Statistical Office
of the Republic of Serbia

WORKING PAPER

Estimation of standard error of an index in case of coordinated samples

Number 102
Year LIII

Belgrade, 2017

REPUBLIC OF SERBIA
STATISTICAL OFFICE OF THE REPUBLIC OF SERBIA

WORKING PAPER

ISSN 1820 – 0141

Estimation of standard error of an index in case of coordinated samples

Year – LIII Belgrade, 2017 Number 102

Estimation of standard error of an index in case of coordinated samples

Publisher: Statistical Office of the Republic of Serbia, Belgrade, 5, Milana Rakića
For the publisher: Dr Miladin Kovačević, Director

Authors: Olga Melovski Trpinac and Gordana Zamaklar

Translation into English: Gordana Nedeljković

Technical editor: Irena Dimić

Foreword

This working paper presents the results of the simulation study on precision estimation of a parameter of change when samples are coordinated with permanent random numbers.

The work is based on ideas and methods proposed by our Swedish colleagues, Annika Lindblom and Stefan Berg, during their expert missions. Their missions took place in the period 2013 – 2016 within the Component on Survey Methodology of the project Partnership in Statistics funded by the Swedish International Development Cooperation Agency – SIDA.

Determining the procedure for estimation of precision of an index is in line with the activities whose aim is to develop business survey methodology at the Statistical Office of the Republic of Serbia. In 2013, also with the help of Swedish experts, the coordination was started. Since 2015, sampling frames are constructed on the basis of frozen versions of the Statistical Business Register. Also, sample designs, as well as the methods of sample allocation and selection are standardized: sample designs are stratified with selection of a simple random or a Pareto sample within strata; for most of the surveys, allocation is conducted using combination of Hidirouglu (1986) and Bethel (1989) algorithms; sequential sample selection scheme with the use of permanent random numbers is adopted. Estimation procedures are standardized, too. Sampling weights are corrected for non-response and outliers. As a rule, the computed estimates of parameters are accompanied with measures of precision, non-response and over-coverage errors, as well as with the information on outliers and imputations, etc.

The working paper "Estimation of the standard error of an index in case of coordinated samples" is also presented in electronic form on the internet page www.stat.gov.rs.

Belgrade, 2017

Director
Dr Miladin Kovačević

Contents

1. Introduction

Among the basic tasks of short term statistics (STS) is calculation of the index[1] – relative number that is used to represent the change in value or volume in time. It is estimated simply by computing the ratio of the two estimates of totals, current to the one of the previous occasion. The estimate of this parameter is usually not accompanied by the estimate of the standard error. One reason for this is that the precise estimates of the totals are in a way a guarantee for precise estimate of the corresponding index. Another reason is that it is often difficult to estimate the precision for the parameter of change.

In business surveys, successive samples are selected with some kind of rotation scheme to be positively coordinated – partly overlapping. Also, successive samples are selected from frame populations that differ because of unit births, deaths or their reclassifications in terms of size or industry. It could also happen that samples do not have the same design. These characteristics of business surveys make the problem of variance estimation for the measure of change complex.

Positive coordination of the samples induces positive correlation of the estimators of totals. As a consequence, the variance of the estimator of the ratio (index) is smaller than it would be if the samples were independent and when it depends only upon the totals and variances of the estimators of these totals, but not on their correlation. In case of coordinated samples, for a more precise estimator of the variance of an index, it is necessary to estimate the temporal correlation/covariance between the estimators of two totals.

Sampling coordination has been introduced at the Statistical Office of the Republic of Serbia (SORS) in 2013 (Melovski Trpinac O. et al, 2014). It is based on the Swedish method (SAMU, 2003). To each enterprise of the Statistical Business Register (SBR), this system associates a permanent random number generated from the uniform distribution on the interval (0,1). Sample rotation is achieved by randomly stratifying SBR enterprises into five rotation groups of equal or almost equal size. Permanent random numbers of one rotation group are shifted for 0.10 each year.

At SORS, SBR is used for construction of sampling frames for business surveys. Most often, these surveys employ stratified simple random sample or stratified Pareto sample. Sequential scheme is used for sample selection. Units are ordered by permanent random numbers in ascending sequence and from each stratum, the first n_h units are selected, where n_h is equal to the number of sample units allocated to stratum h. Sample selection is made from the starting point from the interval (0,1), which is determined for each survey in advance so that sample coordination of different surveys is maintained. Under the described framework, successive samples of the same survey overlap in a random number of units and that additionally complicates variance estimation.

The problem of estimating the variance for a measure of change when samples are coordinated by permanent random numbers was addressed during the missions of the Statistics Sweden experts, A. Lindblom and S. Berg (reports SERSTAT 2013:22, SERSTAT 2014:07 and SERSTAT 2016:05). The idea that was exploited was to use in the estimation procedure adequately reduced correlation between estimates of totals on the overlapping part of samples. Several correction factors were proposed and their validity was checked in a simulation study. Simulations were conducted using sampling frame data of the Quarterly Structural Business Survey (SBS03) for the years 2012 and 2013, which were previously updated with data from the financial accounts for 2012 and 2013, respectively. The study variable was turnover.

Following the same idea and goal, Đ. Petković repeated the simulations but with the SBS03 data frames for the years 2013 and 2014 (Petković, 2015). He also considered certain additional correction factors that were suggested by sampling methodologists from SORS.

This paper refers to the results of simulations that were conducted using SBS03 sampling frames for the years 2015 and 2014. The purpose of these last simulations was to systematize all the work done on this problem till now and to come to recommendations – which of the suggested methods are most acceptable for the estimation of the precision of the annual indices of turnover, operating income and operating expenses of the Quarterly Structural Business Survey.

[1] In this paper, ‘index‘and ‘index estimate’ have equal meaning, if not pointed out differently.

There are various suggestions in the literature for solving the problem in question which consider different sample plans and selection procedures, as well as characteristics of the changing population. For the situation similar to ours, Norberg (2000) offered a solution that is based on decomposition of the covariance of the estimators of totals into conditional covariance term and remainder term and Lindblom (2014) explored in a simulation study the effect of design changes on the estimate of precision for a parameter of change. For somewhat different conditions, solutions are given in Laniel (1988), Hidiroglou et al. (1995), Berger (2004), Full and Lewis (2004) and in other documents.

In part 2, the formulas for approximate theoretical variance and its estimators are presented for the parameter of change. Several estimators of the temporal correlation of the estimators of totals, which are based on the correction of the correlation on the overlapping parts of samples, are proposed in part 3. An idea for a different approach to the estimation of error is presented in part 4. Part 5 contains the set up for the simulation study. The overview of the methodology for the Quarterly Structural Business Survey is given in part 6. The results of simulations are presented in part 7. Finally, brief concluding remarks are reported in part 8.

2. Estimation of index precision

The estimator of annual index, the ratio parameter of change, is $\hat{r} = \frac{\hat{t}_1}{\hat{t}_0}$, where: $\hat{t}_0$ is the estimator of total for month/quarter of year at time 0 and $\hat{t}_1$ is the estimator of total for month/quarter of year at time 1.

Using the method of Taylor linearization of the ratio, the variance of an index can be approximated (Sarndall et al, 1992, 178-179) in the following way:

$$V(\hat{r}) \approx AV(\hat{r}) = \left(\frac{t_1}{t_0}\right)^2 \left(\left(\frac{V(\hat{t}_0)}{t_0^2}\right) + \left(\frac{V(\hat{t}_1)}{t_1^2}\right) - \frac{2 \cdot C(\hat{t}_0, \hat{t}_1)}{t_0 \cdot t_1}\right) \quad (1)$$

where $V(\cdot)$ denotes the variance, $AV(\cdot)$ the approximate variance and $C(\hat{t}_0, \hat{t}_1)$ the covariance of estimators of totals. This approximate variance can be expressed by correlation, instead of the covariance:

$$V(\hat{r}) \approx AV(\hat{r}) = \left(\frac{t_1}{t_0}\right)^2 \left(\left(\frac{V(\hat{t}_0)}{t_0^2}\right) + \left(\frac{V(\hat{t}_1)}{t_1^2}\right) - \frac{2 \cdot \rho(\hat{t}_0, \hat{t}_1) \cdot \sqrt{V(\hat{t}_0) \cdot V(\hat{t}_1)}}{t_0 \cdot t_1}\right) \quad (2)$$

where $\rho(\hat{t}_0, \hat{t}_1) = \frac{C(\hat{t}_0, \hat{t}_1)}{\sqrt{V(\hat{t}_0) \cdot V(\hat{t}_1)}}$.

Since the coefficient of variation of an estimator equals the ratio of its standard error to the expected value (it is equal to parameter value for unbiased estimator), formula (2) can be further expressed in the following way:

$$CV^2(\hat{r}) \approx ACV^2(\hat{r}) = \left(CV^2(\hat{t}_0) + CV^2(\hat{t}_1) - 2 \cdot \rho(\hat{t}_0, \hat{t}_1) \cdot CV(\hat{t}_0) \cdot CV(\hat{t}_1)\right) \quad (3)$$

where $CV(\cdot)$ denotes the coefficient of variation, $ACV(\cdot)$ the approximate coefficient of variation. The value of $ACV^2(\hat{r})$ is bounded from above and below

$$\left(CV(\hat{t}_0) - CV(\hat{t}_1)\right)^2 \leq ACV^2(\hat{r}) \leq \left(CV(\hat{t}_0) + CV(\hat{t}_1)\right)^2 \quad (4)$$

implying

$$|CV(\hat{t}_0) - CV(\hat{t}_1)| \leq ACV(\hat{r}) \leq CV(\hat{t}_0) + CV(\hat{t}_1) \quad (5)$$

with limiting values achieved for $\rho(\hat{t}_0, \hat{t}_1) = 1$ and $\rho(\hat{t}_0, \hat{t}_1) = -1$.

In case of non-negative correlation of totals, as is in positively coordinated samples, upper limit for $ACV(\hat{r})$ is lower and is achieved for $\rho(\hat{t}_0, \hat{t}_1) = 0$:

$$|CV(\hat{t}_0) - CV(\hat{t}_1)| \leq ACV(\hat{r}) \leq \sqrt{CV^2(\hat{t}_0) + CV^2(\hat{t}_1)} \tag{6}$$

The last inequality justifies the statement given in the first paragraph of the Introduction: If the totals are estimated precisely then the index is precisely estimated, too. But usually, in statistical inference, a more precise estimate of the error is needed than its limiting values.

Further in this paper, the following estimator of the variance (1) is used:

$$\hat{V}(\hat{r}) = \left(\frac{\hat{t}_1}{\hat{t}_0}\right)^2 \left(\left(\frac{\hat{V}(\hat{t}_0)}{\hat{t}_0^2}\right) + \left(\frac{\hat{V}(\hat{t}_1)}{\hat{t}_1^2}\right) - \frac{2 \cdot \hat{C}(\hat{t}_0, \hat{t}_1)}{\hat{t}_0 \cdot \hat{t}_1}\right) \tag{7}$$

or

$$\hat{V}(\hat{r}) = \left(\frac{\hat{t}_1}{\hat{t}_0}\right)^2 \left(\left(\frac{\hat{V}(\hat{t}_0)}{\hat{t}_0^2}\right) + \left(\frac{\hat{V}(\hat{t}_1)}{\hat{t}_1^2}\right) - \frac{2 \cdot \hat{\rho}(\hat{t}_0, \hat{t}_1) \cdot \sqrt{\hat{V}(\hat{t}_0) \cdot \hat{V}(\hat{t}_1)}}{\hat{t}_0 \cdot \hat{t}_1}\right) \tag{8}$$

When the sample is stratified and within strata simple random samples are selected, the estimates of totals and variances, in formulas (7) and (8), are easily calculated as:

$$\hat{t} = \sum_{h=1}^{H} \frac{N_h}{n_h} \cdot \sum_{i=1}^{n_h} y_{hi}, \quad \hat{V}(\hat{t}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h - 1} \cdot \left[\sum_{i=1}^{n_h} y_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} y_{hi}\right)^2}{n_h}\right] \tag{9}$$

where: H represents the number of strata in the frame; N_h and n_h denote the number of units in the frame and in the sample and y_{hi} the variable value of the sample unit i from stratum h. The formulas (9) are applied for time 0 and time 1.

3. Estimation of temporal correlation/covariance when samples are coordinated with permanent random numbers

The problem comes down to the estimation of correlation in formula (8). It could be estimated using only the overlapping part of samples and stratification of time 1, for example. But in this way, the information on non-overlapping parts of samples is not used and as a consequence the true correlation of the estimators of totals is over estimated. That is why this estimator of correlation should be corrected by a factor that represents the relative size of the common part of the samples.

Further in this document, the following notations are used at the level of domain of estimation:

Sample, stratification and domain of interest as at time 0	
Notation	**Explanation**
N_0	The number of units in the sampling frame
n_0	Sample size
n_{0a}	Number of common sample units at time 0 and time 1
$\widehat{N}_0, \widehat{N}_{0a}$	Estimator of the number of frame units: at time 0; present at both times 0 and 1
$\hat{t}_0, \hat{t}_{0a}, \hat{t}_{0b}$	Estimator of the total: for frame units at time 0; for units present in both frames, at time 0 and time 1; for units present in the frame at time 0 but not in the frame at time 1

$\hat{V}(\hat{t}_0), \hat{V}(\hat{t}_{0a})\ \hat{V}(\hat{t}_{0b})$	Variance estimators of $\hat{t}_0, \hat{t}_{0a}, \hat{t}_{0b}$
Sample, stratification and domain of interest as at time 1	
Notation	**Explanation**
N_1	The number of units in the sampling frame
n_1	Sample size
n_{1a}	Number of common sample units at time 0 and time 1. In case that common units have not changed their domain, from one time point to another, then $n_{0a} = n_{1a} = n_c$.
$\hat{N}_1, \hat{N}_{1a}$	Estimator of the number of frame units: at time 1; present at both times 0 and 1
$\hat{t}_1, \hat{t}_{1a}, \hat{t}_{1b}$	Estimator of the total: for frame units at time 1; for units present in both frames at time 0 and 1; for units present in the frame at time 1 but not in the frame at time 0
$\hat{V}(\hat{t}_1), \hat{V}(\hat{t}_{1a})\ \hat{V}(\hat{t}_{1b})$	Variance estimators of $\hat{t}_1, \hat{t}_{1a}, \hat{t}_{1b}$
Overlapping part of samples at time 0 and at time 1, with stratification and domain as at time 1	
Notation	**Explanation**
$\hat{t}_{0c}, \hat{t}_{1c}, \hat{t}_{0c} + \hat{t}_{1c}$	Estimators of totals based on overlapping part of the samples
$\hat{V}(\hat{t}_{0c}), \hat{V}(\hat{t}_{1c}), \hat{V}(\hat{t}_{0c} + \hat{t}_{1c})$	Variance estimators of $\hat{t}_{0c}, \hat{t}_{1c}, \hat{t}_{0c} + \hat{t}_{1c}$
The union of samples; for overlapping part of samples, domain of interest is as at time 1; for other units, domain depends upon the sample to which the unit belongs	
Notation	**Explanation**
n_{01}	Number of units in the union of samples at time 0 and time 1

A. Llindblom and S. Berg proposed the following factors for correction of the estimate of correlation based on the overlapping part of samples:

1) Correction factor based on variance estimators

$$c_1 = \frac{1}{2}\left(\frac{\hat{V}(\hat{t}_{0a})}{\hat{V}(\hat{t}_{0a}) + \hat{V}(\hat{t}_{0b})} + \frac{\hat{V}(\hat{t}_{1a})}{\hat{V}(\hat{t}_{1a}) + \hat{V}(\hat{t}_{1b})}\right) \quad \text{(M1)}$$

2) Correction factor based on sample sizes

$$c_2 = \frac{1}{2}\left(\frac{n_{0a}}{n_0} + \frac{n_{1a}}{n_1}\right) \quad \text{(M2)}$$

3) Correction factor based on the estimated number of frame units, at time 0 and time 1

$$c_3 = \frac{1}{2}\left(\frac{\hat{N}_{0a}}{\hat{N}_0} + \frac{\hat{N}_{1a}}{\hat{N}_1}\right) \quad \text{(M3)}$$

Correction factors suggested by the SORS staff:

4) Correction factor based on the number of units sampled both times and the union of sampled units (proposed by O. Melovski Trpinac)

$$c_4 = \frac{n_{1a}}{n_{01}} \quad \text{(M4)}$$

5) Correction factor based on the number of units sampled both times and the sum of units sampled (proposed by O. Melovski Trpinac)

$$c_5 = \frac{2 \cdot n_{1a}}{n_0 + n_1} \quad \text{(M5)}$$

M. Ogrizović Bašanac suggested using as the estimate of correlation of the totals of coordinated samples, the un-weighted correlation computed on the overlapping part of samples on the levels of domains of time 1, without any corrections. This method will be denoted as M6.

The estimator of the correlation is defined as a product of a correction factor and the estimator of correlation based on the overlapping part of samples

$$\hat{\rho}_m(\hat{t}_0, \hat{t}_1) = c_m \cdot \hat{\rho}(\hat{t}_{0c}, \hat{t}_{1c}),\ c_m = 1, \dots, 5 \quad (10)$$

The estimator $\hat{\rho}(\hat{t}_{0c}, \hat{t}_{1c})$ is derived using the relation

$$\hat{V}(\hat{t}_{0c} + \hat{t}_{1c}) = \hat{V}(\hat{t}_{0c}) + \hat{V}(\hat{t}_{1c}) + 2 \cdot \hat{\rho}(\hat{t}_{0c}, \hat{t}_{1c}) \cdot \sqrt{\hat{V}(\hat{t}_{0c}) \cdot \hat{V}(\hat{t}_{1c})} \quad (11)$$

as

$$\hat{\rho}(\hat{t}_{0c}, \hat{t}_{1c}) = \frac{\hat{V}(\hat{t}_{0c} + \hat{t}_{1c}) - \hat{V}(\hat{t}_{0c}) - \hat{V}(\hat{t}_{1c})}{2 \cdot \sqrt{\hat{V}(\hat{t}_{0c}) \cdot \hat{V}(\hat{t}_{1c})}} \quad (12)$$

Let $\rho_6(\hat{t}_0, \hat{t}_1)$ be the un-weighted correlation of totals based on the overlapping part of samples (method M6).

The estimators $\hat{V}_m(\hat{r})$ and $sde_m = \sqrt{\hat{V}_m(\hat{r})}$ $(m = 1, \dots, 6)$, of variance and standard error, are obtained by inserting in formula (8) the estimators $\hat{\rho}_m(\hat{t}_0, \hat{t}_1)$, $\hat{t}_0, \hat{t}_1, \hat{V}(\hat{t}_0)$ and, $\hat{V}(\hat{t}_1)$.

4. A different approach to estimation of index precision

A different approach than the previous one is inspired by formulas (3)-(6) that refer to the coefficient of variation of a ratio. With an assumption that the correlation of coordinated samples is not negative, the coefficient of variation could be estimated as

$$\widehat{CV}_7(\hat{r}) = \alpha \left| \widehat{CV}(\hat{t}_0) - \widehat{CV}(\hat{t}_1) \right| + (1 - \alpha) \sqrt{\widehat{CV}^2(\hat{t}_0) + \widehat{CV}^2(\hat{t}_1)} \quad \text{(M7)}$$

where $0 \leq \alpha \leq 1$. Parameter α should be determined in such a way so that it is proportional to the correlation of totals of coordinated samples.

Provided that the parameter α is adequately chosen, formula (M7) provides a quick estimate of the error without the estimation of the correlation on the overlapping part of samples.

Using formula (M7) and formula (3) with parameters replaced with their estimators, the estimator of correlation of totals is

$$\hat{\rho}_7(\hat{t}_0, \hat{t}_1) = \frac{\widehat{CV}^2(\hat{t}_0) + \widehat{CV}^2(\hat{t}_1) - \widehat{CV}_7^2(\hat{r})}{2 \cdot \widehat{CV}(\hat{t}_0) \cdot \widehat{CV}(\hat{t}_1)} \quad (12)$$

and the estimator of standard error is $sde_7 = \widehat{CV}_7(\hat{r}) \cdot \hat{r}$.

5. Simulation study

In order to compare estimators of the variance or of the coefficient of variation of an index, that are determined by one of the suggested methods, it is enough to compare the corresponding estimators of the correlation of the estimators of the two successive totals (see formulas (2), (3)).

Comparisons were conducted in a simulation study. The process of sample selection and estimation of the correlation was repeated K times. For each of the seven proposed methods, the average value of estimates of correlation from K pairs of coordinated samples was computed. At the same time, the simulation results were used to calculate the referent value for correlation – approximate true correlation of the estimators of totals.

The simulation setup is explained below in more details.

In each of the K simulations (K is the total number of simulations), the process of selection of coordinated samples, one from the sampling frame at time 0, the other from the sampling frame at time 1, is carried out in the following manner:

- to each frame unit at time 0 a permanent random numbers from the uniform distribution over the interval (0,1) is assigned;
- for sampling frame units at time 1 that have existed at time 0 (persisting units), permanent random numbers are taken over from time 0, while the permanent random numbers are generated for units of sampling frame 1 that did not exist in the frame at time 0;
- sampling frame units at time 1 are randomly distributed to five rotation groups of equal or almost equal size;
- permanent random numbers of persisting units of the frame at time 1 that belong to the first rotation group are shifted (rotated) for 0.1;
- sample selection follows sequential scheme: sampling frame units are sorted in ascending sequence by permanent random numbers and the first n_h enterprises from stratum h are included in the sample where n_h is the sample size from stratum h, $h=1,...,H$. The procedure is conducted both for time 0 and 1.

The described sampling technique is used in practice at SORS for generation of coordinated samples of enterprises.

- In each of the K simulations (K large enough, K=10,000), totals at time 0 and at time 1, $\hat{t}_{0k}$ and $\hat{t}_{1k}$, are estimated. Based on simulations, their variances and covariance are approximate as

$$V(\hat{t}_0) \approx \hat{V}_{sim}(\hat{t}_0) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{0k} - \hat{t}_{0sim})^2 \quad \text{(S1)}$$

$$V(\hat{t}_1) \approx \hat{V}_{sim}(\hat{t}_1) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{1k} - \hat{t}_{1sim})^2 \quad \text{(S2)}$$

$$C(\hat{t}_0, \hat{t}_1) \approx \hat{C}_{sim}(\hat{t}_0, \hat{t}_1) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{0k} - \hat{t}_{0sim}) \cdot (\hat{t}_{1k} - \hat{t}_{1sim}) \quad \text{(S3)}$$

where

$$\hat{t}_{0sim} = \frac{1}{K}\sum_{k=1}^{K}\hat{t}_{0k} \quad \text{(S4)}$$

and

$$\hat{t}_{1sim} = \frac{1}{K}\sum_{k=1}^{K} \hat{t}_{1k} \tag{S5}$$

Using relations (S1)-(S3), the value of the **approximate true correlation of estimators of the totals** from simulations is

$$\rho(\hat{t}_0, \hat{t}_1) \approx \hat{\rho}_{sim}(\hat{t}_0, \hat{t}_1) = \frac{\hat{C}_{sim}(\hat{t}_0, \hat{t}_1)}{\sqrt{\hat{V}_{sim}(\hat{t}_0) \cdot \hat{V}_{sim}(\hat{t}_1)}} \tag{S6}$$

If all possible outcomes of pairs of coordinated samples are generated by simulations, the true correlation of overlapping coordinated samples would be expressed by formula (S6).

The number of 10,000 replicates was sufficient for the convergence of simulation values (estimates after 1000, 2000... 10,000 iterations were compared).

The approximate true correlation, (S6), can be used as a **benchmark** for comparing estimators of correlation defined by one of the methods M1-M7. The following should be kept in mind:

- The expected value of the correlation estimator should be close to the value (S6).
- If the expected value of the correlation estimator is higher than the value (S6), then the corresponding estimator of variance underestimates the true variance of an index.
- If the expected value of the correlation estimator is lower than the value (S6) then corresponding estimator of variance overestimates true variance of an index.
- To be on the safe side, it is better that the expected value of the correlation estimator is lower than the true value (negatively biased), so that the corresponding estimator of the index variance is conservative (positively biased).
- Among the two estimators whose expected values do not differ much, the one with the smaller mean square is better.

If the approximate value of the correlation coefficient, $\hat{\rho}_{sim}(\hat{t}_0, \hat{t}_1)$ is denoted as $true$, and $\hat{\rho}_{mk}(\hat{t}_0, \hat{t}_1)$, is the estimate of correlation calculated in the k^{th} iteration of the simulation study by method m $(m = 1, \dots, 7)$, then the expected values (mean value) and the bias of an estimator for the method m are:

$$\rho_m(\hat{t}_0, \hat{t}_1) \approx \frac{1}{K}\sum_{k=1}^{K} \hat{\rho}_{mk}(\hat{t}_0, \hat{t}_1) \tag{13}$$

$$B_m\big(\hat{\rho}_m(\hat{t}_0, \hat{t}_1)\big) \approx \rho_m(\hat{t}_0, \hat{t}_1) - true \tag{14}$$

The mean square error of the estimator of correlation for the method m is

$$MSE_m\big(\hat{\rho}_m(\hat{t}_0, \hat{t}_1)\big) \approx \frac{1}{K}\sum_{k=1}^{K} (\hat{\rho}_{mk}(\hat{t}_0, \hat{t}_1) - true)^2 \tag{15}$$

6. Quarterly Structural Business Survey

The aim of Quarterly Structural Business Survey (SBS03) is to provide data on quarterly dynamics of financial operating of enterprises as well as on changes on the structure of economic activities in the field of nonfinancial operating economy. Financial operating of enterprises is based on the data on operating income, operating expenses, stocks and investments in tangible fixed assets. Within the "Operating income", data on "Revenues from the sale of goods, products and services" are collected, which, in turn, represent the realized turnover.

The data obtained by this survey are primarily used for the calculation of quarterly macroeconomic aggregates. In addition, these data are also used for the following purposes: ongoing monitoring and study of movements in non-financial business economy; the ongoing analysis of the effects of economic policy measures in the area of non-financial business economics; studying the development and structural changes in non-financial business economy, especially in the area of business services, as well as fulfilling obligations towards international organizations in the part of short-term indicators on the turnover of business services.

Quarterly Structural Business Survey encompasses all business entities involved in the production and sale of goods and services for the market, i.e. those entities that are mainly classified as non-financial business of NACE Rev. 2 classification of activities (sections A-S, excluding sections K and O). An enterprise is a statistical and reporting unit.

Below is a short description of the SBS03 sampling design and the estimation procedure that refer to the years 2014 and 2015.

The sampling frame is constructed using the Statistical Business Register, the version of 31st of December, year $t-1$, where t is the current year. The frame list consists of the data of active enterprises that have reported annual financial report for the year $t-2$. Turnover and number of employees are key auxiliary variables with values taken over from financial statements. The frame enterprises cover at least 95% of turnover, by NACE Rev. 2 divisions.

The frame consisted of about 27 thousand enterprises in 2014 and about 28 thousand enterprises in 2015.

The sampling frame is stratified according to the NACE classification of activities, number of employees and turnover.

The stratification according to NACE is applied in divisions 01 - 82 and four sections: P, Q, R and S (in all 74 strata).

The stratification of enterprises according to number of employees is implemented in two classes:

- with less than 50 employees and
- 50 and more employees.

The further stratification of frame units is done according to the value of turnover:

- enterprises with smaller turnover that are sampled and
- enterprises with larger turnover that are enumerated completely.

On recommendation by the SORS expert group for business surveys, Serbian Oil Company is divided into nine parts that are classified in special strata according to their activity and the number of employees.

The final stratification is defined by cross classifying activity strata with classes according to number of employees and turnover. In all, there are about 270 strata, of which almost half are census strata (8 census strata are reserved for the Oil Company parts).

The sample allocation by strata is performed with Bethel algorithm (Bethel, 1989). The anticipated coefficients of variation for the estimates of auxiliary variables totals (turnover and number of employees) by domains for 2014 and 2015 are given in Table 1 below.

Table 1. Anticipated errors for estimating auxiliary variable totals by domains in 2014 and 2015

Domain	Number of domains	Coefficient of variation (%)	
		Turnover	Number of employees
NACE divisions 01–82 and sections P, Q, R and S covered by SBS03	74	10	11,5
NACE sections A–S (form the survey coverage)	17	9	9
Merged sections – macro sections, as for quarterly release and working document: A; B–F; G; H; I; J; L–N, P–S	7	8	8

Under these conditions samples of about 2800 units were allocated (including 9 parts of the Oil Company). In each of the two samples, about 50% units were census units.

As described in the Introductory part of the paper, simple random sample is selected by a sequential scheme from each stratum. Sample selection is from the point 0.25.

SBS03 data are collected by mail or by web questionnaire. Enterprises that do not respond are contacted by phone or by email. The collected data are edited thoroughly. In case of some inconsistencies or major errors that could not be solved automatically or by a statistician, the enterprise is contacted in order to provide correct data.

In case of non-response, the data are imputed for about 20 large enterprises. For calculation of the preliminary results the missing data are imputed using data on the presented turnover from the Tax Administration database - data from the tax declaration on value added tax (PPPDV) of the current quarter. For the final estimates, the missing data are imputed using annual financial statements of the corresponding year. Of help are also data from the current and previous SBS03 surveys, as well as from the available annual financial statements.

The basic characteristics of the estimation procedure used in SBS03 are:

- The Horvitz-Thompson estimates, as for stratified simple random sample, are calculated for totals and their standard errors. The sampling weight is corrected for unit nonresponse. Also, units with outlier values for weighted turnover (about 15 units) are put in special census strata and at same time weights of units from initial strata are corrected.
- Transformations of census units (splitting, merging) are registered during the year. Data of enterprises participating in transformations of all quarters are harmonized with the state in the fourth quarter. These data are used for calculation of final estimates. Harmonized data for 2014 and 2015 (period 0 and period 1) has been used in the simulation study too.
- Changes of economic activity that are determined during data collection are utilized for domain estimation.
- Estimate of chain index (current quarter / previous quarter), as well as of annual index (current quarter / the same quarter of the previous year) are simply calculated as quotients of corresponding estimates of totals.
- In order to estimate the standard error of an index, Taylor linearization is used.
- Standard error of an index for total estimates of the same year is calculated using the overlapping part of the realized samples of the two quarters.
- Standard error of an index for total estimates of successive years is estimated using estimates of totals, their variances and of correlation that is computed on the common part of the realized samples of the two quarters in question (with stratification from the current year). The correlation estimate is corrected by a factor. The appropriate choice of this factor is the subject of this paper.

7. Results of the simulation study

The Quarterly Structural Business Survey sampling frames from two successive years, 2014 and 2015, were used in the simulation study. Original frames were updated by adding to frame units' auxiliary variable the values on turnover and number of employees from the annual financial statements of the corresponding year, if they existed. For imputing the missing turnover for 2014 or 2015, the following formula was used:

$$y_{j,t,imp} = y_{t-2} \cdot \frac{\sum_i y_{i,t}}{\sum_i y_{i,t-2}} \qquad (16)$$

where $y_{j,t,imp}$ is the imputated turnover value for unit j and year t. By construction, units of both frames have values for turnover y_{t-2} for year $t-2$. Ratio on the right side of the formula (16) is first calculated for all units i that have the value of turnover for the year t (t =2014, 2015) and belong to the same NACE class as the unit for which the imputed value needs to be calculated. If there are at least 5 such units, the value is imputed; otherwise, the ratio in formula (16) is computed for a higher NACE level. Depending on the number of units with known turnover for year t, imputation is conducted or not. Finally, for a unit with a missing value that belongs to a division with less than 5 units with none missing values, the ratio in formula (17) is calculated on the section level.

We assume that the results and conclusions regarding estimation of correlation that are brought according to the simulation study can be applied in the estimation of the precision of the annual index of operating income, turnover and operating expenses in SBS03 survey.

Table 2 refers to the basic sampling frame and sample characteristics for 2014 and 2015. It also includes the average values from simulations on number of units in the overlapping parts of samples and in the union of samples. These results are given by domains of estimation, sections or aggregated sections of NACE Rev. 2. For the common part of samples, domains are defined according to the economic activity in 2015, and for the rest of the units according to the activity of the corresponding year.

Table 2. Basic sampling frame and sample characteristics, for 2014 and 2015, and average number of units in the overlapping part of samples and in their union, from simulations

	Number of units				Average number of units from 10,000 simulations	
	Frame 2014	**Frame 2015**	**Sample 2014**	**Sample 2015**	**Overlapping part of samples**	**Union of samples**
TOTAL	**27351**	**28245**	**2797**	**2836**	**2015**	**3618**
Agriculture, forestry and fishing	1096	1096	116	118	84	151
Industry and construction[2]	8805	8691	1110	1137	841	1405
Services	17450	18458	1575	1583	1091	2061
Trade	7326	7926	293	294	205	375
Transportation and storage	15195	1530	106	100	81	125
Accommodation and food service activities	926	1054	104	116	69	150
Information and communications	1366	1383	207	200	147	262
Other services[3]	6313	6565	866	874	589	1150

[2] Covers: mining and quarrying; manufacturing; electricity, gas, steam and conditioning supply; water supply, sewerage, waste management and remediation activities and construction

[3] Covers: real estate activities; professional, scientific and technical activities; administrative and support service activities; education, human health and social work activities; arts, entertainment and recreation and other service activities.

Tables 3, 4 and 5 contain the results from 10,000 simulations on approximate true correlation and for the estimator of correlation, its mean value, bias and mean square error by different methods and sectors or aggregated sectors.

Table 3. Approximate true value of correlation and the mean value of the estimator of correlation

	Approx. true correlation	Mean value of the correlation estimator from simulations, by method							
		Overlapping part of samples	M1	M2	M3	M4	M5	M6	M7
TOTAL	**0.463**	**0.922**	**0.632**	**0.659**	**0.563**	**0.513**	**0.659**	**0.582**	**0.422**
Agriculture, forestry and fishing	0.548	0.896	0.631	0.637	0.601	0.496	0.641	0.985	0.442
Industry and construction [2]	0.371	0.949	0.664	0.710	0.587	0.568	0.711	0.585	0.415
Services	0.474	0.915	0.620	0.633	0.551	0.484	0.632	0.570	0.411
Trade	0.472	0.920	0.623	0.653	0.552	0.504	0.647	0.540	0.425
Transportation and storage	0.540	0.956	0.648	0.755	0.601	0.624	0.754	0.989	0.494
Accommodation and food service activities	0.295	0.847	0.567	0.537	0.469	0.390	0.531	0.994	0.357
Information and communications	0.459	0.886	0.613	0.636	0.556	0.497	0.639	0.996	0.431
Other services [3]	0.301	0.622	0.454	0.421	0.373	0.318	0.421	0.075	0.359

Table 4. Bias of the of the estimator of correlation

	Bias for the correlation estimator computed from simulations, by method							
	Overlapping part of samples	M1	M2	M3	M4	M5	M6	M7
TOTAL	**0.459**	**0.169**	**0.197**	**0.100**	**0.050**	**0.197**	**0.119**	**-0.041**
Agriculture, forestry and fishing	0.348	0.082	0.088	0.053	-0.052	0.093	0.437	-0.106
Industry and construction [2]	0.578	0.293	0.339	0.216	0.197	0.340	0.214	0.044
Services	0.441	0.146	0.160	0.077	0.010	0.159	0.097	-0.062
Trade	0.448	0.151	0.180	0.080	0.031	0.174	0.067	-0.047
Transportation and storage	0.415	0.108	0.215	0.060	0.084	0.214	0.448	-0.047
Accommodation and food service activities	0.552	0.272	0.242	0.174	0.095	0.236	0.699	0.062
Information and communications	0.428	0.154	0.178	0.098	0.039	0.180	0.537	-0.028
Other services [3]	0.321	0.153	0.120	0.072	0.017	0.120	-0.226	0.058

Table 5. Mean square error of the estimator of correlation

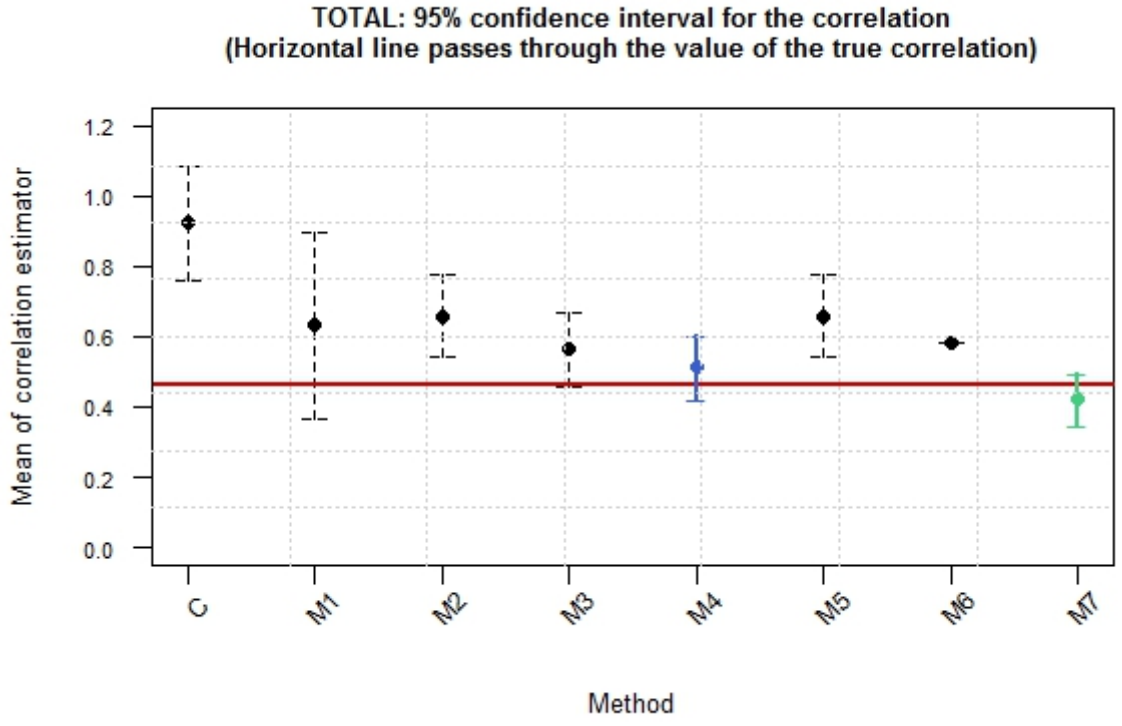
	Mean square error for the correlation estimator, computed from simulations, by method							
	Overlapping part of samples	M1	M2	M3	M4	M5	M6	M7
TOTAL	**0.217**	**0.047**	**0.042**	**0.013**	**0.005**	**0.042**	**0.014**	**0.003**
Agriculture, forestry and fishing	0.125	0.025	0.011	0.007	0.005	0.011	0.191	0.012
Industry and construction [2]	0.335	0.103	0.116	0.048	0.039	0.116	0.046	0.004
Services	0.205	0.046	0.030	0.010	0.003	0.030	0.009	0.005
Trade	0.212	0.049	0.038	0.012	0.005	0.036	0.005	0.004
Transportation and storage	0.175	0.042	0.048	0.010	0.009	0.048	0.201	0.004
Accommodation and food service activities	0.337	0.106	0.072	0.043	0.017	0.069	0.489	0.005
Information and communications	0.193	0.045	0.037	0.015	0.005	0.038	0.288	0.002
Other services [3]	0.120	0.039	0.022	0.011	0.005	0.022	0.051	0.005

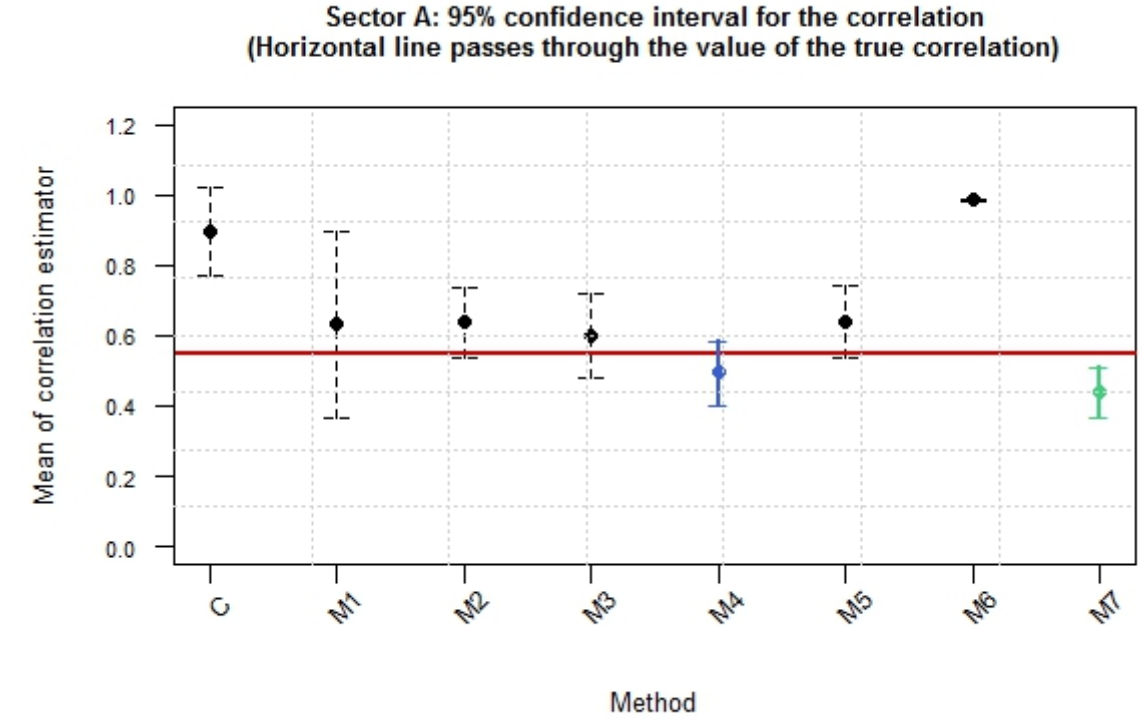
It is clear from Table 3 that the estimator of correlation which is based on the overlapping part of samples is positively biased (over estimates the true correlation) and that it cannot be used without previous correction. The un-weighted correlation on the overlapping part of samples is also not satisfactory. In some of the domains, it overestimates the true correlation even more than in the common correlation.

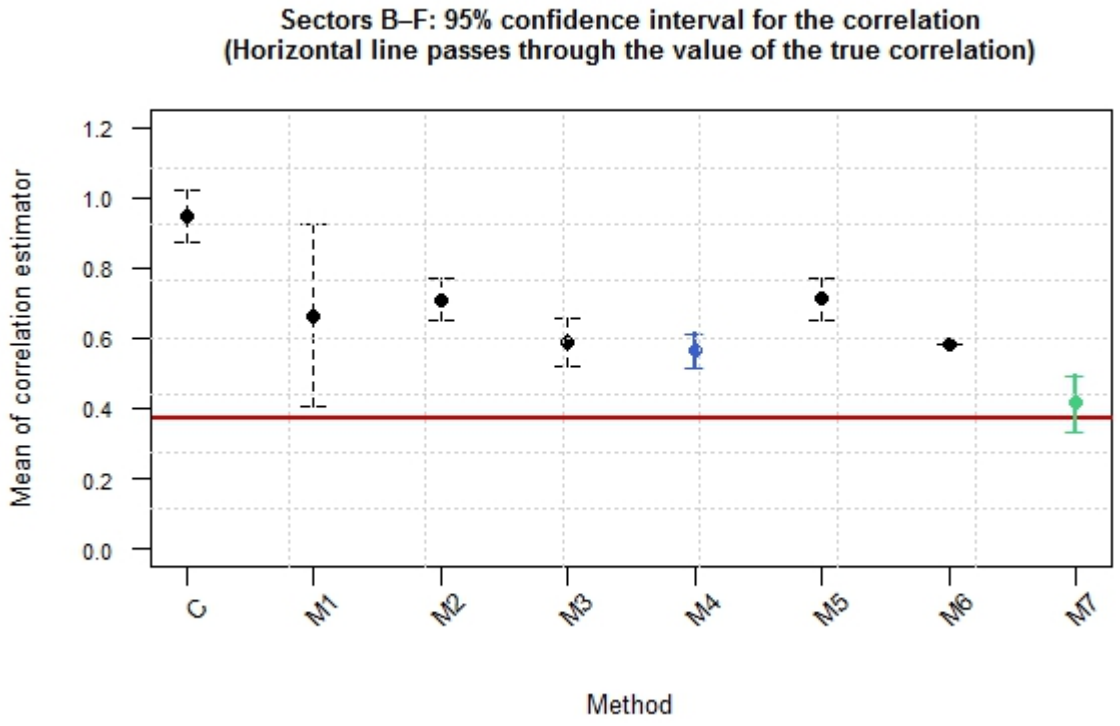
The simulation results for the methods M1-M5, which are based on the reduction of the correlation between two estimated totals calculated from the overlapping part of samples, show that the methods M3 and M4 outperform the others in the sense that these two estimators have smaller bias and standard error than the remaining methods. The method M4 gave slightly better results than the method M3 and correction factor is easy to calculate: the ratio of units sampled both times and the union of sampled units.

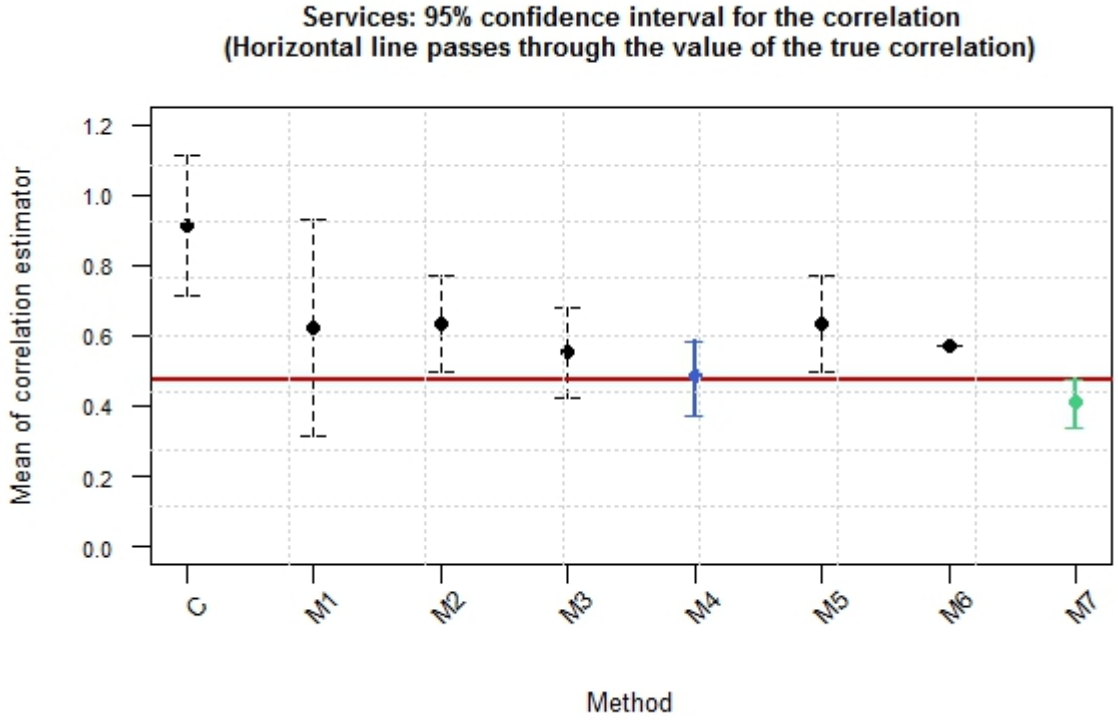
Using the formula M7, coefficient of variation is estimated as a linear combination of terms which depend only on the coefficients of variation of the estimators of totals at time 0 and time 1 (years 2014 and 2015). The parameter α was defined as $\alpha = 0.5 \cdot c_4$, where c_4 is the correction factor of the method M4. Coefficients of variations are easily estimated once the totals and their variances are estimated. According to simulations, this method gave a precise estimator with a small bias, in most cases negative. In spite of these good characteristics, further studies are needed to confirm its validity. The parameter α was determined by trials and for now there is no good argument for this choice.

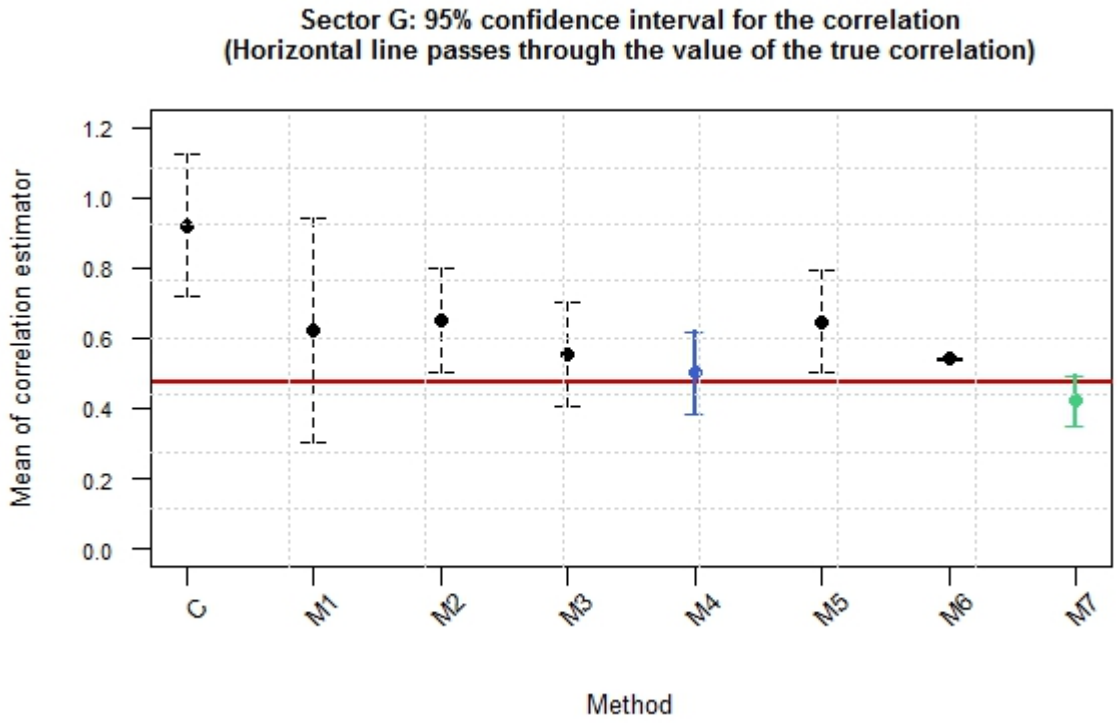
The plots below visualize the results concerning the estimation of temporal correlation. For the methods M4 and M7, the mean values of estimators of correlation and corresponding 95% confidence intervals for correlation are presented on the graphs in blue and green colour, respectively. On the plots, method ‘C’ denotes the mean of the correlation based on the overlapping part of samples.

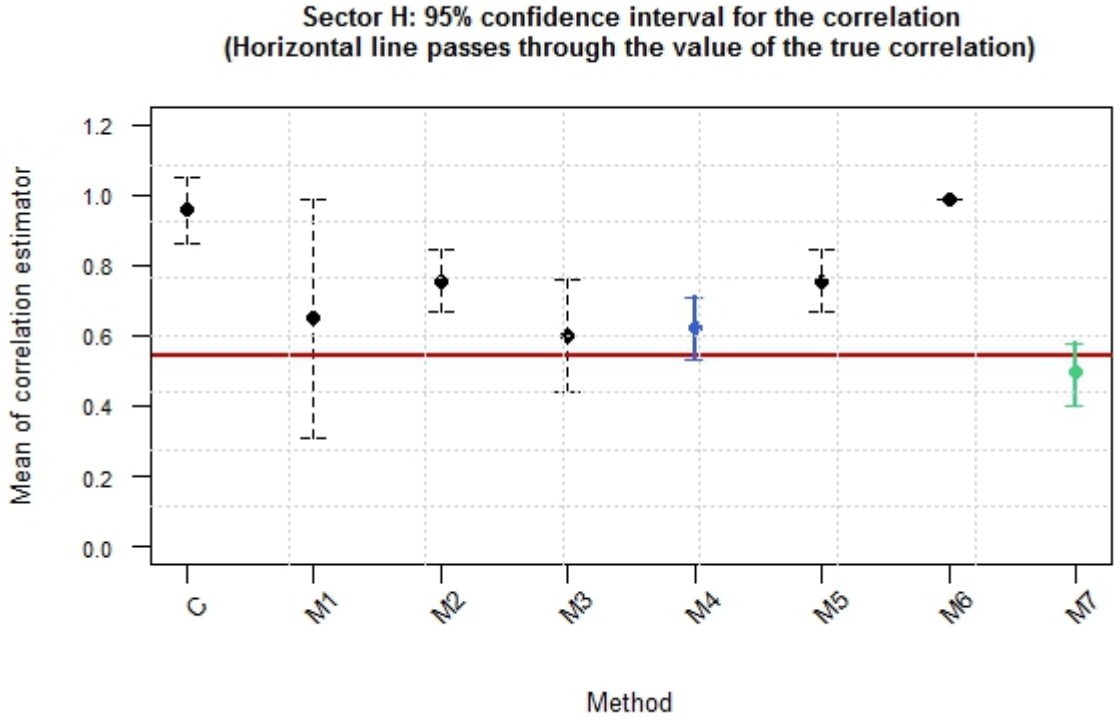


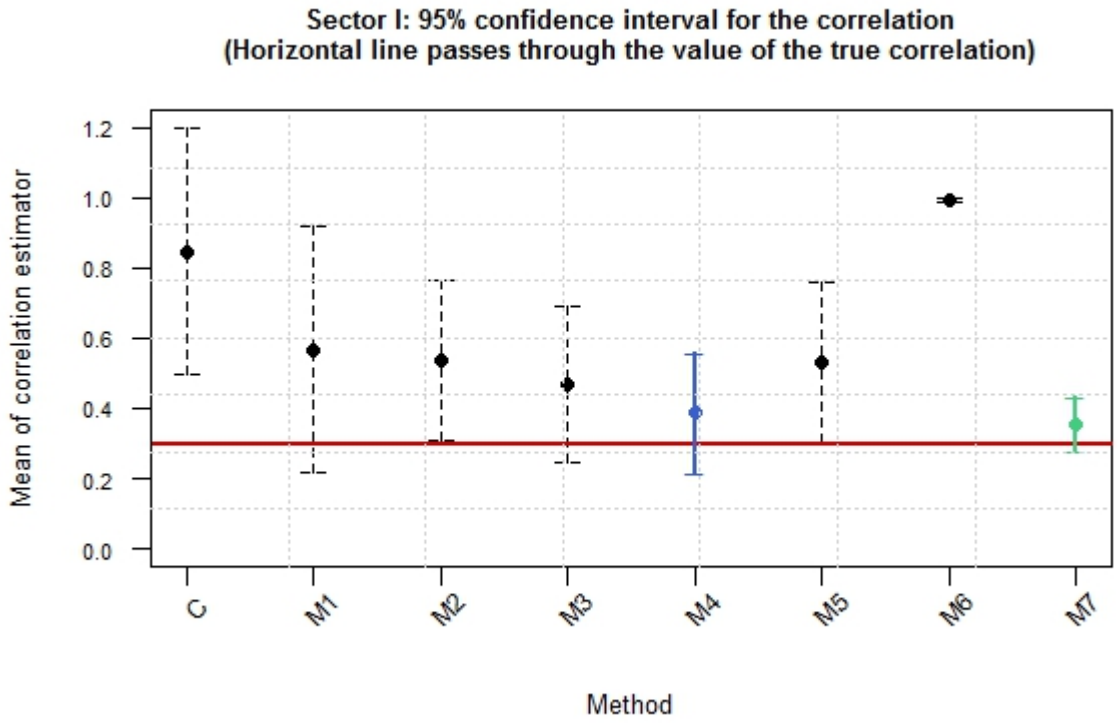

Sector A: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method

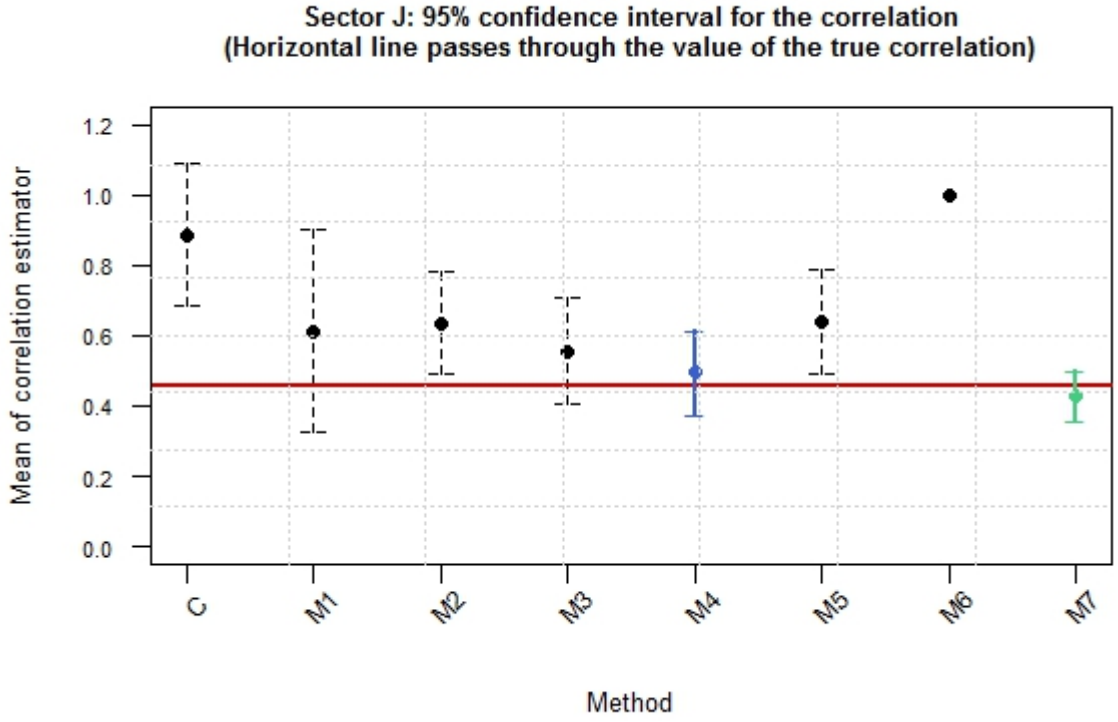

Sectors B–F: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method

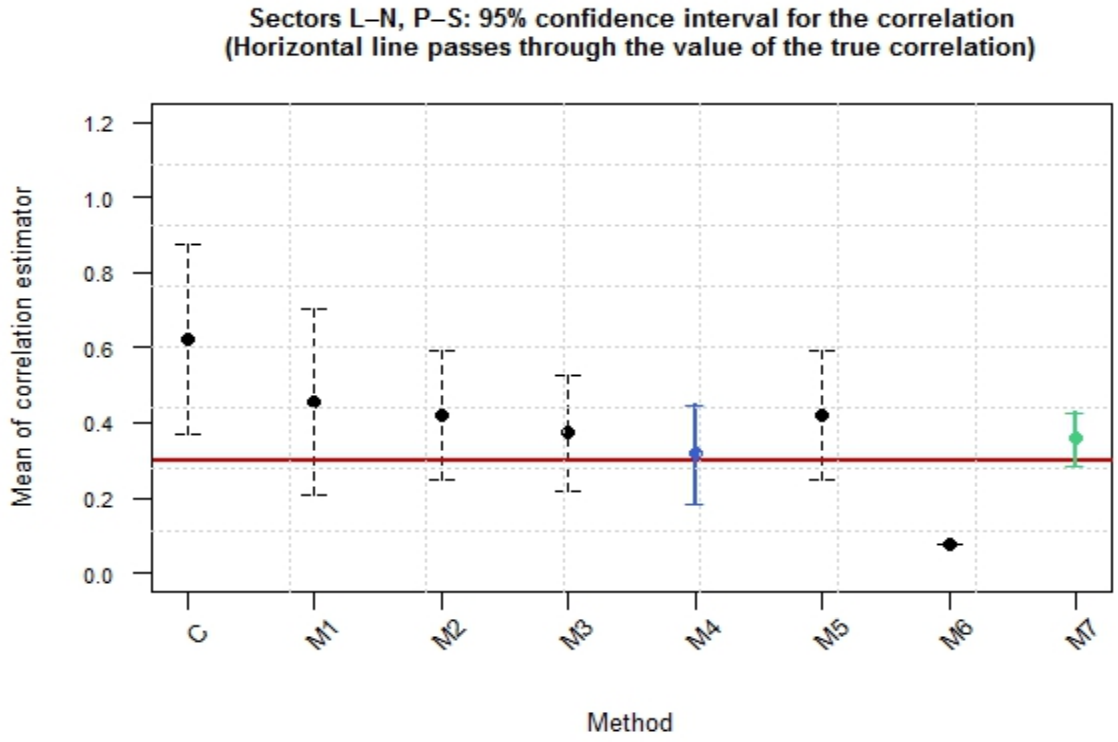

Services: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method


Sector G: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method


Sector H: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method


Sector I: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method


Sector J: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method


Sectors L–N, P–S: 95% confidence interval for the correlation
(Horizontal line passes through the value of the true correlation)
Mean of correlation estimator
1.2
1.0
0.8
0.6
0.4
0.2
0.0
C
M1
M2
M3
M4
M5
M6
M7
Method

8. Concluding remarks

The problem of precision estimation of an index is encountered in short term statistics. It is a complex problem due to changing business population and to the system of coordination built in the method of sample selection.

In a simulation study, a solution to this problem has been sought for annual indices of operating income, turnover and operating expenses of the Quarterly Structural Business Survey. As other business samples in the Statistical Office of Republic of Serbia, successive samples of this survey are coordinated with permanent random numbers and they overlap in a random number of units.

The simulation study has been conducted using the sampling frames of the Quarterly Structural Business Survey for the years 2014 and 2015, and calculation were performed for the turnover. Because operating income and expenses are highly related with turnover, the method adopted for the estimation of the variance of an index on turnover, could be applied to them too.

Since the samples are partly overlapping, the variance expression includes the temporal correlation of the estimators of totals of successive samples. The correlation on the common part of samples overestimates the true correlation because the remaining units of the two samples are not considered.

The idea of this study was to start from the estimator of correlation on the common part of samples and to seek for a correction factor that would trim the initial estimator and provide a satisfactory estimator of the true correlation.

Several factors were proposed and the behaviour of the output correlation estimators was checked in the simulation study.

The method M4, with correction factor defined as the ratio of the number of the common units to the number of units in the union of samples, is recommended to be used for precision estimation of indices of turnover, operating income and expenses by sections or aggregated sections as published in the quarterly release of the Quarterly Structural Business Survey. Compared to other investigated methods, the simulation results for this estimator gave the smallest bias and the variance. The bias was positive and high only for the group of sections B-F, but for this domain other estimators showed similar (method M3) or even worse characteristics.

During the simulation study, an idea for a different approach to precision estimation of an index has emerged (method M7). Instead of basing the estimation procedure on the correlation estimator on the overlapping part of samples, the estimator of the coefficient variation of an index was defined as a linear combination of limits for the approximate coefficient of variation (formula (6))

$$\widehat{CV}_7(\hat{r}) = \alpha\left|\widehat{CV}(\hat{t}_0) - \widehat{CV}(\hat{t}_1)\right| + (1-\alpha)\sqrt{\widehat{CV}^2(\hat{t}_0) + \widehat{CV}^2(\hat{t}_1)}$$

with $0 \leq \alpha \leq 1$. Here the problem is to search for an adequate value of the parameter α. By trial, in this simulation study, α was chosen to be $0.5 \cdot c_4$, where c_4 is the correction factor for the method M4. The simulation results show the corresponding correlation estimator had the smallest variance and a small negative bias for all domains except for B–F; I and L–M, P–S. However, in these two domains, the true value of correlation belonged to the confidence interval. In spite of the good simulation results, this method needs to be further explored before it could be recommended for use.

It would be appreciated that at the time of final estimation, simulation studies are repeated in order to check whether the obtained results and derived conclusions are stable. Perhaps some new ideas will come up or a proof of the good performance of the method M7.

Finally, the purpose of this paper has been to systematically present the results of the work devoted to this subject. The work if founded on ideas and method of simulation that have been proposed by Swedish experts, A. Lindblom and S. Berg, without which this problem would have been set aside.

References

Berger, Y. (2004). “Variance Estimation for Measures of Change in Probability Sampling”. *The Canadian Journal of Statistics,* Volume 32, Issue 4, pp. 451–467. Canada: Wiley, Statistical Society of Canada.

Bethel, J. (1989). “Sample allocation in multivariate surveys”. *Survey methodology* Volume 15, No. 1, pp. 47-57. Canada: Statistic Canada.

Full, S, Lewis, *D.* (2004). “Estimating Sampling Errors for Movements in Business Surveys”. *Seminar on economic statistics*, Session 11: Sampling and Variance Estimation II. Sweden: Statistics Sweden.

Hidiroglou, M.A. (1986). “The Construction of Self-Representing Stratum of Large Units in Survey Design”. *American Statistician*, Volume 40, No. 1, pp. 27-31. United States: Taylor & Francis.

Hidiroglou, M., Sarndal, C.-E, Binder, D. (1995). “Weighting and Estimation in Business Surveys”, *In Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P.S. Scott, pp. 477–502. United States: John Wiley & Sons.

Laniel, N. (1988). “Variances for a Rotating Sample from a Changing Population.” In *Proceedings of the Business and Economic Statistics Section: American Statistical Association*. pp. 246-250. United States: Government publication.

Statistics Sweden (2003) S*AMU – The System for Coordination of Frame Populations and Samples from the Business Register at Statistics Sweden*. Background Facts on Economic Statistics. Sweden: Statistics Sweden Available at:

http://www.scb.se/statistik/OV/AA9999/2003M00/X100ST0303.pdf.

Lindblom, *A.* (2014). “On Precision in Estimates of Change over Time where Samples are Positively Coordinated by Permanent Random Numbers”. *Journal of Official Statistics*, Volume 30, No. 4, pp. 773–785. Sweden: Statistics Sweden. Available at:

http://dx.doi.org/10.2478/JOS-2014-0047 (2014).

Melovski-Trpinac, O. Ninić, M, Panović, M. (2014). *Sample coordination of statistical business surveys*, Working paper of Statistical office of the Republic of Serbia No. 89, Year L. Belgrade, SORS. Available at:

http://pod2.stat.gov.rs/ObjavljenePublikacije/G2014/pdfE/G201410089.pdf).

Mission reports of the Statistics Sweden International Consulting Office:

Berg, S. and Lindblom, A. SERSTAT 2013:22;

Lindblom, A. and Berg, S. SERSTAT 2014 07;

Lindblom, A. SERSTAT 2016:05

Nordberg, L. (2000). “On Variance Estimation for Measures of Change when Samples are Coordinated by the Use of Permanent Random Numbers”. *Journal of Official Statistics,* Vol. 16. No. 4, pp.363-378. Sweden: Statistics Sweden.

Petković, Đ. (2015). “Estimation of standard error of the parameter of change using simulations”. *Romanian Statistical Review* nr. 2 , pp. 90-95. Romania: National Institute of Statistics.

Särndal C-E., Swensson B. & Wretman J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Замаклар, Г., Меловски Трпинац, О. и Брашанац, Б. (2015). *Резултати статистичког истраживања СБС-03*. Радни документ Републичког завода за статистику, год. LI, бр. 91. Београд: РЗС. Available at:

http://www.stat.gov.rs/WebSite/repository/documents/00/01/90/95/Rd-91-kvartalno_poslovanje_privrednih_drustava_2014.pdf

Замаклар, Г. и Меловски Трпинац, О. (2016). *Резултати статистичког истраживања СБС-03, 2015*. Радни документ Републичког завода за статистику, год. LII. бр. 97. Република Србија: РЗС. Available at:

http://www.stat.gov.rs/WebSite/repository/documents/00/02/30/93/RD_97_SBS03_2015_31102016.pdf

Dissemination and public relations unit
Phone: +38111 2401284
Email: stat@stat.gov.rs

Library
Phone: +38111 2412922, ext. 251
Email: biblioteka@stat.gov.rs

Number of pages: 22