Monthly estimates for the Labour Force Survey: an experimental study

Alessio Guandalini[1], Silvia Loriga[1], Mirjana Ogrizovi-Brasanac[2]

Abstract

The Labour Force Survey (LFS), conducted by the Statistical Office of the Republic of Serbia (SORS), provides quarterly labour market figures. The applied sample design allows the calculation of monthly estimates using a longitudinal sample structure. This paper presents the methodology for calculating monthly estimates of the main labour market indicators that comply with Eurostat requirements. The procedure for producing the monthly estimates is illustrated using data for the 4th quarter of 2019.

Keywords: Calibration estimator (CAL estimator), longitudinal constraints, Regression Composite Estimator, Adjusted Regression Estimator, sampling variance.

DOI: https://doi.org/10.1481/ISTATRIVISTASTATISTICAUFFICIALE_1-2-3.2025.01.

1 Alessio Guandalini (alessio.guandalini@istat.it); Silvia Loriga (siloriga@istat.it), Italian National Institute of Statistics - Istat.

2 Mirjana Ogrizovi-Brasanac (m.ogrizovic-brasanac@stat.gov.rs), Statistical Office of the Republic of Serbia - SORS.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

Timely labour market information is essential to the national economy. It enables critical analysis of the workforce, including its behaviour, supply and demand, structural changes, and the impact of policies on it. In the European Union, official labour market information is collected through the Labour Force Survey (LFS). The LFS is based on Regulation (EU) 2019/1700[3] of the European Parliament and Council. The Regulation establishes a common framework for European statistics relating to individuals and households. It also sets out the survey's main methodological requirements, content, and operational constraints. The main goals are the harmonisation of the LFS across the European Union, the improvement of data quality, and the comparability of results.

The survey is designed as a quarterly sample survey. The requirements for the sampling design, the estimation process, and the microdata delivery apply to quarterly data. The detailed information on participation in the labour market, collected through a very large set of variables, is reliable when analysed on a quarterly and annual basis.

For short-term analytical needs, Regulation (EU) 2019/2241 requires European countries to submit monthly unemployment data to Eurostat, limited to a few indicators. The deadline for data transmission falls between 25 and 27 days after the calendar month. To submit the data within the required period, a two-stage procedure is typically required. Provisional monthly estimates are computed from a partial sample for each month of the quarter. Final monthly estimates are computed from the full monthly sample, once data for the entire quarter are collected and quarterly estimates are computed. Final estimates differ slightly from provisional ones and are calculated to be consistent with quarterly estimates.

The Serbian LFS fulfils the European requirements for calculating quarterly estimates, and the adopted sampling design also enables the estimation of monthly labour market indicators. In this study, the procedure used to compute the monthly estimates is tested on Serbian LFS data for the 4th quarter of 2019.

3 Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019, amending Regulations (EC) No 808/2004, (EC) No 452/2008 and (EC) No 1338/2008 of the European Parliament and of the Council, and repealing Regulation (EC) No 1177/2003 of the European Parliament and of the Council and Council Regulation (EC) No 577/98.

The steps are as follows: computation of sampling weights for provisional and final monthly estimates; estimation of monthly figures for the main labour market indicators (employed, unemployed, and inactive); calculation of standard errors for these estimates; and implementation of the statistical procedures in R.

The paper is organised as follows: Section 2 provides an overview of the Serbian LFS; Section 3 addresses the production of provisional estimates; Section 4 examines variance estimation for the adopted estimator; Section 5 is dedicated to consistency with the quarterly estimates and the production of final monthly estimates. Each of Sections 3-5 contains three subsections: methodological aspects, application to the Serbian LFS, and focus on the R procedures. Finally, Section 6 discusses the main results.

2. Sample design of the Serbia LFS 2019

The Serbian LFS is a quarterly household sample survey. The target population comprises all private (non-institutional) households and persons in the territory of the Republic of Serbia, excluding the region of Kosovo and Metohija, which constitutes the usual population. Persons in collective households are excluded. The survey population is restricted to households living in the 2011 Census enumeration areas that contained at least 20 households at the time of the 2011 Census. The survey population is 1.5% smaller than the target population.

The sampling design is a two-stage stratified sample: enumeration areas (EAs, i.e., clusters of households) represent the primary sampling units (PSUs), and households are the final sampling units (SSUs). EAs are stratified by settlement type (urban/other)[4] and by territory at the NUTS 3 level (25 areas).

A sample of enumeration areas is selected at the first stage, with a probability proportional to the number of persons aged 15 and over. In each selected EA, a sample of 10 households is selected using simple random sampling.

A 2-(2)-2 rotation scheme is adopted for the sample of households. Each household participates in the survey four times over an 18-month period; i.e., each household is in the sample for two consecutive quarters, then out of the sample for the next two quarters, and back in the sample for two further consecutive quarters.

Each quarter includes four independent subsamples (rotating groups): one new rotating group (households interviewed for the first time), two rotating groups from the previous quarter (households receiving the second and fourth interviews), and one rotating group from three quarters earlier (households interviewed for the third time).

Under this rotation scheme, the expected overlap of the sample is 50% between two consecutive quarters and between the same quarter in two consecutive years. This improves the efficiency of estimating changes relative to the previous quarter and to the same quarter of the previous year.

4 Official statistics in Serbia do not include a specific definition of rural settlements. Instead, an "administrative-legal" criterion is applied that designates settlements as either "Urban" or "Other". Urban settlements are recognised as such by an act of the local self-government, with all other settlements falling into the category of "Other".

Furthermore, data from previous rounds can inform current estimates.

Each subsample (rotation group) comprises 4,810 households across 481 enumeration areas. In each enumeration area, 10 households are selected.

The continuous conduct of the survey requires the sample distribution over time. Each subsample, consisting of one rotating group, is uniformly and randomly distributed across the 13 weeks of the quarter[5].

Data are collected using a mixed-mode design combining Computer-Assisted Personal Interviewing (CAPI) and Computer-Assisted Telephone Interviewing (CATI): the first interview is always conducted by CAPI, and subsequent interviews are mainly conducted by CATI, except for households not found in previous rounds (phone numbers were not collected, etc.), which are interviewed by CAPI.

Data collection is fast. The time period to which the information on labour market participation refers (the reference week) is randomly assigned to each household. After each reference week, interviewers have two weeks to collect data. Owing to the use of computer-assisted techniques and the speed of data collection, the data are affected by few errors and are available in a timely manner for treatment.

The quarterly estimates are calculated using weights obtained through calibration (Deville and Särndal 1992). In the calibration procedure, design weights (for each stratum, the weight is the reciprocal of the product of the inclusion probabilities at each stage) are first corrected for non-response. The data are based on current demographic estimates: the distribution of the population by gender (two groups) and five-year age classes (14 groups) at the NUTS 3 level (25 groups), and the distribution of households by the number of household members (six groups) at the NUTS 3 level. The ReGenesees package (Zardetto 2015) in R is used to compute the calibrated weights. The logit function is used as the distance function. The same final weight is assigned to each individual in the household to ensure consistent estimates for both households and individuals.

5 When the year is composed of 53 weeks instead of 52, the sample of one quarter (first or fourth) is distributed over 14 weeks. This occurs once every four years.

3. Provisional estimates

To ensure data are published no later than 30 days after the end of the observed month, it is customary to publish provisional monthly estimates first. Thirty days after the end of the corresponding quarter, these provisional monthly estimates are usually replaced by final monthly estimates that are consistent with the quarterly estimates.

The data collection phase for the Serbian LFS usually ends two weeks after the end of the month, and the data are available for processing five days later. Almost all units involved in each reference month are interviewed and can be included in the calculation of the provisional monthly estimates.

The dataset used to produce provisional monthly estimates must include the variables (without missing values) listed in Table 3.1.

Table 3.1 - List and features of the variables in the input file for computing monthly estimates

VARIABLE NAME	Description
KEY_INDIV	Key variable for the individual
KEY_HH	Key variable for the household
KEY_PSU	Key variable for the Enumeration Area (EA)
STRATUM	Stratum variable (NUTS2 * urban/other * rotation group)
SEX	Sex
AGE	Age
NUTS2	Region code
NUTS3	Province code
ROTGR	Rotation group
EMP	Occupational status: employed (1=yes; 0=no)
UNEMP	Occupational status: unemployed (1=yes; 0=no)
INAC	Occupational status: inactive (1=yes; 0=no)

Source: Authors' processing

3.1 Design weights

In the LFS sample, selected households from each stratum are randomly assigned to the 13 weeks of the quarter. Because of this sampling feature, the households surveyed in a given month can be treated as a random sample and used to produce reliable monthly estimates.

The computation of the monthly sample design weight (dk) is the starting point in the computation of the weights. The design weight for the household k,

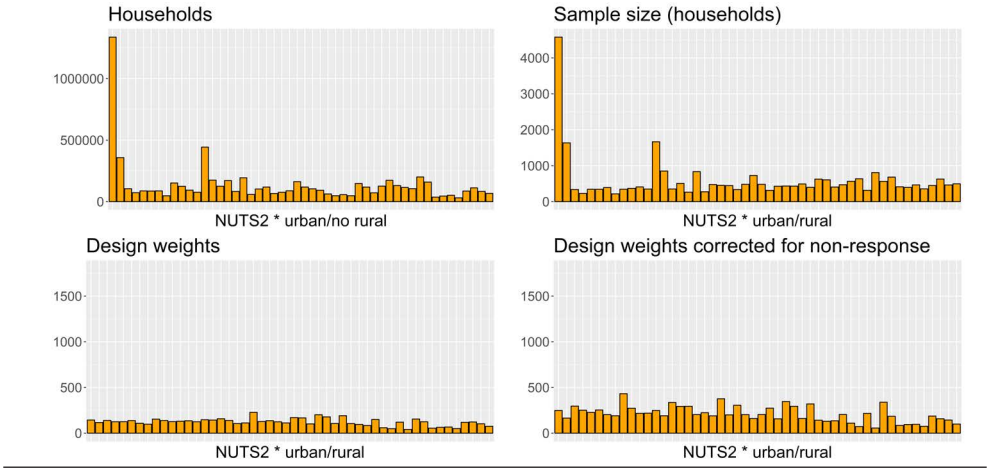
which is also shared by all individuals in the household, is equal to the inverse of the first-order inclusion probability, π_k. The first-order inclusion probability is computed at the NUTS 3 domain level for the type of settlement (urban/other), and is equal to the ratio of the planned number of households to be surveyed in the domain to the overall number of households in the population for the domain. This procedure approximates the real selection process for the monthly sample and is crucial for preventing extreme weights and enabling their calibration.

The design weight has to be adjusted for non-response among households. Non-response correction is applied at the domain level by multiplying the design weight by the inverse of the observed non-response rate in each domain. More complex weighting adjustments could be evaluated depending on the availability of auxiliary variables related to non-response (Bethlehem et al. 2011; Särndal and Lundström 2005; Little and Rubin 2002; Groves and Couper 1998).

The design weight corrected for non-response ($d_{k,nrc}$) represents the starting point for the monthly estimates and will be used in the work.

Figure 3.1 shows that the distribution of households in the sample is approximately equal to that in the domain. Looking at the values $d_{k,nrc}$ of the weights, it can be seen that in some domains their value after the correction for the non-response of households is significantly higher than the weights of the plan.

Figure 3.1 - Number of households, planned sample size, design weights, and design weights corrected for non-response at domain level (NUTS2 * urban/no urban). October 2019



Source: Authors' processing of the Serbian Labour Force Survey data

3.2 The Adjusted-Regression and the Regression Composite Estimators

3.2.1 Theoretical bases

The estimator proposed in the literature for producing monthly estimates is the regression composite (RC) estimator (Singh 1994 and 1996; Fuller 1999; Singh et al. 2001). The RC estimator incorporates auxiliary information from previous survey waves, making the estimates more stable. For this estimator, auxiliary information from previous survey waves is required for each unit in the sample. For the part of the sample for which these pieces of information are not available, data imputation should be performed. In the Canadian LFS, the RC estimator is used because there is a large overlap between the household samples for two successive months (5/6 of the sample overlaps), so imputation is needed for only 1/6 of households.

In the Serbian LFS, under the 2-(2)-2rotation scheme, the theoretical overlap between the samples referring to months m and m-3 is ½, and the same is for the months m and m-12. When applying the RC estimator, it is necessary to impute half of the sample that does not overlap for months m-3 and m-12. Attrition and non-response reduce the planned overlap rate. In the 4th quarter of 2019, the overlap rate was 36% with the 3rd quarter of 2019, and 30% with the 4th quarter of 2018. The use of the RC estimator would introduce too much variability because it would require data imputation for 64-70% of the sample, respectively. Therefore, this estimator cannot be used, and a different estimator is recommended, such as the Adjusted-Regression (AR) estimator (Renssen and Nieuwenbroek 1997).

The AR estimator is known in the literature also as the Extended regression (E) estimator (Ballin et al. 2000; Rancourt 2001; Berger et al. 2009). For practical purposes and ease of use, its "calibrated version", named AC estimator, has been defined (see, e.g., Särndal and Traat 2011; Ceccarelli and Guandalini 2012; Guandalini 2014). An AC estimator is used in this study to calculate the Serbian LFS monthly estimates.

The AC estimator lies on the calibration estimator (Deville and Särndal 1992), the most commonly used estimator in official statistics. The calibration estimator (CAL) may be written as follows

$$\hat{t}_{y_{CAL}} = \sum_{k \in s} w_k y_k \qquad (1)$$

and it represents a way to include auxiliary information in the estimation process based on a random sample (s) selected from a finite population. Through the calibration process, a system of weights, w_k, for the sample units is obtained, thereby making the sample consistent with the known distributions of structural variables in the reference population. This improves the efficiency and coherence of the estimates and can be used to produce all estimates. That is why it is widely used in official statistics (Särndal 2007).

The design weights or, the design weights corrected for non-response (as suggested in Deville et al. 1993), are modified so that estimates of totals for the P auxiliary variables ($x_1, x_2,\ldots, x_P$) collected on the sample, are equal to the known totals, $\underline{t}_x=(t_{x1}, t_{x2},\ldots, t_{xP})$. The totals, $t_{x1}, t_{x2},\ldots, t_{xP}$, are usually retrieved from registers or administrative sources and are, by definition, assumed to be error-free.

The calibrated weights, w_k, are obtained by solving the following optimisation problem

$$\begin{cases} min_{w_k}\left\{\sum_{k \in s} G(w_k, d_k)\right\} \\ \sum_{k \in s} w_k \underline{x}_k = \underline{t}_x \end{cases} \qquad (2)$$

where $\underline{x}_k$ is a vector of P auxiliary variables collected on the unit k. They are as close as possible to the d_k, an average sense, with respect to a pseudo-distance $G(\cdot)$, therefore, the CAL estimator keeps some good properties of the Horvitz-Thompson (HT) estimator, $\hat{t}_{y_{HT}} = \sum_{k \in s} d_k y_k$ (Horvitz and Thompson 1952), which uses the design weights d_k. Furthermore, by definition, the calibrated weights enable to exactly reproduce the totals of auxiliary variables known in the population.

Different pseudo-distances $G(\cdot)$ can be used (see, e.g., Deville and Särndal 1992; Singh and Mohl 1996). Each metric determines a different system of calibrated weights with different ranges. The most used metric in official statistics is the truncated logarithmic distance,

$$\left(\frac{w_k}{d_k} - L\right) ln\left(\frac{\frac{w_k}{d_k} - L}{1-L}\right) + \left(U - \frac{w_k}{d_k}\right) ln\left(\frac{U - \frac{w_k}{d_k}}{U-1}\right)$$

because it prevents negative and extreme weights. The ratio w_k/d_k is defined between L (lower bound) and U (upper bound), and the bounds are usually set by the researcher.

When the Chi-Squared pseudo-distance is used, the CAL estimator is equivalent to the Generalised REGression (GREG) estimator (Bethlehem and Keller 1987; Cassel et al. 1976; Isaki and Fuller 1982; Särndal 1980; Särndal et al. 1989; Wright 1983)

$$\hat{t}_{y_{GREG}} = \hat{t}_{y_{HT}} + \left(\underline{t}_x - \underline{\hat{t}}_{xHT}\right)^T \underline{\hat{\beta}},$$

where the HT estimate is corrected by the difference between the vector of known totals and the vector of the totals of the same variables estimated on the sample with the HT estimator, weighted $\underline{\hat{t}}_{xHT} = (\hat{t}_{x1_{HT}}, \hat{t}_{x2_{HT}}, \dots, \hat{t}_{x_{PHT}})$ by the vector of the regression coefficient $\underline{\hat{\beta}} = (\hat{\beta}_{x_1}, \hat{\beta}_{x_2}, \dots, \hat{\beta}_{x_P})^t$.

As long as the sample size and the population size are large, the CAL estimator is equivalent to the GREG estimator even when other pseudo-distances are used. This result is important because the sampling variance of CAL estimator is automatically equivalent to the sampling variance of GREG estimator:

$$AV\left(\hat{t}_{y_{CAL}}\right) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\hat{e}_k w_k)(\hat{e}_l w_l) \cong \hat{V}\left(\hat{t}_{y_{GREG}}\right) \tag{3}$$

where π_{kl} are the second order inclusion probabilities and $\hat{e}$ are the estimated residuals of the superpopulation model, $\xi: y = \underline{X}\underline{\beta} + e$. This model is implicitly assumed by the calibration estimator, therefore, it is possible to write the CAL estimator as a GREG estimator.

Because of the auxiliary totals are not all available from registers or administrative sources, survey-estimated totals can also be considered. In the case of surveys with rotating schemes with partial overlapping samples, their use it is even advocated. It enables to include in the estimation auxiliary variables collected on the same units on a different time. These variables are

usually highly correlated with the target variables and therefore, they help in increasing the efficiency of the estimates or at least in making them more stable and less volatile.

The expression of the CAL estimator in (1) holds, even if in this case Särndal and Traat (2011) suggest naming this estimator as AC estimator, pointing out the double nature of the considered auxiliary totals. The final weights can be obtained by solving the optimisation problem:

$$\begin{cases} min_{w_k}\left\{\sum_{k\in s} G(w_k, d_k)\right\} \\ \sum_{k\in s} w_k \underline{u}_k = \underline{\tilde{t}}_u \end{cases} \tag{4}$$

where $\underline{u}_k=\underline{x}_k,z_k$ is a vector of $P+M$ auxiliary variables in which the first P are $(x_1, x_2,\ldots, x_P)$ and the last M are $(z_1, z_2,\ldots, z_M)$. Totals of the x auxiliary variable, $\underline{t}_x=(t_{x1}, t_{x2},\ldots, t_{xP})$, are exactly known from registers or administrative data, while totals of the z auxiliary variables, $\underline{\tilde{t}}_z=(t_{z1}, t_{z2},\ldots, t_{zM})$, are estimated by the sample survey of another survey or previous waves of the same survey. The vector $\underline{\tilde{t}}_u=(\underline{t}_x, \underline{t}_z)$ consists of the $P+M$ components of which the last M are affected by sampling errors, while the first P are not.

The AC estimator has its equivalent GREG version, which is called the AR estimator

$$\hat{t}_{y_{AR}} = \hat{t}_{y_{HT}} + \left(\underline{\tilde{t}}_u - \underline{\tilde{t}}_{u_{HT}}\right)\underline{\hat{\beta}} \tag{5}$$

where $\hat{t}_{uHT}$ are the HT estimates in the sample of the $P+M$ auxiliary variables and

$$\underline{\hat{\beta}} = \left(\underline{\hat{\beta}}_x, \underline{\hat{\beta}}_z\right)^t = \left(\hat{\beta}_{x_1}, \hat{\beta}_{x_2}, \ldots, \hat{\beta}_{x_P}, \hat{\beta}_{z_1}, \hat{\beta}_{z_2}, \ldots, \hat{\beta}_{z_M}\right)^t$$

the vector of the regression coefficients in which the first P elements are related to the x auxiliary variables, while the last M are related to z auxiliary variables.

3.2.2. Application to the Serbian LFS

To produce monthly estimates that are efficient and stable, it is important to include in the estimation process a good set of auxiliary variables correlated with the target variables. This is crucial especially for monthly estimates because the sample size is ⅓ of the quarterly sample, then such auxiliary variables are fundamental to increase the accuracy.

The AC estimator (or equivalently the AR) enables to entail in the estimation auxiliary variables for which the totals can be derived both from a register or estimated on a survey. Three kinds of auxiliary variables have been considered for the present purpose: household, individual and longitudinal auxiliary variables. In particular:

1. rotation group[6];
2. number of components of the household[7];
3. sex (SEX, M=male, F=female);
4. age class (AGE_CLASS, 0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+);
5. age class 25-74 (0=no, 1=yes);
6. employed a quarter before (1=yes, 0=no) [emp_1];
7. unemployed a quarter before (1=yes, 0=no) [unemp_1];
8. inactive a quarter before (1=yes, 0=no) [inac_1];
9. employed four quarters before (1=yes, 0=no) [emp_4];
10. unemployed four quarters before (1=yes, 0=no) [unemp_4];
11. inactive four quarters before (1=yes, 0=no) [inac_4].

Auxiliary variables in 1 and 2 are collected for the individuals in the sample but refer to household characteristics. The solution proposed by Lemaître and Dufour (1987) is adopted. In particular, the same value is assigned to all individuals belonging to the same household. Furthermore, since the auxiliary variables are originally constructed as dummy variables (1 = yes, 0 = no), they are divided by the number of household members. In this way, it is possible to simultaneously consider both household- and individual-level auxiliary variables. The rationale behind the use of these variables is that they can help address non-response and attrition problems, since these issues are strongly related to household composition and the number of interviews conducted within the household.

Auxiliary variables 3, 4, 5 are characteristics of each individual; variable 5 is useful to introduce at least this auxiliary information on age 25-74 at

6 Four modalities: first rotation group [RG1], second rotation group [RG2], third rotation group [RG3], fourth rotation group [RG4].

7 Six modalities: household with one component [CLASS_COMP1], two components [CLASS_COMP2], three components [CLASS_COMP3], four components [CLASS_COMP4], five components [CLASS_COMP5], six or more components [CLASS_COMP6].

the NUTS 3 level. Age and sex are very important variables for explaining the occupational status. Moreover, for NSIs, it is important to disseminate estimates consistent with the main population distributions. For household and individual auxiliary, variables totals are considered known.

Auxiliary variables 6, 7, 8, 9, 10 and 11 are the so-called longitudinal auxiliary variables. That is, they are the occupational status observed on the same individuals three months before and twelve months before. These auxiliary variables are obtained via a simple linkage operation with the data collected in the previous quarter and four quarters before. For this purpose, just the individual identifier (KEY_INDIV) is sufficient.

Due to the rotating scheme, this information is available for almost ½ of the quarter, being ½ the theoretical overlapping fraction between the two following quarters and between the same quarter in two following years. As already said, this leans the choice towards the AC (or equivalently the AR) estimator instead of the RC estimator.

The totals of the longitudinal auxiliary variables are based on estimates obtained in previous rounds of the LFS. They play as the z variables. Their use is very strategic for increasing the accuracy of monthly LFS estimates, because they are strongly correlated with the current occupational status, that is the target variable. However, since their totals are estimated, they introduce additional variability in the estimates that must be accounted for.

The implicit model assumed by the AC estimator that involves listed auxiliary variables is defined for each NUTS 2 level and can be written as:

(AGE_CLASS*SEX)+	Individual	x
RG1+RG2+RG3+RG4+ CLASS_COMP1+CLASS_COMP2+ CLASS_COMP3+CLASS_COMP4+ CLASS_COMP5+CLASS_COMP6+	Household	x
(SEX+SEX*CLASS2574)*NUTS3+	Individual	x
SEX*(emp_1+unemp_1+inac_1+ emp_4+unemp_4+inac_4)	Longitudinal	z
- 1	no intercept	

It should be pointed that the model is without an intercept.

3.2.3 Population totals

In the LFS, besides the variables on the occupational status, auxiliary information on sex (male and female), age and geographical areas of residence (NUTS2: regions, NUTS3: provinces) are collected. These auxiliary variables are twofold important because, including them in the estimator, makes the weights consistent with the population distribution by sex and age in specific geographical areas; furthermore, they help in improving the estimates, because the occupational status is strongly correlated with sex, age and geographical areas of the individuals. According to the auxiliary variables in Section 3.2.2, five distributions based on current demographic estimates that are updated every quarter are observed:

- number of households by rotation group at NUTS2 level (block1, bl1);
- number of households by number of members at NUTS2 level (block2, bl2);
- population by sex and age class at NUTS2 level (block3, bl3);
- population by sex at NUTS3 level (block4, bl4);
- population by sex and age class 25-74 at NUTS3 level (block5, bl5).

3.2.4 Longitudinal totals

Totals of longitudinal auxiliary variables are used in this context to increase the efficiency of the estimates, reduce the volatility of the monthly series and lighten the impact of the small sample size of the monthly samples. They relate to the occupational status of individuals referred three months before (one quarter) and twelve months before (four quarters). They are usually highly correlated with the current occupational status (especially for employment), and they can reduce the sampling variance of the estimates.

This information is not available for all the individuals in the sample. The number of units in the sample for which these auxiliary variables are available depends on the rotation scheme adopted. In the Serbian LFS is 2-(2)-2. Theoretically, this represents 50% of the current sample. However, attrition and non-response reduce the actual overlapping rate (for instance, in the fourth quarter of 2019, the overlapping rate was 36% with the third quarter of 2019 and 30% with the fourth quarter of 2018). In this context, the use of the RC estimator

would introduce too much variability, as it would require the imputation of auxiliary variable values for approximately 70–74% of the sample, respectively. Therefore, the use of the AC estimator (and, equivalently, the AR estimator) is more appropriate. It is worth noting that, in this case, the totals of these variables, corresponding to the estimates of the occupational status referred to three and twelve months before, cannot be directly used. They have to be adjusted according to the overlapping rate (o) and the population change rate of individuals aged 15 years and over, between the periods (rc) and $\underline{\tilde{t}}_z$=o rc $\hat{t}_z$. Since the population totals are computed at the NUTS 2 level, all these quantities are computed at the same territorial level for the corresponding month. To determine the sampling variance of the estimates, it shall be ensured that these totals are estimates, so that the sample variance has to be adjusted, see Section 4.1.

3.3 Calibration

The use of the AC estimator consists of defining a calibration system as (4). It is preferable to use the weights already corrected for non-response. The distance function used is the truncated logarithmic function, which prevents negative and extreme weights. The auxiliary totals on which the calibrated weights are from current demographic estimations and from previous rounds of the survey. Overall, 344 auxiliary totals (86 auxiliary variables by 4 domains) are used in the calibration (Figure 3.2).

Figure 3.2 - Calibration system for the provisional estimates

NUTS2 (region)	b11	b12	b13	b14	tot_long		b15
	1-4	5-10	11-38	39-56	57-62	63-68	69-86
1 2 3 4	Households by rotation group at NUTS2 level	Households by number of components at NUTS2 level	Population by age class and sex at NUTS2 level	Population by sex at NUTS3 level	Population by occupational status and sex 3 months before at NUTS2 Level	Population by occupational status and sex 12 months before at NUTS2 level	Population by age class 25-74 and sex at NUTS3 level

Household auxiliary variables - Totals at the NUTS2 level available from the demographics estimates
Individual auxiliary variables - Totals at the NUTS2 level available from the demographics estimates
Individual auxiliary variables - Totals at the NUTS3 level available from the demographics estimates
Longitudinal auxiliary variables - Totals at the NUTS2 level available from previous rounds of the LFS

Source: Authors' processing

Table 3.2 presents the comparison between the design weights, the design weights corrected for non-response and the calibrated weights. The non-response treatment and the calibration increase the variability of the sampling weights. It can be seen according to the Kish index (Kish 1965) which passes from 1.07 to 1.11 (by correction for non-response) and finally to 1.78 (by calibration).

Table 3.2 - Comparison between the design weights, the design weights corrected for non-response and the calibrated weights. October 2019

SAMPLING WEIGHTS	Min	Percentile 25%	Median	Mean	Percentile 75%	Percentile 90%	Percentile 95%	Percentile 99%	Max
Design weights (dk)	195	458	518	492	568	605	605	606	606
Design weights corrected by non-response (dk_cnr)	187	403	551	538	686	699	737	921	1,796
Calibrated weights (dk_cnr.cal)	40	295	447	613	754	1,228	1,612	2,449	6,197

Source: Authors' processing

3.4 Point estimates

The calibrated weights (w_k, dk_cnr.cal), named as provisional weights, are used for deriving the provisional estimates with expression (1). Auxiliary variables and totals described in Section 3.2.2 are used for the calculation of the monthly estimates for the number of employed, unemployed, active, and inactive people by sex and by region. Table 3.3 shows the provisional monthly estimates at the national level for October 2019.

Table 3.3 - LFS Provisional monthly estimates. October 2019 (a)

Occupational Status	Sex	Estimate (in thousand)
Active		3,285
Employed		2,970
Unemployed		315
Inactive		2,627
Active	Male	1,834
Active	Female	1,450
Employed	Male	1,670
Employed	Female	1,299
Unemployed	Male	163
Unemployed	Female	151
Inactive	Male	1,018
Inactive	Female	1,609

Source: Authors' processing
(a) Please note that the estimates are not the official estimates. They are just related to this experimental study.

3.5 R procedures

The methodology and all steps described in Section 3 have been implemented in R software. The R package ReGenesees (Zardetto 2015) is used. This package enables to perform the calibration, the computation of the estimates of parameters and of the variance for sample surveys. The main stages of the R script are explained below.

After preparing the monthly dataset and the blocks of auxiliary totals, it is necessary to define the design weights and the adopted sampling design.

The first step is to define the sampling design for the dataset (data) through the function e.svydesign. The sampling design is a stratified two-stage sample where the PSUs are the EAs (KEY_PSU) and the SSUs (KEY_HH) are the households. The PSUs are stratified by region, type of settlement (urban/no urban) and rotation group (STRATUM). The sampling fractions at the first stage (fpc1, fraction based on the persons in the stratum) and at the second stage (fpc2, fraction based on the households in the stratum for the selected EAs) must be indicated. The design weights corrected for non-response (dk_nrc) are used. The output is a design object (des) that contains all the relevant information on the sampling design. The following syntax is used.

Invocating the object des displays the message explaining that a two-stage stratified sample plan is defined and that for the month October 2019 there are 200 strata, 724 PSUs that are the enumeration areas and 4426 SSUs that are the households. In the calculation there will be 733 PSUs introduced, instead of 724 which are involved in the survey.

```
> des <- e.svydesign(data=data,
+                    ids=~KEY_PSU+KEY_HH,
+                    strata=~STRATUM,
+                    weights=~dk_nrc,
+                    fpc=~fpc1+fpc2,
+                    check.data = TRUE)
> des
Stratified 2 - Stage Cluster Sampling Design
- [200] strata
- [724, 4426] clusters

Call:
e.svydesign(data = data, ids = ~KEY_PSU + KEY_HH, strata = ~STRATUM,
   weights = ~dk_nrc, fpc = ~fpc1 + fpc2, check.data = TRUE)
```

The reduction of the PSUs is due to the collapsing step, a necessary pre-arrangement for properly managing the variance computation for the first and second stage of the sample selection (Rust and Kalton 1987). Strata that contain one PSU (enumeration area) and PSUs that contain one SSU (household) are collapsed (jointed with the corresponding, closest stratum or primary sampling unit) to enable the computation of the variance within the stratum and within the PSU.

The function collapse.strata is used for collapsing strata with one PSU. The parameter block defines inside which domain the strata can be collapsed, while an ad-hoc function manages the case in which just one SSU is observed in one PSU. The output des.c is an update of the design object des in which the additional information on collapsed strata is added. PSUs are grouped by an ad-hoc procedure.

```
> des.c <- collapse.strata(des,block=~NUTS3)
```

In order to perform the calibration, the following three steps are required: i) definition of the template for the auxiliary totals (pop.template); ii) filling the template; iii) calibration (e.calibrate).

The calibration model is defined with the pop.template function. The auxiliary variables to be used are listed and the level at which the totals are provided is defined (partition).

The output, tot.cal_AC, is an empty dataframe with 4 rows (one for each domain according to the variables specified in partition) and 86+1 columns (the header row plus one column for each variable included in the model).

```
> contrasts.off()

# Contrasts treatment has been switched off:

$contrasts
  unordered     ordered
"contr.off" "contr.off"

> tot.cal_AC <- pop.template(data=des.c,
+                        calmodel=~(AGE_CLASS:SEX)+
+                                 RG1+RG2+RG3+RG4+
+                                 CLASS_COMP1+CLASS_COMP2+
+                                 CLASS_COMP3+CLASS_COMP4+
+                                 CLASS_COMP5+CLASS_COMP6+
+                                (SEX+SEX:CLASS2574):COUNT_PROV+
+                                 SEX:(emp_1+unemp_1+inac_1+
+                                     emp_4+unemp_4+inac_4)-1
+                        partition=~NUTS2)
```

The command contrasts.off() considerably simplifies the procedure of compiling the output of pop.template. When it is used, it is sufficient to add, beside the blocks of auxiliary totals defined in the previous Sections, the R command with the cbind function (function for collapsing together two data frames when the number of rows in both data frames is equal).

```
> tot.cal_AC[,-1] <- cbind(bl1[,-1],bl2[,-1],bl3[,-1],bl4[,-1],
+                          tot_long[,-1],bl5[,-1])
```

Calibrated weights are obtained using the function e.calibrate. For the use of this function, it is necessary to assign first the values to its parameters, as it can be seen in the program below. So, the design is defined by the object des.c, the data frame with auxiliary totals is tot.cal_AC, the calibration model is calmodel, the level at which the totals are planned is defined with the function partition, calfun is a logit distance function, with the lower and upper bounds given by the parameter bounds and aggregate.stage is the function which defines the level on which calibrated weights are equal.

In particular, calfun="logit" and bounds=c(0.1, 9.6) means that the truncated logarithmic distance with L=0.1 and U=9.6 is used. It is important to point out that bound cannot be defined a priori and depends by the sample and the system of constraints. While, aggregate.stage=2 means that all the individuals in the same household (the units of stage 2) will have the same calibrated weights.

```
> cal_AC <- e.calibrate(design=des.c,
+                       df.population=tot.cal_RC,
+                       calmodel=~(AGE_CLASS:SEX)+
+                                 RG1+RG2+RG3+RG4+
+                                 CLASS_COMP1+CLASS_COMP2+
+                                 CLASS_COMP3+CLASS_COMP4+
+                                 CLASS_COMP5+CLASS_COMP6+
+                                (SEX+SEX:CLASS2574):COUNT_PROV+
+                                 SEX:(emp_1+unemp_1+inac_1+
+                                     emp_4+unemp_4+inac_4)-1,
+                       partition=~NUTS2,
+                       calfun = "logit",
+                       bounds = c(0.1, 9.6),
+                       aggregate.stage = 2,
+                       force=FALSE)
> check.cal(cal_AC)
All Calibration Constraints (344) fulfilled (at tolerance level epsilon = 0.0000001).
```

When the calibration is successful, the message of check.cal(cal_AC) states that all the calibration constraints (344 in this case) are fulfilled and a calibrated object is created. It can be seen as an update of des.c.

All the information related to the calibration are included in this new object and, the calibrated weights (w_k, dk_ncr.cal) are added to the dataset. The weights dk_ncr.cal are the provisional weights and they will be used for computing the provisional estimates.

The function svystat is used for calculating the provisional estimates. Since that one should calculate the estimates for the total number of active (act), employed (emp), unemployed (unemp) and inactive (inac) persons by region (NUTS2) and (SEX), the parameter kind (type of estimator) is set to "TM" and the parameter estimator is "Total".

The parameter combo=2 enables the computation of the estimates at national level, by region, by sex and by region and sex.

```
> out <- svystat(design=cal_AC,
+              kind ="TM", estimator="Total",
+              y=~act+emp+unemp+inac,
+              by=~NUTS2+SEX,
+              combo=2)
```

4. Sampling variance estimation

4.1 Theoretical bases

The AC estimator used for producing the monthly estimates includes auxiliary variables for which totals can be desumed both from a register, then exactly known, and from another survey or from a previous round of the survey, then affected by sampling errors. Therefore, its sampling variance depends on two sources of variability. The first source is the survey itself and entails stratification, stadification, sample size, sampling allocation. While, the second source is the variance of the estimated auxiliary totals.

Expression (3) related to the calculation of the CAL estimator variance, which does not include the variance of the estimated auxiliary totals, is not recommended for the calculation of the AC estimator variance (Berger et al. 2009; Dever and Vaillant 2010), as it can lead to underestimation of the sampling variance.

The suitable estimator for the variance is proposed in Berger et al. (2009), Renssen and Nieuwenbroek (1997), Ceccarelli and Guandalini (2012) and Guandalini (2014). These authors use the link between the CAL estimator and the GREG estimator as well as the AC estimator and the AR estimator.

The variance estimator of the AC estimator is derived starting from the expression of the AR estimator in (5), by adding and subtracting the expression t_u

$$\begin{aligned}\hat{t}_{y_{AR}} &= \hat{t}_{y_{HT}} + (\tilde{\underline{t}}_u - \underline{t}_u + \underline{t}_u - \hat{\underline{t}}_{u_{HT}})\hat{\underline{\beta}} \\ &= \hat{t}_{y_{HT}} + (\underline{t}_u - \hat{\underline{t}}_{u_{HT}})\hat{\underline{\beta}} - (\underline{t}_u - \tilde{\underline{t}}_u)\hat{\underline{\beta}}\end{aligned} \tag{6}$$

where $\underline{t}_u = (\underline{t}_X, \underline{t}_z)$ is the vector of the auxiliary totals, assuming that totals of the z auxiliary variables are not affected by sampling errors. In (6) $\hat{t}_{yHT} + (\underline{t}_u - \hat{\underline{t}}_{uHT})\widehat{\beta}$ is the GREG estimator, $\hat{t}_{\mathrm{yGREG}}$ when both x and z variables are considered, assuming that the related totals are not affected by sampling error.

Therefore, the AR estimator can be written as

$$\hat{t}_{y_{AR}} = \hat{t}_{y_{HT}} + (\underline{t}_u - \hat{\underline{t}}_{u_{HT}})\hat{\underline{\beta}} - (\underline{t}_u - \tilde{\underline{t}}_u)\hat{\underline{\beta}}$$
$$= \hat{t}_{y_{GREG}} - (\underline{t}_u - \tilde{\underline{t}}_u)\hat{\underline{\beta}}$$

and the estimator of the sampling variance is equal to

$$\hat{\mathbb{V}}(\hat{t}_{y_{AR}}) = \hat{\mathbb{V}}(\hat{t}_{y_{GREG}}) + \hat{\mathbb{V}}\left((\underline{t}_u - \tilde{\underline{t}}_u)\hat{\underline{\beta}}\right) - 2\widehat{COV}\left(\hat{t}_{y_{GREG}}, (\underline{t}_u - \tilde{\underline{t}}_u)\hat{\underline{\beta}}\right)$$
$$= \hat{\mathrm{A}}_1 + \hat{\mathrm{A}}_2 - \hat{\mathrm{A}}_3 \qquad (7)$$
$$\cong \widehat{\mathrm{AV}}(\hat{t}_{y_{AC}})$$

Expression (7) approximates the variance of the AC estimator. The variance determined by

$$\hat{\mathrm{A}}_1 = \hat{\mathbb{V}}(\hat{t}_{y_{GREG}}) \qquad (8)$$

represents the variance of the GREG estimator (expression 3). The variance of the auxiliary variables *x* and *z* is:

$$\hat{\mathrm{A}}_2 = \hat{\underline{\beta}}^t \hat{\mathbb{V}}(\tilde{\underline{t}}_u)\hat{\underline{\beta}} \qquad (9)$$

$$= \hat{\underline{\beta}}_z^t \hat{\mathbb{V}}(\tilde{t}_z)\hat{\underline{\beta}}_z \qquad (10)$$

as, only the totals $\tilde{\underline{t}}_z$ are estimated. The component $\hat{\mathrm{A}}_3$ of the AC eastimator variance is:

$$\hat{\mathrm{A}}_3 = 2\widehat{COV}\left(\hat{t}_{y_{GREG}}, \tilde{\underline{t}}_u\hat{\underline{\beta}}\right) \qquad (11)$$

$$= 2\sqrt{\hat{\mathbb{V}}(\hat{t}_{y_{GREG}})}\left[\widehat{CORR}(\hat{t}_{y_{GREG}}, \tilde{\underline{t}}_z)\right]\sqrt{\hat{\mathbb{V}}(\tilde{\underline{t}}_z)}\hat{\underline{\beta}}_z \qquad (12)$$

and represents the covariance between $\widehat{\mathrm{A}}_1$ and $\widehat{\mathrm{A}}_2$. It takes into account that the totals of the *z* auxiliary variables have been estimated on a not negligible fraction of the units surveyed in the current round. The covariance in expression (11) is computed in terms of correlation (expression (12), according to relation $\widehat{\mathrm{COV}}(\mathrm{X},\mathrm{Y}) = \widehat{\mathrm{CORR}}(\mathrm{X},\mathrm{Y})\sqrt{\widehat{\mathrm{V}}(\mathrm{X})}\sqrt{\widehat{\mathrm{V}}(\mathrm{Y})}$.

Expression (7) summarised all the steps needed for properly assessment of the variance of the AC (equivalently AR) estimator. Expression (8) is the sampling variance when considering all the auxiliary totals exactly known and that are error-free.

For the auxiliary totals that are estimates based on sample and which are affected by the sampling error, variance (8) has to be increased for the value given by expression (9), respectively by expression (10), as the totals of *x* auxiliary variables have no sampling error.

4.2 Application to the Serbian LFS

For the interpretation of results based on a random sample, the estimate of their accuracy is of great importance. This is particularly relevant for monthly estimates, as analysing their standard errors can determine which part of the estimate is based on the change of an indicator and which part is based on the sampling error (see, for example, Wolter 2007).

The sampling variance of the Serbian LFS monthly estimates is estimated using expression (7). As the auxiliary totals referred three and twelve months before, are adjusted according to the overlapping rate (o) and the population change rate of persons aged 15 and over, between the periods (rc), then the sample variance is adjusted as well. According to the relation valid for the variance $\widehat{\mathrm{V}}(abX) = a^2b^2\widehat{\mathrm{V}}(X)$, and the requirement that $\tilde{t}_z = o\,r\,c\,\hat{t}_z$, the variance of auxiliary totals $\tilde{t}_z$ is calculated based on the expression $\widehat{\mathrm{V}}(\tilde{t}_z) = \widehat{\mathrm{V}}(orc\,\hat{t}_z) = o^2r^2c^2(\widehat{\mathrm{V}})(\hat{t}_z)$.

In Table 4.1 to the estimates at the national level presented in Table 3.3 attached are: standard error (SE), the coefficient of variation (CV), the coefficient of variation in percents (CV%) and limits (LB and UB) for 95% confidence interval.

To show the influence of longitudinal auxiliary variables on the sample variance, all components in expression (7) are compared, as well as the sample variance in the case when only *x* variables are used in the calculation. Four variance estimates were observed:

CALx is the estimate when only the x variables are considered, the CAL estimator in (1) is used and the sampling variance is computed using expression (3).

CALxz considers only the component $\widehat{A}_1$ in (7), the z auxiliary variables are considered and their totals are assumed to be exactly known, as well as those of the x auxiliary variables.

ACxz (covariance = 0) takes into account the sampling error introduced in the estimates because of the estimated totals of the z auxiliary variables, but it does not consider the covariance term and only $\widehat{A}_1$ and $\widehat{A}_2$ in (7) in the expression are accounted.

ACxz (covariance $\neq$ 0) considers all the components in expression (7), that is $\widehat{A}_1$, $\widehat{A}_2$ and $\widehat{A}_3$.

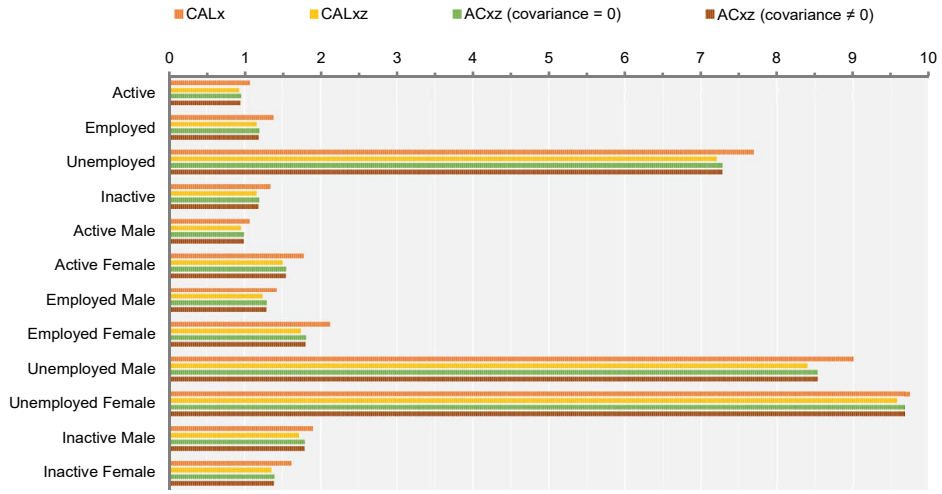
Table 4.1 - LFS monthly estimates (provisional) with accuracy measures. October 2019
(a) (b)

Occupational Status	Sex	Estimate (in thousand)	SE	VAR	CV	CV%	LB	UB
Active		3,285	31.040	963.479	0.009449	0.944941	3,224	3,346
Employed		2,970	35.171	1236.981	0.011842	1.184209	2,901	3,039
Unemployed		315	22.930	525.775	0.072793	7.279304	270	360
Inactive		2,627	31.027	962.706	0.011811	1.181125	2,566	2,688
Active	Male	1,834	18.166	329.995	0.009905	0.990519	1,798	1,869
Active	Female	1,450	22.371	500.444	0.015428	1.542833	1,407	1,494
Employed	Male	1,670	21.480	461.371	0.012862	1.286241	1,628	1,712
Employed	Female	1,299	23.412	548.116	0.018023	1.802349	1,253	1,345
Unemployed	Male	163	13.907	193.414	0.085321	8.532076	136	190
Unemployed	Female	151	14.614	213.567	0.096781	9.678099	122	180
Inactive	Male	1,018	18.184	330.640	0.017862	1.786211	982	1,054
Inactive	Female	1,609	22.339	499.047	0.013884	1.388386	1,565	1,652

Source: Authors' processing
(a) (SE) standard error; (CV) coefficient of variation; (CV%) coefficient of variation in percents; (LB and UB) extremes of the confidence interval at 95%.
(b) Please note that the estimates are not the official estimates. They are just related to this experimental study.

Figure 4.1 shows the coefficients of variation for four variance estimates. It can be seen that including the z auxiliary variables in the estimation almost always increases the accuracy. Of course, the increase is higher when ignoring that the totals of the z auxiliary variables are estimates. But, even when the sampling variance of the AC estimator is properly managed, there is a general advantage in considering the longitudinal auxiliary variables. The possible disadvantage associated with some estimates is paid off by an important increase of the accuracy, for instance, for the unemployed status.

Figure 4.1 - Coefficient of variation of the LFS provisional estimates. October 2019
(percentage values)



Source: Authors' processing

4.3 R procedures

The computation of the sampling variance of the AC estimator in expression (7) has been implemented in R.

The software ReGenesees directly provides the component $\widehat{A}_1$. The output of the function svystat, besides the estimates (Y), provides the sampling variance (VAR). VAR is the variance of the CAL estimator (equivalent to the GREG estimator), that is, is computed with expression (3), assuming that totals for the x and z auxiliary variables are not affected by sampling error. Therefore, the components $\widehat{A}_2$ and $\widehat{A}_3$ have to be computed and included.

```
> out <- svystat(design=cal_RC,
+              kind ="TM", estimator="Total",
+              y=~act+emp+unemp+inac,
+              by=~NUTS2+SEX,
+              combo=2)
```

For $\widehat{A}_2$ in expression (10) the regression coefficients $\underline{\beta}$ of the implicit superpopulation model $\xi : y = \underline{X}\,\underline{\beta} + e$ with $U = (X, Z)$ must be estimated on the sample. A model involving all the 344 auxiliary variables (296 x and

48 *z*) for each of the y variable (such as active, employed, …, active male, active female, …) has to be defined and for each of them, the regression coefficients have to be estimated. For making the computation faster, the regression coefficients are computed using the normalised calibrated weights (w) instead of the calibrated weights.

```
> b <- matrix(nrow=344,ncol=ncol(y))
> w <- (data$dk_nrc.cal/sum(data$dk_nrc.cal))*nrow(data)
> for(i in 1:ncol(y)){
+   model <- as.formula(paste(colnames(y)[i],
+                             "~((AGE_CLASS:SEX)+
+                               RG1+RG2+RG3+RG4+
+                               CLASS_COMP1+CLASS_COMP2+
+                               CLASS_COMP3+CLASS_COMP4+
+                               CLASS_COMP5+CLASS_COMP6+
+                              (SEX+SEX:CLASS2574):COUNT_PROV+
+                               SEX:(emp_1+unemp_1+inac_1+
+                               emp_4+unemp_4+inac_4)):NUTS2-1",
+                       sep=""))
+ beta <- lm(model,data,weights=w)$coefficient
+ b[,i] <- beta
+ }
> colnames(b) <- colnames(y)
> rownames(b) <- names(beta)
> b <- data.frame(b)
```

Only the regression coefficients related to the *z* auxiliary variables are needed. Since the sampling variance of $\tilde{\underline{t}}_z$ has already been produced three and twelve months before, the component $\widehat{A}_1$ is obtained just by multiplying these quantities.

For $\widehat{A}_3$ in expression (11) the correlation should be estimated. There are several ways for estimating the correlation between two totals under complex sampling designs (see, e.g., Berger and Priam 2016). In this study, a simpler approximation is used in order to reduce the computational burden, and only the correlation between $\hat{\underline{t}}_{yGREG}$ and $\tilde{\underline{t}}_z$. was computed. $\widehat{V}(\hat{t}_{yGREG})$ and $\widehat{V}(\hat{t}_{yGREG})$ have already been computed and $\widehat{A}_3$ is obtained just by multiplying these quantities as in expression (12).

```
> corrYZ <- matrix(nrow=ncol(y), ncol=ncol(z))
> for(i in 1:ncol(y)){
+     e <- y[,i]
+     for(j in 1:ncol(z)){
+         corrYZ[i,j] <- ifelse(is.na(cov(e,z[,j])),0,cov(e,z[,j]))
+     }
+ }
```

5. Final estimates consistent with the quarterly estimates

The main outputs of the LFS are the quarterly estimates. Quarterly data allow to produce very detailed information on the labour market, data are carefully checked and treated and the sample dimension allows to obtain more precise estimates related to the monthly estimates. For the easier interpretation of the results, and to stabilize monthly series it is important that the monthly estimates are consistent with the quarterly ones. Therefore, at the end of each quarter, there is a revision of the provisional monthly estimates and the dissemination of the final monthly estimates adjusted to be consistent with the quarterly ones.

Each quarterly estimate, $\hat{t}_y^Q$, can be seen as a weighted mean of the monthly estimates, $\hat{t}_y^{Mm}$ (m=1, 2, 3), where the weights are the number of the weeks of the month (4 or 5). For instance, in the 4th quarter 2019, the first month has 5 weeks, while the second and the third 4, then

$$\hat{t}_y^Q = \frac{5\hat{t}_y^{M1} + 4\hat{t}_y^{M2} + 4\hat{t}_y^{M3}}{13}$$

This result can be easily obtained through a calibration of the monthly estimates on the estimates referred to the quarter for the main labour market indicators. The consistence of the estimates is then guaranteed by definition. To avoid too much modification of the provisional monthly estimates most of the auxiliary variables already used for providing the provisional monthly weights are included in this calibration step. The auxiliary variables are mainly those related to the population distribution by sex, age classes and regions. The output is a system of weights that for each month enables to provide the final monthly estimates consistent with the quarterly ones.

5.1 Application to the Serbian LFS

To produce the final monthly estimates it is necessary to define a new calibration system as in expression (2), on which the calibrated weights that were used in computation for the provisional monthly estimates (provisional weights dk_nrc.cal) will be modified. All monthly data are treated together. Auxiliary variables included in the calibration process are the following:

12. sex (SEX, M=male, F=female);
13. age class (AGE_CLASS, 0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+);
14. month (month, M1=October, M2=November, M3=December);
15. age class 15-24 (0=no, 1=yes);
16. age class 25-74 (0=no, 1=yes);
17. number of the weeks in the month (nweek, 4, 5);
18. employed (emp=yes, 0=no) [emp];
19. unemployed (unemp=yes, 0=no) [unemp];
20. inactive (inac=yes, 0=no) [inac];
21. employed (emp*nweek/13=yes, 0=no) [Xemp];
22. unemployed (unemp*nweek/13=yes, 0=no) [Xunemp];
23. inactive (inac*nweek/13=yes, 0=no) [Xinac].

The auxiliary variables 1-9 refer to the observed month and serve to preserve the coherence of the monthly estimates with the corresponding reference population. This coherence was already assured by the provisional monthly weights and must be maintained in the final monthly weights.

The auxiliary variables 10-12 serve to reach the coherence between the monthly estimates of the main indicators and the corresponding quarterly estimates, through a weighted mean. A practical solution to obtain this coherence is a weighting of the auxiliary variables 7-9 by the number of the weeks in the observed month for obtaining auxiliary variables 10-12.

The implicit model assumed at the NUTS 2 level is:

```
(AGE_CLASS*SEX*month)+                  Individuals
((Xemp+Xunemp+Xinac)*                   occupational status
(CLASS1524+CLASS2574):SEX)
-1                                      no intercept
```

The definition of this model enables to obtain monthly data for the occupational status (employed, unemployed and inactive), that is in accordance with corresponding monthly population, and all together they determine the parameter estimate that is equal to the already calculated one.

The totals used are (Figure 5.1):

- population by sex and age class at the NUTS2 level (block1, bl1);
  - for the first month of the quarter (block1a, bl1a);
  - for the second month of the quarter (block1b, bl1b);
  - for the third month of the quarter (block1c, bl1c);
- male 15-24 by status at NUTS2 level from quarterly estimates (block2 bl2);
- female 15-24 by status at NUTS2 level from quarterly estimates (block3, bl3);
- male 25-74 by status at NUTS2 level from quarterly estimates (block4, bl4);
- female 25-74 by status at NUTS2 level from quarterly estimates (block5, bl5).

Figure 5.1 - Calibration system for the final estimates

NUTS2 (region)	bl1			bl2	bl3	bl4	bl5
	bl1.a	bl1.b	bl1.c				
	1-28	29-56	57-84	85-87	88-90	91-93	94-96
1 2 3 4	Population by age class and sex at NUTS2 level Month 1	Population by age class and sex at NUTS2 level Month 2	Population by age class and sex at NUTS2 level Month 3	Male pop. 15-24 y.o. by occupational status at NUTS2 level	Female pop. 15-24 y.o. by occupational status at NUTS2 level	Male pop. 25-74 y.o. by occupational status at NUTS2 level	Female pop. 25-74 y.o. by occupational status at NUTS2 Level

Monthly auxiliary variables – Totals at NUTS2 level available from current demographic estimates.
Quarterly auxiliary variables – Totals at NUTS2 level available from the quarterly estimates

Source: Authors' processing

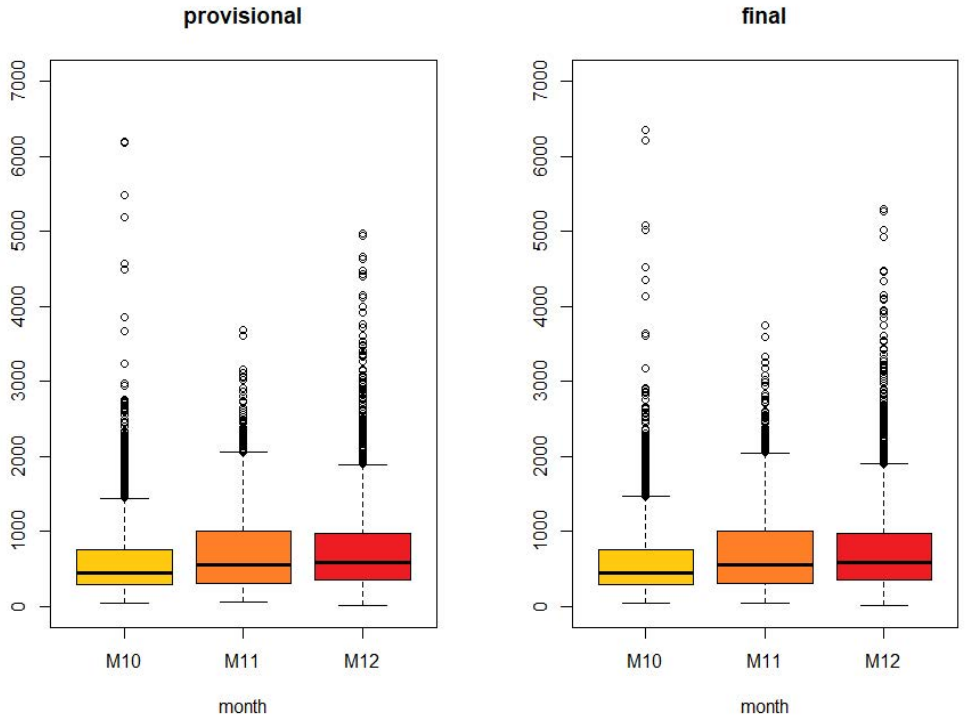
In this experimental study for the 4th quarter 2019, consistency of monthly and quarterly data is reached. The calibration process converges setting the limits L=0.7 and U=1.2 in logit distance function. The final weights (dk_cnr.cal.cal) are the calibrated weights obtained according to the provisional weights (dk_cnr.cal). This step of calibration slightly changes the provisional weights, which can be seen in the comparative overview in Table 5.1 and Figure 5.2.

Table 5.1 - Comparison between the final weights of the three months of the quarter. October-December 2019 (a)

Weights	Month	Min	Perc. 25%	Mean	Median	Perc. 75%	Perc. 90%	Perc. 95%	Perc. 99%	Max
Provisional (dk_cnr.cal)	October	40	295	447	613	754	1,228	1,612	2,449	6,197
	November	49	302	544	730	1,004	1,609	1,903	2,480	3,688
	December	9	348	584	776	967	1,551	2,065	3,488	4,975
Final (dk_cnr.cal.cal)	October	35	293	447	613	759	1,227	1,591	2,532	6,349
	November	42	303	546	730	1,001	1,605	1,893	2,503	3,747
	December	9	348	582	776	969	1,543	2,014	3,542	5,292

Source: Authors' processing
(a) Perc.=Percentile.

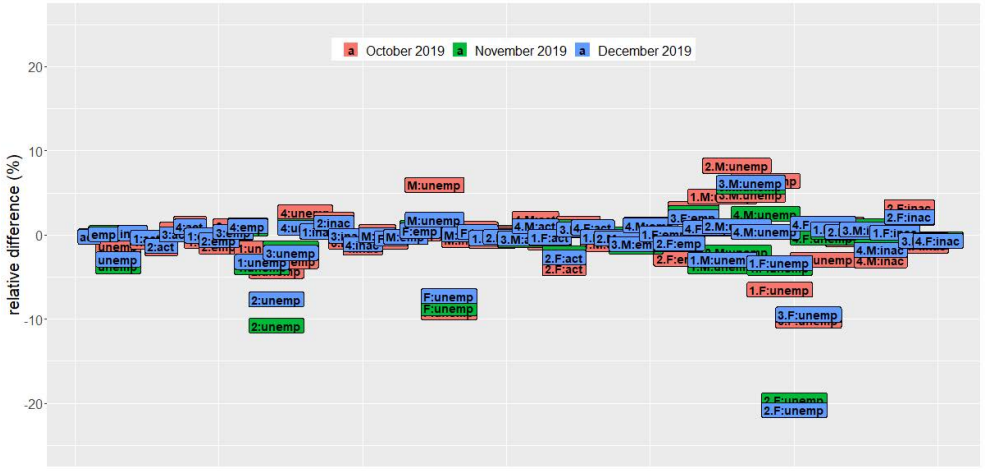
Figure 5.2 - Boxplot of the provisional weights (dk_cnr.cal) and the final weights (dk_cnr. cal.cal) for each month of the fourth quarter. Year 2019



Source: Authors' processing

Final monthly estimates are in average close to the provisional monthly estimates. This can be seen in Figure 5.3. Most of the estimates, i.e. 99 estimates over 180 estimates have a relative change lower than 1%, 64 estimates change less than 5%, while the remaining 17 estimates change more than 5%, and only the estimates of the number of unemployed women in region 2 are modified around the 20% in all months.

Figure 5.3 - Relative difference between final and provisional monthly estimates. October-December 2019 (percentage values)



Source: Authors' processing

Considering the accuracy of the final monthly estimates, it should be pointed out that only some estimates of totals are used in the calibration to achieve consistency with the quarterly data.

Contrary to the process, which was done for the provisional estimates, here their sampling error was not accounted for assessing the sampling variance of the final estimates.

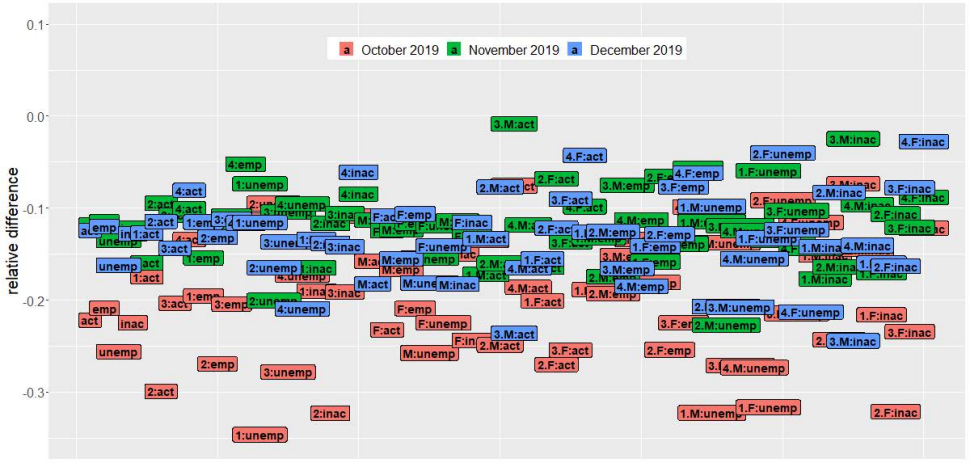
Considering that quarterly estimates are reliable, it is assumed that this approximation has not a great impact on the sampling error of the final monthly estimates.

Comparing the coefficient of variation of the final monthly estimates with the coefficients of variation of the provisional monthly estimates, it can be

seen in Figure 5.4, a general decrease of the sampling error (most of the relative percentage differences are negative).

The decrease of the sampling error is proportional to the reduction of variability of the monthly sampling weights due to the reduction in calibration constraints from NUTS3 to the NUTS2 level.

Figure 5.4 - Relative difference between the coefficient of variation of the provisional and final monthly estimates. October- December 2019



Source: Authors' processing

5.2 R procedures

The first step to achieve consistency between monthly and quarterly estimates is to create a quarterly dataset putting together all monthly datasets with their provisional weights (dk_nrc.cal).

For the calibration procedure, the totals of the quarterly estimates (totals 85-96) has to be derived by the quarterly weights (CALWEIGHT), while, the population totals for each of the three months (totals 1-84) have already been calculated. Using the following commands, the calibration system and the matrix of the auxiliary totals will be defined:

```
> contrasts.off()

# Contrasts treatment has been switched off:

$contrasts
  unordered     ordered
"contr.off" "contr.off"

> # definition of the calibration system
> tot.cal <- pop.template(data=des,
+                         calmodel=~(AGE_CLASS:SEX:month)
+                                 +((Xemp+Xunemp+Xinac):(CLASS1524+CLASS2574):SEX) -1,
+                         partition=~NUTS2)
> tot.cal[,-1] <- tot[,-1]
```

Then the calibration can be performed by function e.calibrate explained in Section 3.5. The truncated logarithmic function, calfun = "logit" is used. The bounds for the weights L and U are set by the function bounds = c(0.7, 1.2). The final weights are calculated to be equal for all the individuals in the same household, by the function aggregate.stage = 2.

```
> cal <- e.calibrate(design=des,
+                    df.population=tot.cal,
+                    calmodel=~(AGE_CLASS:SEX:month)+
+                             ((Xemp+Xunemp+Xinac):
+                              (CLASS1524+CLASS2574):
+                              SEX)
+                             -1,
+                    partition=~NUTS2,
+                    calfun = "logit",
+                    bounds = c(0.7, 1.2),
+                    aggregate.stage = 2,
+                    force=FALSE)
> check.cal(cal)
All Calibration Constraints (384) fulfilled (at tolerance level epsilon = 0.0000001).
```

The message of check.cal(cal) in the previous set of commands states that all the calibration constraints (384 in this case) are fulfilled and the calibrated object is created. The cal object contains the data for all the months of the quarter and the final weights to be used for providing the final estimates (dk_ncr.cal.cal). It means that provisional weights dk_ncr.cal, are calibrated.

After the calibration is performed, using the function ext.calibrate, the des object is defined and contains the information for the sample design plan for each monthly sample, the calibration model used for each month, and the calibration model used to achieve consistency with quarterly data.

In practice, this function is a combination of e.svydesign and e.calibrate and is demonstrated by the following set of commands.

```
> des <- ext.calibrated(data=cal$variables,
+                       ids=~KEY_PSUc+KEY_HH,
+                       strata=~STRATUM,
+                       weights=~dk_nrc,
+                       fpc=~fpc1+fpc2,
+                       check.data=TRUE,
+                       weights.cal=~dk_nrc.cal.cal,
+                       calmodel=~month:((AGE_CLASS:SEX)+
+                                        RG1+RG2+RG3+RG4+
+                                        CLASS_COMP1+CLASS_COMP2+
+                                        CLASS_COMP3+CLASS_COMP4+
+                                        CLASS_COMP5+CLASS_COMP6+
+                                        (SEX+SEX:CLASS2574):COUNT_PROV+
+                                         SEX:(emp_1+unemp_1+inac_1+
+                                              emp_4+unemp_4+inac_4)
+                                        )+
+                                   ((Xemp+Xunemp+Xinac):
+                                    (CLASS1524+CLASS2574):
+                                     SEX)
+                                   -1,
+                       partition=~NUTS2)
```

For computing the final monthly estimates, the object des is an input of the function svystat. The estimate of the totals for the number of active (act), employed (emp), unemployed (unemp) and inactive (inac) persons by region (NUTS 2) and sex (SEX) is performed by the parameter kind="TM" (Totals and Means) and the estimator="Total". Setting the parameter group=~month enables to immediately compute the estimates for all the three months. The output object is a list object, and each element contains the estimates for the observed month.

```
> out <- svystat(design=des,kind ="TM",estimator="Total",
+              y=~act+emp+unemp+inac, by=~NUTS2+SEX, combo=2,
+              group=~month)
```

In the process of variance estimation, the impact of the quarterly estimates of totals in the last calibration step is ignored because it is assumed to be negligible, contrary to the impact of the longitudinal auxiliary totals. The procedure already shown in Section 4.3 has to be applied individually to each monthly data. The final estimates and the related measures of accuracy, such as sampling variance (VAR), standard error (SE), coefficient of variation (CV), coefficient of variation percent (CVpct) and the limits (LB, UB) of the 95% confidence intervals are computed and saved to be used for the calculation of the monthly estimates in the subsequent quarters.

6. Concluding remarks

The aim of this study was to define a methodology for calculating monthly Labour Force Survey estimates that incorporates data from previous survey waves (longitudinal data). In this way, the accuracy of the estimate would be improved, as the estimate is calculated on the basis of one-third of the quarterly sample size, while also using available historical data.

Considering the LFS sample plan, an estimation method was used that, in addition to the known auxiliary totals, introduces estimates of totals from previous waves of the survey into the calibration process (AC estimator). When defining the matrix of auxiliary calibration totals, it is necessary to correct the estimated totals according to the rate of overlap and the rate of change of the population aged 15 and over for the period between the previous and current survey. In addition, the AC method differs from the standard calibration procedure because the estimated variance includes the error of the estimated totals, which must be included in the estimates of the estimations of the accuracy.

The defined method refers to the estimates that would be calculated immediately after the end of the month and need to be published within the provided period. The study also defines the procedure for creating final monthly estimates after the end of the quarter, which are consistent with the quarterly estimates, differ slightly from the initial estimates, and are also published.

The practical goal of using the Serbian Labour Force Survey data to calculate quarterly estimates, as well as timely monthly estimates of key labour market indicators that are in accordance with Eurostat requirements, is achieved.

In calculating the provisional monthly estimates, 344 auxiliary variables were used, of which 296 relate to households and persons and their totals are known based on current demographic estimates, while 48 auxiliary variables were estimated based on the LFS three and twelve months earlier.

To calculate the final monthly estimates that are to be consistent with the quarterly estimates, 384 auxiliary variables were used, of which 336 related to persons and were also used in the calibration step to calculate the provisional monthly estimates, while 48 auxiliary variables were estimated based on LFS quarterly data.

The calculated provisional and final monthly estimates presented in this paper are not official, but their accuracy is in line with the accuracy of the monthly estimates of European countries published for monthly LFS data. The provisional monthly estimates are very close in value to the final monthly estimates, and it can be concluded that the applied methodology is appropriate and can be used for the calculation of the monthly estimates, especially when the volatility of the monthly estimates is taken into account.

References

Ballin, M., P.D. Falorsi, e A. Russo. 2000. "Condizioni di coerenza e metodi di stima per le indagini campionarie sulle imprese". *Rivista di statistica ufficiale/Review of official statistics*, Volume 2/2000: 31-52.

Berger, Y.G., J.F. Mũnoz, and E. Rancourt. 2009. "Variance estimation of survey estimates calibrated on estimated control totals - An application to the extended regression estimator and the regression composite estimator". *Computational Statistics & Data Analysis*, Volume 53, N. 7: 2596-2604. https://doi.org/10.1016/j.csda.2008.12.011.

Berger, Y.G., and R. Priam. 2016. "A simple variance estimator of change for rotating repeated surveys: an application to the European Union Statistics on Income and Living Conditions household surveys". *Journal of the Royal Statistical Society, Statistics in Society*, *Series A*, Volume 179, N. 1: 251-272. https://doi.org/10.1111/rssa.12116.

Bethlehem, J.G., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ, U.S.: John Wiley & Sons. https://doi.org/10.1002/9780470891056.ch14.

Bethlehem, J.G., and W.J. Keller. 1987. "Linear weighting of sample survey data". *Journal of Official Statistics*, Volume 3, N. 2: 141-153.

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1976. "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations". *Biometrika*, Volume 63, N. 3: 615-620. https://doi.org/10.2307/2335742.

Ceccarelli, C., and A. Guandalini. 2012. "Increasing the accuracy of It-SILC estimates through the use of auxiliary information from Labour Force Survey". *Statistica Applicata - Italian Journal of Applied Statistics*, Volume 24, N. 1: 103-115.

Dever, J.A., and R. Valliant. 2010. "A comparison of variance estimators for poststratification to estimated control totals". *Survey Methodology*, Volume 36, N. 1: 45-56. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11251-eng.pdf?st=L2HC2935.

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, Volume 87, N. 418: 376-382. https://doi.org/10.2307/2290268.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling". *Journal of the American Statistical Association*, Volume 88, N. 423: 1013-1020. https://doi.org/10.2307/2290793.

Fuller, W.A. 1999. "The Canadian Regression Composite Estimation". *Unpublished Manuscript*.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. Hoboken, NJ, U.S.: John Wiley & Sons. https://doi.org/10.1002/9781118490082.

Guandalini, A. 2014. *Coerenza ed ottimalità dello stimatore calibrato*. PhD dissertation.

Horvitz, D.G., and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe". *Journal of the American Statistical Association*, Volume 47, N. 260: 663-685. https://doi.org/10.2307/2280784.

Isaki, C.T., and W.A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model". *Journal of the American Statistical Association*, Volume 77, N. 377: 89-96. https://doi.org/10.2307/2287773.

Kish, L. 1965. *Survey sampling*. New York, NY, U.S.: John Wiley & Sons.

Little, R.J.A., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ, U.S.: John Wiley & Sons. https://doi.org/10.1002/9781119013563.

Lemaître, G., and J. Dufour. 1987. "An Integrated Method for Weighting Persons and Families". *Survey methodology*, Volume 13, N. 2: 199-207. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1987002/article/14607-eng.pdf?st=FcCf9qHs.

Rancourt, É. 2001. "La régression étendue : un ensemble de pratiques d'estimation qui poussent constamment la théorie". In Droesbeke, J.-J., et L. Lebart (sous la direction de). *Enquêtes, modèles et applications*: 334-343. Malakoff, France: Dunod.

Renssen, R.H., and N.J. Nieuwenbroek. 1997. "Aligning Estimates for Common Variables in Two or More Sample Surveys". *Journal of the American Statistical Association*, Volume 92, N. 437: 368-374. https://doi.org/10.2307/2291482.

Rust, K., and G. Kalton. 1987. "Strategies for Collapsing Strata for Variance Estimation". *Journal of Official Statistics*, Volume 3, N. 1: 69-81. https://www.proquest.com/openview/9648e5a196880c3765a96af6e4b50ae2/1?pq-origsite=gscholar&cbl=105444.

Särndal, C.-E. 2007. "The calibration approach in survey theory and practice". *Survey Methodology*, Volume 33, N. 2: 99-119. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488-eng.pdf?st=YEcKmKFm.

Särndal, C.-E. 1980. "On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling". *Biometrika*, Volume 67, N. 3: 639-650. https://doi.org/10.2307/2335134.

Särndal, C.-E., and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. Hoboken, NJ, U.S.: John Wiley & Sons.

Särndal, C.-E., B. Swensson, and J.H. Wretman. 1989. "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total". *Biometrika*, Volume 76, N. 3: 527-537. https://doi.org/10.2307/2336118.

Särndal, C.-E., and I. Traat. 2011. "Domain Estimators Calibrated on Information from Another Survey". *Acta et Commentationes Universitatis Tartuensis de Mathematica*, Volume 15, N. 2: 43-60.

Singh, A.C. 1996. "Combining information in Survey Sampling by Modified Regression". In *Proceedings of the Section on Survey Research Methods*, Volume 1: 120-129. Alexandria, VA, U.S.: American Statistical Association.

Singh, A.C. 1994. "Sampling design based estimating functions for finite population totals. Invited paper". In *Abstracts of the Annual Meeting of the Statistical Society of Canada*, *Banff*: 48. Calgary, Alberta, Canada, 8-11 May 1994.

Singh, A.C., B. Kennedy, and S. Wu. 2001. "Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design". *Survey Methodology*, Volume 27, N. 1: 33-44. https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5852-eng.pdf.

Singh, A.C., and C.A. Mohl. 1996. "Understanding Calibration Estimators in Survey Sampling". *Survey Methodology*, Volume 22, N. 2: 107-115. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996002/article/2973-eng.pdf?st=B9mYnII6.

Wolter, K.M. 2007. *Introduction to Variance Estimation*. New York, NY, U.S.: Springer, Series in Statistics. https://doi.org/10.1007/978-0-387-35099-8.

Wright, R.L. 1983. "Finite Population Sampling With Multivariate Auxiliary Information". *Journal of the American Statistical Association*. Volume 78, N. 384: 879-884. https://doi.org/10.2307/2288199.

Zardetto, D. 2015. "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys". *Journal of Official Statistics*, Volume 31, N. 2: 177-203. https://doi.org/10.1515/jos-2015-0013.