65–69	11	7	5	5	5
Males 70–74	12	7	5	5	5
Females 15–19	13	8	6	6	6
Females 20–24	14	9	7	7	7
Females 25–29	15	10	8	8	8
Females 30–34	16	10	8	8	8
Females 35–39	17	11	8	8	8
Females 40–44	18	11	8	8	8
Females 45–49	19	12	9	9	9
Females 50–54	20	12	9	9	9
Females 55–59	21	13	9	9	9
Females 60–64	22	13	9	9	9
Females 65–69	23	14	10	10	10
Females 70–74	24	14	10	10	10
Strata	25	15	11	11	.
Region	26	16	12	.	.

The imputation is not done in a class c if $0 \leq |D_c| < 10$. The units imputed at one level are not re-imputed any more at the succeeding imputation levels. The units not imputed at one level will be imputed at one of succeeding imputation levels.

The imputation of households as units at the first five levels allows one to keep demographic and economic composition of households the same as observed in the survey data. The specification of the classes at the first five levels is hierarchical. The classification of the classes is the most detailed at the first level. The classes are merged by each succeeding level. Economic activity status is imputed for 82.2% of all individuals from the register data at the first five levels. The imputation for all individuals can not be done in this manner because there are classes of households in the register data which have not been observed in the survey data or have been observed only in few cases (less than 10).

Economic activity status is imputed for the rest of individuals at the last two levels with the same imputation technique except imputation units are individuals and other specification of classes is used. The classification of the classes here is based on the same auxiliary information as used at the first five levels, though it is used at the individual level rather than at the household level. The specification of the classes is hierarchical here as well. It is possible to impute economic activity status for all remaining individuals at the last two imputation levels.

3.2. Dynamic population data

A dynamic population according to the description in Section 2 is generated. A variable – economic activity status is extrapolated from the static population to the dynamic population. Let y_i be the economic activity status of an individual v_i from the static population. A Markov chain model is used to generate the dynamic population. The economic activity status y_i can take any of three different values, $y_i \in \{1, 2, 3\}$:

- $y_i = 1$ if an individual v_i is employed,
- $y_i = 2$ if an individual v_i is unemployed,
- $y_i = 3$ if an individual v_i is economically inactive.

The value of y_i is defined once in a week by the LFS methodology. Let $y_{i,w}$ be the economic activity status for an individual v_i on week $w \in \{0, 1, 2, \ldots\}$. Let $y_{i,w}$ be random variables and sequence $y_{i,0}, y_{i,1}, y_{i,2}, \ldots$ be a time-inhomogeneous Markov chain for an individual v_i. The state space of the Markov chain is $\{1, 2, 3\}$. The probability of going from a state k to a state l after a week for an individual v_i is

$$p_{i,w,w+1,k,l} = P\left(y_{i,w+1} = l \mid y_{i,w} = k\right).$$

Constant transition probabilities for all v_i are assumed

$$p_{i,w,w+1,k,l} = p_{w,w+1,k,l},$$

and a time-dependent transition matrix the same for every individual v_i is

$$\boldsymbol{P}_{w,w+1} = \begin{pmatrix} p_{w,w+1,1,1} & p_{w,w+1,1,2} & p_{w,w+1,1,3} \\ p_{w,w+1,2,1} & p_{w,w+1,2,2} & p_{w,w+1,2,3} \\ p_{w,w+1,3,1} & p_{w,w+1,3,2} & p_{w,w+1,3,3} \end{pmatrix}.$$

The estimate of $\boldsymbol{P}_{w,w+1}$ is necessary to generate artificial dynamic population data. It is assumed there are 52 weeks in each year, and 52 weeks are split in four seasonal quarters by 13 weeks in each.

The first quarter is shown as an example here. It is assumed that all 13 weekly transition matrices are equal for the first quarter. Thus, the following equivalence holds for the 13 weekly transition matrices:

$$\boldsymbol{P}_{0,1} = \boldsymbol{P}_{1,2} = \ldots = \boldsymbol{P}_{12,13}.$$

In general the transition matrix after 13 weeks is equal to the product of 13 weekly transition matrices: $\boldsymbol{P}_{0,13} = \prod_{w=0}^{12} \boldsymbol{P}_{w,w+1}$. Because of the previous equivalence we can write

$$\boldsymbol{P}_{0,13} = \boldsymbol{P}_{w,w+1}^{13} \quad \text{for all} \quad w \in \overline{0,12}.$$

It follows from the previous equation

$$\boldsymbol{P}_{w,w+1} = \sqrt[13]{\boldsymbol{P}_{0,13}} \quad \text{for all} \quad w \in \overline{0,12}.$$

The LFS is a rotating panel survey. There is a 50% overlap between the succeeding quarterly samples. The individuals are interviewed with 13 weeks shift between the succeeding quarterly samples. Theoretically it is possible to estimate $\boldsymbol{P}_{0,13}$ from the LFS data, because there are respondents who are observed both at week $w = 0$ and week $w = 13$. Practically the estimation of $\boldsymbol{P}_{0,13}$ will not be precise if only data from overlapping respondents of weeks $w = 0$ and $w = 13$ are used. It is because the number of such respondents is small.

Thus, the decision was made to estimate $\boldsymbol{P}_{0,13}$ using the LFS data from overlapping respondents of the first and the second quarter:

$$\hat{\boldsymbol{P}}_{0,13} = \hat{\boldsymbol{p}}_{1,2}$$

where $\hat{\boldsymbol{p}}_{1,2}$ is the estimate of transition matrix from the first quarter to the second quarter using the LFS data. This estimation is introducing some bias to the estimate of $\boldsymbol{P}_{0,13}$, but it is more stable estimate.

Thus, the estimate of the weekly transition matrix for the first quarter is estimated as

$$\hat{\boldsymbol{P}}_{w,w+1} = \sqrt[13]{\hat{\boldsymbol{p}}_{1,2}} \quad \text{for all} \quad w \in \overline{0,12}.$$

Similarly, the weekly transition matrices for the second quarter are estimated as

$$\hat{\boldsymbol{P}}_{w,w+1} = \sqrt[13]{\hat{\boldsymbol{p}}_{2,3}} \quad \text{for all} \quad w \in \overline{13,25},$$

where $\hat{\boldsymbol{p}}_{2,3}$ is the estimate of a quarterly transition matrix from the second quarter to the third quarter and so on.

A time-inhomogeneous Markov chain is used to introduce a seasonal component in dynamic population data as it is observed in the survey data with respect to the changes of economic activity status of individuals. The estimates of the quarterly transition matrices and the weekly transition matrices are available in Table 3. The estimated weekly transition matrices are used to generate the dynamic population data by weeks. The variable of economic status from the static population is used as the initial state ($w = 0$) for each individual.

Table 3. Estimates of Transition Matrices

q	w	$\hat{\boldsymbol{p}}_{q,q+1}$	$\hat{\boldsymbol{P}}_{w,w+1}$
1	$\overline{0,12}$	$\begin{pmatrix} 0.950 & 0.021 & 0.029 \\ 0.251 & 0.541 & 0.209 \\ 0.058 & 0.052 & 0.890 \end{pmatrix}$	$\begin{pmatrix} 0.996 & 0.002 & 0.002 \\ 0.025 & 0.952 & 0.022 \\ 0.004 & 0.006 & 0.990 \end{pmatrix}$
2	$\overline{13,25}$	$\begin{pmatrix} 0.944 & 0.021 & 0.035 \\ 0.253 & 0.540 & 0.206 \\ 0.055 & 0.055 & 0.891 \end{pmatrix}$	$\begin{pmatrix} 0.995 & 0.002 & 0.003 \\ 0.026 & 0.952 & 0.022 \\ 0.004 & 0.006 & 0.990 \end{pmatrix}$
3	$\overline{26,38}$	$\begin{pmatrix} 0.937 & 0.028 & 0.035 \\ 0.199 & 0.609 & 0.192 \\ 0.048 & 0.042 & 0.910 \end{pmatrix}$	$\begin{pmatrix} 0.995 & 0.003 & 0.003 \\ 0.019 & 0.962 & 0.019 \\ 0.004 & 0.004 & 0.992 \end{pmatrix}$
4	$\overline{39,51}$	$\begin{pmatrix} 0.930 & 0.033 & 0.037 \\ 0.183 & 0.596 & 0.221 \\ 0.042 & 0.043 & 0.915 \end{pmatrix}$	$\begin{pmatrix} 0.994 & 0.003 & 0.003 \\ 0.018 & 0.960 & 0.022 \\ 0.003 & 0.004 & 0.993 \end{pmatrix}$

4. Cost efficiency

Assume an arbitrary population parameter θ. There is a probability sample s_p drawn by a sampling design $p\left(s\right)$. The parameter θ is estimated by an estimator $\hat{\theta}_p$. The variance of $\hat{\theta}_p$ is denoted by $\mathrm{Var}_p\left(\hat{\theta}_p\right)$. There is a cost function $c\left(s_p\right)$. The operational cost of a sample s_p is computed by the cost function $c_p = c\left(s_p\right)$. The result of the cost function is a random variable because s_p is a random sample. The expectation of c_p under a sampling design $p\left(s\right)$ is denoted as $\mathrm{E}\left(c_p\right) = C_p$. Definition 1 is used to compare two sampling designs with respect to cost efficiency where γ is a survey budget available.

Definition 1. A sampling design $p\left(s\right)$ is more cost efficient than a sampling design $q\left(s\right)$ for estimation of a population parameter θ with a survey budget γ if

$$\mathrm{Var}_p\left(\hat{\theta}_p \middle| C_p \approx \gamma\right) < \mathrm{Var}_q\left(\hat{\theta}_q \middle| C_q \approx \gamma\right).$$

The parameter γ can be replaced by a parameter vector $\boldsymbol{\gamma}$ denoting budget allocation by operational domains in Definition 1. Specifying the budget as a vector is useful in practice if the allocation of a budget by operational domains is important. The practical application of Definition 1 to analyse cost efficiency of sampling designs is achieved by the following steps:

- selection of sampling designs to be analysed with respect to the cost efficiency,
- definition of a cost function $c\left(s\right)$,

- setting the total budget γ or a budget allocation $\boldsymbol{\gamma}$,
- setting specific sample design parameters for each chosen sample design to achieve the expected total cost or cost allocation for all designs approximately equal to γ or $\boldsymbol{\gamma}$ accordingly,
- selection of population parameters for analysis,
- calculation of variance for the estimators of parameters selected,
- determination of the most cost efficient sample design using Definition 1.

4.1. Sampling designs

A modified simple random sampling design (mSRS) is introduced as an alternative to the current LFS sampling design. The notation of Section 2 is used here. The set of sampled units is denoted by $\tilde{s} \subseteq V$. The set of sampled elements in week w is denoted by $s_w \subseteq U_w$. The set of sampled elements over W weeks is denoted by $s = \cup_{w=1}^{W} s_w \subseteq U$. The weekly sample size is denoted by m. The total sample size n is computed as mW. The value of m has to be chosen so that $n = mW \leq M$ because each unit can be sampled only once during W weeks. The goals of the mSRS are:

- all elements of U have sampling probabilities equal to $\pi_k = \frac{n}{N} = \frac{m}{M}$,
- weekly samples for W weeks are drawn,
- all weekly samples are drawn with equal sample size, $|s_w| = m$ for all w, making the total sample size equal to $n = mW$,
- all n sampled elements refer to n different units, one and only one element $u_{i,w}$ may be sampled for a unit v_i.

There are several techniques to achieve the sample by the mSRS. An example is presented here. The sample is selected in two steps. The first step is to select n units by simple random sampling without replacement from M units. The sampled units are sorted in a random order. The ordered sample of units is systemically split into W blocks with length m. The units of the first block determine the sampled elements for the first week, the units of the second block determine the sampled elements for the second week and so on until the units of the last block determine the sampled elements for the week W.

A probability to sample a unit v_i at the first step is equal to $\frac{n}{M}$. The probability of a unit v_i to be located in a block w after the random ordering is equal to $\frac{1}{W}$. A sampled element is determined by the index i of a sampled unit v_i and the index w of a block containing the unit v_i. Therefore, the sampling probability of an element is equal to $\pi_{i,w} = \pi_k = \frac{n}{M}\frac{1}{W} = \frac{n}{N} = \frac{m}{M}$.

A stratified mSRS sampling design is realised if units are stratified in H strata and mSRS is applied independently in each stratum with sample size n_h. The stratified mSRS is denoted as mSSRS.

Three sampling designs are chosen for the cost efficiency study. The first design is mSSRS with individuals as sampling units (denoted as mSSRSi). Each sampled individual is interviewed by a household questionnaire and an individual questionnaire. This is a similar sampling design used for LFS in Sweden and Denmark – stratified random sampling of individuals, and only sampled individuals take part in a survey (European Commission, 2012a).

The second sampling design is mSSRS with households as sampling units (denoted as mSSRSh). Each sampled household is interviewed by a household questionnaire and all household members are interviewed by an individual questionnaire. This is a similar sampling design used for LFS in Malta, Austria and United Kingdom – stratified random sampling of dwellings or households and all members of a sampled dwelling or household take part in a survey (European Commission, 2012a).

The third sampling design is two-stage sampling design (denoted as TSSh) used in practice for the Latvian LFS. The primary sampling units (PSUs) are census counting areas at the first stage. Census counting areas are geographically compact areas with low variation by size (here and afterwards the size of PSU is measured as the number of dwellings in PSU) making them useful for sampling purposes. The average PSU size is 238 in Riga (capital city), 219 in other cities (excluding Riga), 190 in towns and 141 in rural areas.

PSUs are stratified in four strata by the level of urbanisation (Riga – the capital of Latvia, other cities, towns and rural areas). PSUs are sampled by systematic πps sampling with random starting point and sampling probabilities proportional to PSU size. PSUs are ordered in "serpentine" order in each stratum allowing for implicit stratification by administrative territories. The systematic sampling of PSUs allows the implementation of the chosen rotation scheme 2-(2)-2 (European Commission, 2012a, p.7).

Dwellings are the secondary sampling units sampled by simple random sampling with fixed sample size in each stratum. Usually there is only one household in each dwelling. Each sampled dwelling is interviewed by a household questionnaire and all household members are interviewed by an individual questionnaire. More details about the TSSh design are available at Liberts (2010).

The two-stage sampling design using census counting areas as PSUs has been used for the Latvian LFS since 2002. Several questions about the chosen sampling design have been raised quite often: *Why should Central Statistical Bureau of Latvia (CSB) use such complex (two-stage) sampling design? Why CSB are not switching to more simpler (one-stage) sampling design?* One of the main reasons for these questions was the fact that design is using census counting areas as PSUs. The frame of census counting areas (PSUs) has to be updated using the resources of the CSB (it is because the census counting areas are not available in any administrative register). Thus, the question regarding the most appropriate sampling design for

LFS has been open for quite a long time. This explains the choice of the alternative sampling designs for this study.

It is not obvious which of the selected sampling designs is the most cost efficient in the case of the LFS. The mSSRSi and the mSSRSh could provide more precise estimates with smaller sample sizes because of lower cluster effect (in the case of the LFS). However, the TSSh requires lower fieldwork cost per unit because of shorter travelling distances for interviewers allowing to select larger sample size.

Other sampling designs can be analysed as well, for example, mixed designs where one-stage sampling is used for high density areas and two-stage sampling for low density areas (to reduce travelling cost). This kind of sampling design was used for the Latvian LFS in 1995–2001 (Lapiņš et al., 2002, p.628). On the one hand this kind of sampling design could have good cost efficiency properties.

On the other hand, the complexity of the design is higher making the estimators of population parameters and estimators of precision more complex. This could be an obstacle for the external users of survey micro-data or for automatic precision estimation systems assuming unified sampling design used throughout the survey. It will be possible to observe further in the paper that mixed sampling design (with chosen stratification) would not be more cost effective compared to the three chosen sampling designs.

4.2. Cost function

Assume a survey done by face to face personal interviews where interviewers are travelling to respondents. Two components of fieldwork cost are assumed – travel cost and interview cost. Travel cost is approximated by a function $c_1(s) = dK_fC_fk_d$ where d is the total travelling distance done by interviewers expressed in kilometres, K_f is the average fuel consumption expressed in litres per kilometre, C_f is the average price of fuel expressed in lats per litre (lats is the national currency of the Republic of Latvia, 1 lats = 0.702804 euro), and k_d is an adjustment coefficient specified by a statistician.

There are G interviewers available and there is an interviewer assigned to each unit in population. Sampled units for week w are split by interviewers according to the predefined interviewer assignment in population. Geographical coordinates are known for the sampled units and also for the residence places of interviewers. Distances between sampled units and interviewers residence are computed as the Euclidean distance.

The shortest path connecting the residence of an interviewer g and the sampled units assigned to an interviewer g is found by solving a travelling salesperson problem (TSP). The TSP is solved by the nearest insertion algorithm (Rosenkrantz, Stearns, & Lewis, 1977, p.572). The total travel distance d is computed by

$$\sum_{g=1}^{G}\sum_{w=1}^{W} d_{g,w}$$

where W is the total number of weeks observed and $d_{g,w}$ is the length of the path found by solving a TSP for an interviewer g in week w. The constants K_f, C_f and k_d are set.

The interview cost is computed by a function $c_2(s) = aC_a + bC_b$ where a is the total number of individuals in a sample s, b is the total number of households in a sample s, C_a is the interview cost for an individual questionnaire, and C_b is the interview cost for a household questionnaire. A cost function $c(s) = c_1(s) + c_2(s) = K_f C_f k_d \sum_{g=1}^{G} d_g + aC_a + bC_b$ is used further in the study.

4.3. Fieldwork budget allocation

The field work budget $\boldsymbol{\gamma}$ is set equal to the survey budget necessary to run the LFS by the current sampling design (TSSh) for a quarter allocated by three operational domains: “Riga”, “Cities” and “Towns and rural areas”. The estimation of $\boldsymbol{\gamma}$ is done by a Monte Carlo simulation experiment.

The expected values of d_l, a_l and b_l are estimated by a Monte Carlo simulation experiment where l is the operational domain index. A sample is selected by the TSSh and the values of d_l, a_l and b_l are computed in each iteration. The total number of the iterations of the simulation is 6000. The values of K_f, C_f, C_a, C_b and k_d are set according to the available information about the LFS fieldwork organisation.

The resulting total survey cost for a quarter with TSSh design is 36 004.8. The allocation of the survey cost by operational domains and resulting field work budget is set as $\boldsymbol{\gamma} = \{5395.1, 7719.5, 22\,890.1\}$ (for “Riga”, “Cities” and “Towns and rural areas” accordingly).

4.4. Design parameters of alternative sampling designs

The mSSRSi and mSSRSh are chosen as alternative sampling designs. The stratification of both designs is set equal to the operational domains of TSSh. Therefore, three strata (“Riga”, “Cities”, “Towns and rural areas”) are created for each design. Units are individuals for the mSSRSi and units are households for the mSSRSh. A sample size is estimated independently for each design and each stratum (six cases). A stratum sample size n_h is the only parameter for the designs. The valid values of n_h are

$$\{n_h : (0 < n_h \leq M_h \; \& \; n_h \bmod 13 = 0)\}$$

where M_h is the total number of units stratum h. The aim is to find n_h so that $C(n_h) \approx \gamma_h$ where $C(n_h)$ is the expected survey cost with sample size n_h, and γ_h

is the survey budget for a stratum h. The solution is defined as

$$n_h^* = \underset{\{n_h : C(n_h) > \gamma_h\}}{\arg\min} C\left(n_h\right).$$

The solution is found by a stepwise procedure for each design and each stratum independently:

- Eight values of n_h widely spread in the interval of valid sample sizes are selected and $C\left(n_h\right)$ is estimated for each selected n_h with a Monte Carlo simulation.
- The relation between the expected cost and sample size is approximated by a non-linear regression $C\left(n_h\right) \sim \beta_0 + \beta_1 n_h + \beta_2 \sqrt{n_h}$. The regression coefficients β_0, β_1 and β_2 are estimated from the eight pairs of $\left\{n_h, \hat{C}\left(n_h\right)\right\}$.
- An approximate solution $\hat{n}_h^*$ is computed from the regression equation by
$$\hat{n}_h^* = \frac{\left(\sqrt{\hat{\beta_2}^2 - 4\hat{\beta_1}\left(\hat{\beta_0} - \gamma_h\right)} - \hat{\beta_2}\right)^2}{4\hat{\beta_1}^2}.$$
- It has been observed that the exact solution n_h^* is close to $\hat{n}_h^*$. The exact solution is found by another Monte Carlo simulation experiment estimating the cost for a sampling design with seven different sample sizes close to $\hat{n}_h^*$. The sample sizes chosen for the simulation are $\hat{n}_h^* - 39, \hat{n}_h^* - 26, \hat{n}_h^* - 13, \hat{n}_h^*, \hat{n}_h^* + 13, \hat{n}_h^* + 26, \hat{n}_h^* + 39$.

The resulting sample size and survey cost for each stratum and sampling design are available in Table 4, where table columns are: `n.PSU` – number of PSUs, `n.h` – number of households in sample, `n.i` expected number of individuals in sample, `c.travel` – expected travel cost, `c.interview` – expected interview cost, `c.total` – expected total survey cost (the total survey cost is slightly higher than the budget available for mSSRSi and mSSRSh sampling designs to preserve a conservative position with respect to the TSSh).

4.5. Parameters of interest

There are six parameters considered:

- `a.empl` – the average of weekly totals of employed individuals,
- `a.unem` – the average of weekly totals of unemployed individuals,
- `a.inact` – the average of weekly totals of economically inactive individuals,
- `r.act` – the activity rate (the total number of employed and unemployed individuals by the total number of working-age individuals),

Table 4. Sample size and survey cost by stratum and sampling design

stratum	design	n.PSU	n.h	n.i	c.travel	c.interview	c.total
Riga	mSSRSi	.	.	1 261	403.2	5 036.6	5 439.8
Riga	mSSRSh	.	1 001	2 105	351.7	5 107.6	5 459.4
Riga	TSSh	104	1 040	2 185	90.6	5 304.6	5 395.1
Cities	mSSRSi	.	.	1 781	660.1	7 099.0	7 759.2
Cities	mSSRSh	.	1 404	2 963	581.9	7 174.6	7 756.5
Cities	TSSh	208	1 456	3 073	278.8	7 440.8	7 719.5
Other	mSSRSi	.	.	2 834	11 631.4	11 301.7	22 933.1
Other	mSSRSh	.	2 340	5 554	10 356.7	12 573.7	22 930.4
Other	TSSh	416	3 536	8 318	3 964.2	18 925.9	22 890.1

- `r.empl` – the employment rate (the total number of employed individuals by the total number of working-age individuals),
- `r.unem` – the unemployment rate (the total number of unemployed individuals by the total number of employed and unemployed individuals).

Six parameters are estimated for the whole target population and also in breakdowns by domains. Three sets of domains are considered:

- geographical domain (4) – Riga, cities (excluding Riga), towns, and rural areas,
- age group (2) – individuals aged 15–24 and 25–74 years,
- geographical domain (4) × age group (2).

It makes 90 parameters (45 averages of weekly totals and 45 ratios of two totals) selected for the cost efficiency analysis.

4.6. Variance of parameter estimators

The variance of $\hat{Y}_q$ (1) by the mSSRSi and the mSSRSh is computed by

$$\text{Var}\left(\hat{Y}_q\right) = \frac{1}{169}\sum_{h=1}^{H}\left(\frac{M_h^2}{m_h}\sum_{w} S_{w,h}^2\left(y\right) - M_h\sum_{w}\sum_{v} S_{w,v,h}\left(y\right)\right)$$

where h is a stratum index, H is the total number of strata, M_h is the total number of units in the unit population of a stratum h, m_h is the total number of units in the sample of a stratum h, $S_{w,h}^2\left(y\right)$ is the variance of a variable y in week w and a stratum h, and $S_{w,v,h}\left(y\right)$ is the covariance of a variable y between weeks w and v in a stratum h. The approximate variance of $\hat{R}_q$ (2) by the mSSRSi and the mSSRSh is computed by

$$\text{AVar}\left(\hat{R}_q\right) = \frac{1}{Z_q^2}\sum_{h=1}^{H}\left(\frac{M_h^2}{m_h}\sum_{w} S_{w,h}^2\left(u\right) - M_h\sum_{w}\sum_{v} S_{w,v,h}\left(u\right)\right)$$

where Z_q is the denominator of R_q, and u is the so called linearised variable for the ratio of two totals (Särndal et al., 1992, p.178). The variance of $\hat{Y}_q$ and $\hat{R}_q$ by the TSSh is estimated by a Monte Carlo simulation experiment.

4.7. Cost efficiency analysis

The three selected designs are compared by their cost efficiency using Definition 1 for the estimation of each selected parameter. A hypothesis testing is used in the case when the estimate of the variance by the TSSh is compared to the variance by the mSSRSi or the mSSRSh. An assumption is made that the estimates of the parameters by the TSSh are normally distributed:

$$\hat{\theta} \sim N\left(\mu, \sigma^2\right)$$

where σ^2 is unknown and is estimated by $s^2 = s^2\left(\boldsymbol{x}\right)$ from the data $\boldsymbol{x}$ of the simulation experiment. The length of $\boldsymbol{x}$ is equal to the total number of iterations in the simulation, $|\boldsymbol{x}| = J = 20\,000$ in this case. The aim is to compare σ^2 by the TSSh with the known σ_0^2 under alternative design. A one-sided hypothesis testing (Wasserman, 2004) is done:

$$\begin{aligned} H_0 &: \sigma^2 \geq \sigma_0^2, \\ H_1 &: \sigma^2 < \sigma_0^2. \end{aligned} \tag{3}$$

A test statistic is computed as

$$T\left(\boldsymbol{x}\right) = \frac{\left(J-1\right)s^2}{\sigma_0^2},$$

and a rejection region R is defined as

$$R = \left\{\boldsymbol{x} : T\left(\boldsymbol{x}\right) \leq c\right\}$$

where $c = F_{J-1}^{-1}\left(\alpha\right)$ is the value of the inverse cumulative distribution function of χ_{J-1}^2 at α. The following statements with respect to H_0 are set:

$$\begin{aligned} &T\left(\boldsymbol{x}\right) \leq c \Rightarrow \text{reject } H_0, \\ &T\left(\boldsymbol{x}\right) > c \Rightarrow \text{retain (do not reject) } H_0. \end{aligned}$$

The smallest α which rejects H_0 is called p-value, and p-value is equal to the value of the cumulative distribution function of χ_{J-1}^2 at the point $\frac{(J-1)s^2}{\sigma_0^2}$.

The most cost efficient sampling design for the estimation of a parameter is determined by the following procedure:

1. The value of σ_0^2 is computed as $\min\left(\sigma_{mSSRSi}^2, \sigma_{mSSRSh}^2\right)$.
2. The hypothesis testing (3) is done by computing p-value.

3. The TSSh is chosen as the most cost efficient sampling design for a parameter and the procedure stops here if p-value is less than 0.01. The procedure is continued to the step 4 if p-value is equal or greater than 0.01.
4. The mSSRSi is chosen as the most cost efficient sampling design for a parameter if $\sigma^2_{mSSRSi} < \sigma^2_{mSSRSh}$, and the mSSRSh is chosen as the most cost efficient sampling design for a parameter otherwise.

The expected precision of parameter estimates by the three sampling designs and the most efficient sampling design determined is given in Tables 5, 6, 7, and 8. The columns of the tables are:

- `param`: the name of parameter,
- `dom`: five geographical domains – "Latvia", "Riga", "Cities" (excluding city Riga), "Towns" or "Rural" (rural areas),
- `age`: three age groups – "15–74", "15–24" or "25–74",
- `value`: the true value of a population parameter computed from the artificial population data,
- σ_1: the expected standard error of an estimate by the mSSRSi,
- σ_2: the expected standard error of an estimate by the mSSRSh,
- σ_3: the estimated standard error of an estimate by the TSSh,
- p-`val`: p-value of the hypothesis testing (3),
- `des.eff`: the most cost efficient sampling design determined by the framework – "mSSRSi", "mSSRSh" or "TSSh".

Table 5. Precision of the estimates for the average of weekly totals in Latvia

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
a.empl	Latvia	15–74	972 327	11 034	12 061	11 437	1.000	mSSRSi
a.unem	Latvia	15–74	133 746	6 173	4 958	4 654	0.000	TSSh
a.inact	Latvia	15–74	545 052	10 513	9 109	8 605	0.000	TSSh
a.empl	Latvia	15–24	102 838	5 410	4 344	4 097	0.000	TSSh
a.unem	Latvia	15–24	27 693	2 868	2 191	2 034	0.000	TSSh
a.inact	Latvia	15–24	157 176	6 487	5 373	5 078	0.000	TSSh
a.empl	Latvia	25–74	869 489	11 204	10 802	10 150	0.000	TSSh
a.unem	Latvia	25–74	106 054	5 565	4 393	4 121	0.000	TSSh
a.inact	Latvia	25–74	387 876	9 499	7 800	7 282	0.000	TSSh

The efficiency of the sampling designs strongly depends on a domain and the type of a parameter. The mSSRSi is selected as the most efficient design only for three parameters – "the average of weekly totals of employed individuals" in the domains "Latvia", "Riga" and "Cities". The mSSRSh is reasonably efficient for the estimation of the averages of totals in the domain "Riga" – it has been selected as the most efficient design in five out of nine cases. There are five other parameters

Table 6. Precision of the estimates for the ratio of two totals in Latvia

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
r.act	Latvia	15–74	0.670	0.0064	0.0049	0.0045	0.000	TSSh
r.empl	Latvia	15–74	0.589	0.0067	0.0051	0.0048	0.000	TSSh
r.unem	Latvia	15–74	0.121	0.0055	0.0043	0.0040	0.000	TSSh
r.act	Latvia	15–24	0.454	0.0161	0.0122	0.0113	0.000	TSSh
r.empl	Latvia	15–24	0.357	0.0155	0.0118	0.0109	0.000	TSSh
r.unem	Latvia	15–24	0.212	0.0197	0.0148	0.0138	0.000	TSSh
r.act	Latvia	25–74	0.716	0.0067	0.0053	0.0050	0.000	TSSh
r.empl	Latvia	25–74	0.638	0.0072	0.0057	0.0053	0.000	TSSh
r.unem	Latvia	25–74	0.109	0.0056	0.0044	0.0040	0.000	TSSh

in the domains “Riga” and “Cities” which are the most efficiently estimated by the mSSRSh.

The TSSh is the most efficient design for the estimation of ratios in the domain “Riga” and also for the estimation of totals and ratios in the domain “Cities”. The TSSh dominates in the domains “Towns” and “Rural areas” – all parameters in these domains are the most efficiently estimated by the TSSh. It is because travelling distances are longer in these domains compared to the domains “Riga” and “Cities”. The TSSh is the most efficient also for the estimation of the parameters representing the domain “Latvia” (only one parameter for the domain “Latvia” is more efficiently estimated by the mSSRSi).

The cost efficiency analysis is done from a conservative position with respect to the TSSh. Firstly, the total sample size of each stratum for the mSSRSi and the mSSRSh is chosen slightly larger compared to the TSSh (Section 4.4).

Secondly, the TSSh is chosen as the most efficient design only in the cases when it is supported by strong evidence (p-value of the hypothesis testing is less than 0.01). The mSSRSi and the mSSRSh are preferred in the cases when there is uncertainty in the determination of the most efficiency design. For example, there are several cases when the precision of estimates achieved by the mSSRSh and the TSSh is quite similar.

The TSSh sampling design can be used reasonably well in some of these cases even if the mSSRSh has been chosen as the most efficient design, for example, in cases for the estimation of the average of weekly totals of inactive individuals in the domain “Riga” and the average of weekly totals of employed individuals aged 25–74 in the domain “Riga” (these are the cases when p-value is slightly higher than 0.01).

The TSSh has achieved the highest precision of estimates in most cases despite the conservative position with respect to it. Therefore, it is recommended to use the currently used two-stage sampling design for the Latvian LFS to achieve the

Table 7. Precision of the estimates for the average of weekly totals

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
a.empl	Riga	15–74	330 855	7 381	8 272	8 329	1.000	mSSRSi
a.unem	Riga	15–74	47 160	4 284	3 569	3 504	0.000	TSSh
a.inact	Riga	15–74	160 949	6 938	6 062	6 009	0.040	mSSRSh
a.empl	Riga	15–24	31 245	3 543	2 903	2 960	1.000	mSSRSh
a.unem	Riga	15–24	8 152	1 851	1 452	1 435	0.011	mSSRSh
a.inact	Riga	15–24	40 138	3 980	3 300	3 301	0.508	mSSRSh
a.empl	Riga	25–74	299 610	7 533	7 509	7 430	0.017	mSSRSh
a.unem	Riga	25–74	39 007	3 928	3 222	3 184	0.008	TSSh
a.inact	Riga	25–74	120 810	6 322	5 329	5 250	0.001	TSSh
a.empl	Cities	15–74	196 200	3 870	4 304	4 126	1.000	mSSRSi
a.unem	Cities	15–74	26 352	2 125	1 746	1 713	0.000	TSSh
a.inact	Cities	15–74	110 307	3 703	3 250	3 261	0.754	mSSRSh
a.empl	Cities	15–24	19 779	1 860	1 532	1 500	0.000	TSSh
a.unem	Cities	15–24	5 362	991	782	764	0.000	TSSh
a.inact	Cities	15–24	30 430	2 267	1 903	1 846	0.000	TSSh
a.empl	Cities	25–74	176 421	3 926	3 878	3 736	0.000	TSSh
a.unem	Cities	25–74	20 990	1 913	1 536	1 510	0.000	TSSh
a.inact	Cities	25–74	79 877	3 360	2 839	2 850	0.784	mSSRSh
a.empl	Towns	15–74	166 623	5 991	6 139	3 325	0.000	TSSh
a.unem	Towns	15–74	23 376	2 493	1 935	1 395	0.000	TSSh
a.inact	Towns	15–74	96 256	4 808	4 206	2 549	0.000	TSSh
a.empl	Towns	15–24	17 418	2 160	1 687	1 203	0.000	TSSh
a.unem	Towns	15–24	5 101	1 179	873	639	0.000	TSSh
a.inact	Towns	15–24	29 682	2 797	2 284	1 593	0.000	TSSh
a.empl	Towns	25–74	149 205	5 749	5 487	2 967	0.000	TSSh
a.unem	Towns	25–74	18 275	2 212	1 676	1 224	0.000	TSSh
a.inact	Towns	25–74	66 574	4 085	3 361	2 167	0.000	TSSh
a.empl	Rural	15–74	278 650	7 004	7 761	5 583	0.000	TSSh
a.unem	Rural	15–74	36 859	3 103	2 405	2 085	0.000	TSSh
a.inact	Rural	15–74	177 540	6 129	5 698	4 516	0.000	TSSh
a.empl	Rural	15–24	34 396	3 001	2 401	2 043	0.000	TSSh
a.unem	Rural	15–24	9 078	1 568	1 165	1 023	0.000	TSSh
a.inact	Rural	15–24	56 926	3 802	3 252	3 013	0.000	TSSh
a.empl	Rural	25–74	244 254	6 779	6 787	4 821	0.000	TSSh
a.unem	Rural	25–74	27 781	2 710	2 043	1 754	0.000	TSSh
a.inact	Rural	25–74	120 615	5 285	4 461	3 473	0.000	TSSh

Table 8. Precision of the estimates for the ratio of two totals

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
r.act	Riga	15–74	0.701	0.0129	0.0101	0.0099	0.000	TSSh
r.empl	Riga	15–74	0.614	0.0137	0.0109	0.0106	0.000	TSSh
r.unem	Riga	15–74	0.125	0.0111	0.0090	0.0088	0.000	TSSh
r.act	Riga	15–24	0.495	0.0366	0.0287	0.0281	0.000	TSSh
r.empl	Riga	15–24	0.393	0.0358	0.0281	0.0277	0.001	TSSh
r.unem	Riga	15–24	0.207	0.0422	0.0329	0.0329	0.542	mSSRSh
r.act	Riga	25–74	0.737	0.0134	0.0109	0.0107	0.000	TSSh
r.empl	Riga	25–74	0.652	0.0145	0.0119	0.0116	0.000	TSSh
r.unem	Riga	25–74	0.115	0.0113	0.0092	0.0090	0.000	TSSh
r.act	Cities	15–74	0.669	0.0111	0.0088	0.0086	0.000	TSSh
r.empl	Cities	15–74	0.589	0.0116	0.0092	0.0091	0.001	TSSh
r.unem	Cities	15–74	0.118	0.0093	0.0075	0.0073	0.000	TSSh
r.act	Cities	15–24	0.452	0.0288	0.0227	0.0221	0.000	TSSh
r.empl	Cities	15–24	0.356	0.0277	0.0218	0.0213	0.000	TSSh
r.unem	Cities	15–24	0.213	0.0352	0.0275	0.0269	0.000	TSSh
r.act	Cities	25–74	0.712	0.0117	0.0097	0.0095	0.013	mSSRSh
r.empl	Cities	25–74	0.636	0.0125	0.0103	0.0102	0.069	mSSRSh
r.unem	Cities	25–74	0.106	0.0095	0.0076	0.0074	0.000	TSSh
r.act	Towns	15–74	0.664	0.0146	0.0105	0.0079	0.000	TSSh
r.empl	Towns	15–74	0.582	0.0153	0.0111	0.0082	0.000	TSSh
r.unem	Towns	15–74	0.123	0.0125	0.0092	0.0069	0.000	TSSh
r.act	Towns	15–24	0.431	0.0359	0.0263	0.0195	0.000	TSSh
r.empl	Towns	15–24	0.334	0.0342	0.0249	0.0184	0.000	TSSh
r.unem	Towns	15–24	0.227	0.0462	0.0334	0.0246	0.000	TSSh
r.act	Towns	25–74	0.716	0.0154	0.0116	0.0087	0.000	TSSh
r.empl	Towns	25–74	0.637	0.0165	0.0125	0.0092	0.000	TSSh
r.unem	Towns	25–74	0.109	0.0126	0.0093	0.0070	0.000	TSSh
r.act	Rural	15–74	0.640	0.0113	0.0083	0.0072	0.000	TSSh
r.empl	Rural	15–74	0.565	0.0117	0.0086	0.0074	0.000	TSSh
r.unem	Rural	15–74	0.117	0.0095	0.0070	0.0061	0.000	TSSh
r.act	Rural	15–24	0.433	0.0259	0.0188	0.0166	0.000	TSSh
r.empl	Rural	15–24	0.343	0.0248	0.0180	0.0156	0.000	TSSh
r.unem	Rural	15–24	0.209	0.0323	0.0234	0.0204	0.000	TSSh
r.act	Rural	25–74	0.693	0.0122	0.0093	0.0081	0.000	TSSh
r.empl	Rural	25–74	0.622	0.0128	0.0098	0.0084	0.000	TSSh
r.unem	Rural	25–74	0.102	0.0096	0.0071	0.0061	0.000	TSSh

highest overall precision under the current budget constrains. Switching to a simpler sampling design will result in one of two negative effects. The first possible negative effect is the loss of overall precision if the survey cost is kept in the current budget level. The second possible negative effect is the increase in the survey cost if overall precision level is kept equal to the current level.

5. Conclusions

The aim of this paper was to develop a mathematical framework to compare chosen sampling designs with respect to the expected precision of estimates and the data collection cost. The framework has been developed and its application in case of Latvian Labour Force Survey has been demonstrated. The framework presented in the paper utilises Monte Carlo simulation experiment techniques when analytical methods can not be applied.

The framework allows the user to gain information about the sampling design properties (for example, the expected fieldwork cost or the expected precision of estimates) in a relatively short time and with relatively low cost. This information is very valuable for survey planning and the decision making processes. The advantage of the framework is that no extra data collection is required. The framework utilises data already available to a statistical agency (administrative records, population census data or sample survey data).

A set of procedures is developed to support the implementation of the framework in practice. The aim of the procedures is to run Monte Carlo simulations of sampling designs. The procedures are developed in R which is a free software environment for statistical computing and graphics (R Core Team, 2013). The code of the procedures is available online at the “GitHub” repository (Liberts, 2013). The procedures are developed as modular functions. It allows for the extension of the procedures with additional functions if necessary. There is no limitation on the types of design that can be analysed by the procedures. The only requirement is that it must be possible to describe the sampling process of a design as an R function.

The cost efficiency of three sampling designs is analysed using the framework. The properties of the chosen sampling designs are explored and recommendations with respect to an appropriate sampling design for the Latvian LFS are given. It is proven that the two-stage sampling design used currently for the LFS provides more precise parameter estimates under the condition of equal fieldwork cost when compared to two other simpler sampling designs.

The developed framework for cost efficiency analysis is flexible. It can be applied for different surveys and arbitrary sampling designs. There are broad possibilities of tuning the framework to specific aspects under analysis, for example, the survey cost estimation can be extended to take into account other processes from real fieldwork operations. The developed framework can be used both by national statistical agencies and private companies organising sample surveys.

The research can be continued by extending the framework with non-response modelling. The set of the developed R procedures has to be extended with additional procedures. The additional procedure is necessary to simulate the process of the non-response of sampled units. The cost function has to be adjusted to take into account the actions done by interviewers in the case of non-response. The procedure for estimation of the population parameters in the case of non-response is necessary.

Acknowledgements

ESF
EIROPAS SOCIĀLAIS FONDS

EIROPAS SAVIENĪBA

LATVIJAS UNIVERSITĀTE
ANNO 1919

IEGULDĪJUMS TAVĀ NĀKOTNĒ

This work has been supported by the European Social Fund within the project **«Support for Doctoral Studies at University of Latvia – 2»**.

I would like to thank the anonymous referees for their constructive comments and suggestions, helping to improve the manuscript.

REFERENCES

Central Statistical Bureau of Latvia., (2012). *Employment and unemployment* [Metadata]. Riga. Retrieved 15.12.2012, from http://ej.uz/CSB-LFS

CHEN, B.-C., (2008). *Stochastic simulation of field operations in surveys* (Research report). Washington: U. S. Census Bureau. Retrieved from https://www.census.gov/srd/www/byyear.html

COX, L., (2012). The case for simulation models of federal surveys. In *Research conference papers of federal committee on statistical methodology research conference 2012*. Washington.
Retrieved from http://www.fcsm.gov/events/papers2012.html

European Commission. (2012a). *Labour force survey in the EU, candidate and EFTA countries – Main characteristics of national surveys, 2011* (Tech. Rep.). Luxembourg: Eurostat. Retrieved from http://epp.eurostat.ec.europa.eu/

European Commission. (2012b). *Quality report of the European Union Labour Force Survey – 2010* (Tech. Rep.). Luxembourg: Eurostat. Retrieved from http://epp.eurostat.ec.europa.eu/

GROVES, R. M., (1989). *Survey errors and survey costs*. New Jersey: Wiley.

HANSEN, M. H., HURWITZ, W. N., & MADOW, W. G., (1953). *Sample survey methods and theory* (Vol. I). New-York: Wiley.

JESSEN, R. J., (1942). *Statistical investigation of a sample survey for obtaining farm facts* (Research Bulletin No. 304). Iowa State College of Agriculture and Mechanic Arts. KISH, L. (1965). *Survey sampling*. New-York: John Wiley & Sons.

LAPIŅŠ, J., (1997). Sampling surveys in Latvia: Current situation, problems and future development. *Statistics in Transition*, *3*(2), 281–292.

LAPIŅŠ, J., VASKIS, E., PRIEDE, Z., & BĀLIŅA, S., (2002). Household surveys in Latvia. *Statistics in Transition*, *5*(4), 617–641. Retrieved from http://www.stat.gov.pl/pts/15_ENG_HTML.htm

LIBERTS, M., (2010). The redesign of Latvian Labour Force Survey. In M. Carlson, H. Nyquist, & M. Villani (Eds.), *Official statistics – methodology and applications in honour of Daniel Thorburn* (pp. 193–203). Stockholm, Sweden: Stockholm University. Retrieved from http://officialstatistics.wordpress.com/

LIBERTS, M., (2013). *Survey-design-simulation* [Online code repository]. Retrieved from https://github.com/djhurio/Survey-Design-Simulation

MAHALANOBIS, P. C., (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics*, *4*(4), 511–530.

R CORE TEAM., (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org

ROSENKRANTZ, D., STEARNS, R., & LEWIS, P., II., (1977). An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, *6*(3), 563–581.

SÄRNDAL, C.-E., SWENSSON, B., & WRETMAN, J., (1992). *Model assisted survey sampling*. New-York: Springer.

United Nations. (2010). *Handbook on population and housing census editing: Revision 1*. New York: United Nations.

WASSERMAN, L.,(2004). *All of statistics*. New-York: Springer.