Optimal Design of a Master Sample: the Case of the Household Surveys in the Republic of Moldova

Giulio Barcaroli, Loredana Di Consiglio, Alessio Guandalini, Marco Dionisio Terribili [1]

Abstract

This article describes the results of the activities carried out in the framework of the "Project of Technical Assistance to Support the National Bureau of Statistics (NBS) of the Republic of Moldova", in particular those regarding the Activity A14 (Design of Master Sample): the identification of the household surveys that will make use of the new master sample and, for each of them, a newly optimised sample design; the calculation of the total amount of final sampling units to be selected from the master sample in its whole life span; the selection of Primary Selection Units (PSUs); the definition of two different possible configurations of the master sample, depending on the coordination or not of the PSUs samples of the different surveys; the evaluation of pros and cons of the two solutions, taking into account the constraints derived by the current organisation of the NBS data collection network. The above activities have been carried out making use of the methodologies implemented in three generalised software (ReGenesees, R2BEAT and FS4) developed and used by the Italian National Institute of Statistics - Istat.

Keywords: Statistical cooperation, Master Sample, two-stage sample design, optimal allocation, social surveys, sample coordination, R2BEAT, ReGenesees.

1 Giulio Barcaroli (gbarcaroli@gmail.com); Loredana Di Consiglio (diconsig@istat.it); Alessio Guandalini (alessio.guandalini@istat.it); Marco Dionisio Terribili (terribili@istat.it), Italian National Institute of Statistics.

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[2]

Starting from December 2019 to June 2020 several cooperation missions were carried out in the framework of the Project of Technical Assistance to Support the National Bureau of Statistics of the Republic of Moldova, within the Activity A14 (Design of master sample)[3]. Activity A14 includes the following:

1. Identification of household surveys that make use of the master sample; then, determination for each survey, the initial allocation of units planned for each wave of the master sample;
2. Calculation of the total amount of units to be selected from the master sample in its whole life span;
3. Selection of Primary Selection Units (PSUs);
4. Refinement of sample designs for all surveys and selection of samples;
5. Improvement and documentation of the estimation procedures and sampling design;
6. Training on the ReGenesees and its use at the NBS, or other software tools;
7. Assistance to NBS in the production of a complete methodology.

The in-use version of the master sample was designed in 2018 and was intended to work during the years 2019 and 2020. Due to the exhaustion of included PSUs, the new version of the master sample had to be designed in the first half of 2020 and implemented in the second half, ready to be used at the beginning of 2021 for a period of four or five years.

The 2019-2020 master sample was implemented by selecting 150 PSUs in eight different strata, resulting from the cross-product of the four statistical regions (North, Center, South and Chisinau) and the PSUs rural/urban feature. There is an important difference between PSUs in the stratum Chisinau urban and in the other strata:

2 The authors wish to sincerely thank the entire NBS staff and in particular Lilian Galer, Olga Moraru and Galina Ostapenco for their precious collaboration during the activities of the cooperation project, without whom the results obtained would not have been reached.

3 EU funded Project ENI/2019/406-262, "Technical Assistance to Support the National Bureau of Statistics of the Republic of Moldova".

- in Chisinau urban the PSUs correspond to the EAs in the Population and Housing Census 2014 (PHC 2014), for a total of 1,718 PSUs;
- in the rest of the country the PSUs are an aggregation of EAs from PHC 2014 in order to reach a minimum size of 420 households, for a total of 1,742 PSUs.

The first stage sample size was equal to 150 PSUs, with a simple random sampling in the Chisinau urban stratum, and a PPS (probability proportional to size) sample in the other strata. For each selected PSU a complete enumeration of households was carried out, and a simple random sample of them was drawn to guarantee the interviewers (150, one for each selected PSU) a predetermined number of interviews for the Labour Force Survey and the Household Budget Survey.

In the Chisinau urban stratum the sample PSUs for LFS and HBS do not match, whereas in the rest of the country the same sample PSUs are used both for LFS and for HBS, taking care that the sampled households do not overlap. As for the rotation schemes, to implement the 2-2-2 scheme of the LFS, in the Chisinau urban stratum only, each quarter, one-fourth of the sample -the entering LFS panel- is renewed and new PSUs are used for it; after 6 quarters - when the cycle is completed - each LFS PSU is removed.

Also in the Chisinau urban, PSUs used by HBS are removed once used for a round of the survey. In the other strata, there is no rotation of LFS PSUs. For both, LFS and HBS PSUs are removed once exhausted.

The multivariate optimal allocation carried out in 2018 was based on the following steps:

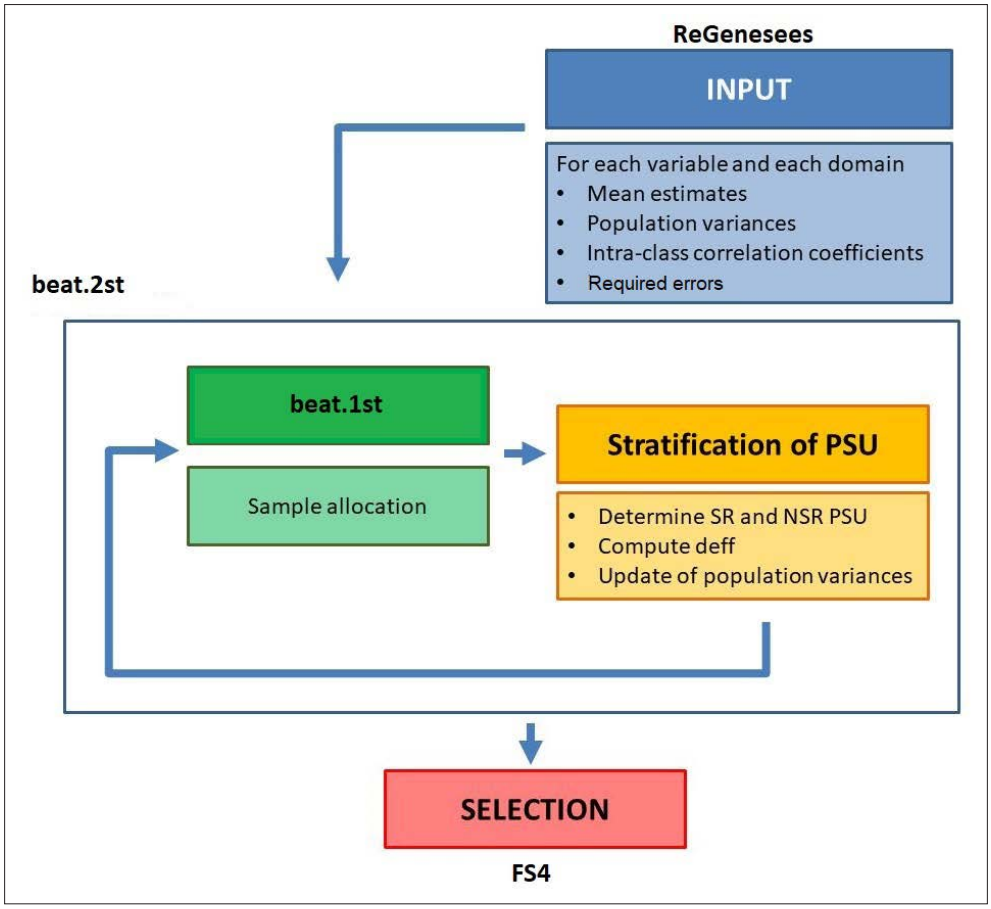
- the allocation of SSUs (households) in the 8 strata was based on the variability in these strata of the LFS and HSB target variables;
- taking the total amount of the PSUs as fixed (determined by the number of interviewers, *i.e.* 150), and setting a minimum number of 30 and 24 households per PSU respectively for LFS and HBS, the allocation of these 150 PSUs was determined by applying the methodology implemented in the R package R2BEAT that will be illustrated in the following sections.

According to the Inception Report of the Project, in December it was decided to perform the new master sample design taking into account not only the aforementioned surveys (LFS and HBS), but all those planned to be carried out during the life span of the MS. Consequently, beyond these two ones, two more were taken into consideration: the Energy Consumption and the Domestic Violence surveys. Differently from LFS and HBS, which are carried out quarterly, the two additional surveys are to be carried only once in a date to be still defined, but included in the interval 2021-2024.

So, to determine the total amount of Secondary Stage Units (SSUs) that are needed to conduct the four surveys and therefore that the new master sample is expected to satisfy over its lifetime, new sample designs for each survey have been produced. For each survey the following steps have been performed (see Figure 1):

1. previously available rounds of the surveys have been taken into consideration;
2. for each round, sampling estimates have been calculated together with the information needed by the next steps, *i.e.* strata variances, design effects and estimator effects: this has been done using the Istat software ReGenesees;
3. taking the output of the previous step as input, optimal allocation in terms of PSUs and SSUs has been obtained by using the R package R2BEAT, developed by Istat (Fasulo *et al.,* 2020);
4. the first stage selection of PSUs has been performed by using the R package FS4 (developed by Istat): this package stratifies further the PSUs in terms of their size and classifies them as Self-Representative (SR, *i.e.* with only one unit, with inclusion probability equal to 1), and Non-Self-Representing (NSR); in the latter strata, PSUs are selected with a PPS scheme.

Figure 1 - General framework of the work performed for each survey



Source: Our processing

The sample designs of the four surveys have been optimised concerning these important criteria:

1. the total amount of final units (households) have been minimised under the precision constraints (expected coefficients of variation not exceeding given values) specified by NBS, but tentatively modified in order to get feasible solutions;
2. in doing that, total non-response rates observed in previous survey rounds have been taken into account and the planned sample sizes have been inflated accordingly (oversampling);

3. First and second stage probability of inclusions are such that the overall inclusion probabilities are almost constant for each unit, resulting in self-weighted samples.

Each selected PSU may or may not be used by more than one survey: in the former case we say there is no coordination of PSUs (the only overlap is by chance), in the latter case we say there is coordination of PSUs.

In our specific case, coordination of samples has been obtained by considering LFS PSUs sample as pivotal, and coordinating with it the HBS sample (and then the Domestic Violence and Energy Consumption surveys), favouring the PSUs with a higher number of households. Anyway, positive coordination has never been applied for PSUs in the Chisinau urban stratum, due to the very low average number of households.

The master sample obtained without coordination is characterised by the highest number of PSUs, both in terms of selected ones and joint use in each quarter. In contrast, the master sample obtained by coordinating PSUs is characterised by the lowest number of initial PSUs, but a higher rate of exhaustion.

This paper is organised as follows: in section 2 an illustration of methodologies employed and of tools used is given; in Section 3 the sample design for each one of the four involved surveys is reported; in Section 4 the methodological approach to coordinate PSUs is described and the results obtained are illustrated; in Section 5 the two scenarios of implementation of the master sample (with and without coordination of PSUs) are illustrated by means of simulating the master sample use in the life span 2021-2025. Some conclusions are reported in the Section 6.

2. Methods and tools

This section briefly describes the main statistical methods that are needed for the determination of a sampling design and the software tools used for their application as illustrated in Figure 1.

2.1 Sampling estimates and sampling errors (ReGenesees)

The first step to study the new allocation for a sample survey is, when previous occasions of the survey are available, to estimate means and variances for its main target variables.

In order to compute sampling estimates and their precision, the R package ReGenesees[4] can be used. The package is the prneoduct of a long term Istat project.

The package enables computation of the calibration estimators and their variance estimation (see Deville & Särndal, 1992; Särndal, 2007) based on the linearisation method (see Woodroof, 1971). The package also contains utilities for calculating population totals, that are used for calibration estimation, or to aggregate strata, in case of unique sample units within strata. It also allows estimating precision measures of complex indicators that can be linearised.

The main functions are the following:

- e.svydesign describes the adopted sampling design;
- e.calibrate computes the calibration weights (Deville and Särndal, 1992);
- svystatTM computes weighted estimates for totals and means using suitable weights depending on the class of design;
- svystat is a general function to compute weighted estimates of totals, ratios, quantiles, *etc.*

See Zardetto (2015) for details on the use of the software. An illustration of its use is given with some detail in the following sections 3.1, 3.2, 3.4, 7 where the previous functions are specified accordingly to the contexts they are applied.

4 The software is not on R-CRAN, but can be downloaded from the Istat website: https://www.istat.it/en/ methods-and-tools/methods-and-it-tools/process/processing-tools/regenesees

2.2 Multivariate allocation in two-stage sampling design (R2BEAT)

Two-stage sampling design with stratification of PSUs are very common for household surveys in Official Statistics. However, their sample allocations are usually a non-easy task, because the household surveys have multi-domains and multi-purpose objectives, so they have to provide accurate estimates for different variables and different domains (*e.g.* geographical areas such as national, regional, *etc.*).

The whole sample size of a survey, n, is often an exogenous detail because it is usually defined by budget and, sometimes, also by logistic constraints.

Instead, the allocation of the sample among the strata ($h = 1, \ldots, L$), generally at the lower level at which estimates are required, can be defined in different ways (mainly with uniform, proportional or optimal allocation). In Official Statistics, the optimal allocation is the most widely used method, especially because information on the size of the strata and the variance of target variables in the strata can be obtained. In particular, the variances of the target variables or at least of proxy variables for each stratum can be computed, for instance, from register data or previous occasions of the same survey.

Since August 2019, a package called R2BEAT, developed by Istat, is available on the R-CRAN repository (Fasulo *et al.*, 2020). R2BEAT easily manages all the complexity due to the optimal sample allocation in two-stage sampling design and provides several outputs for evaluating the allocation[5]. Its name stands for R "to" Bethel Extended Allocation for Two-stage. It is an extension of another open-source software called Mauss-R (Multivariate Allocation of Units in Sampling Surveys), implemented by Istat researchers[6]. Mauss-R determines the optimal sample allocation in multivariate and multi-domains estimation, for one-stage stratified samples. It extends the Neyman (1934) - Tschuprow (1923) allocation method to the case of several variables, adopting a generalisation of Bethel's proposal (1989).

5 To complete the suite of tools developed by Istat in order to cover the stratified sample design, we cite SamplingStrata (Barcaroli *et al.*, 2020), which allows to jointly optimise both the stratification of the sampling frame and the allocation, still in the multivariate multidomain case (only for one-stage designs), and MultiWay. Sample. Allocation, that allows determining the optimal sample allocation for multi-way stratified sampling designs and incomplete stratified sampling designs. See https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/ multiwaysampleallocation for more details.

6 https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r.

The main idea for optimal allocation is that strata with larger size and larger variability with respect to the target variables need a larger sample size to provide better estimates. Since, in our case, data from previous surveys occasions were available, an optimal allocation could be deployed.

R2BEAT develops Bethel's proposal but, besides, enables the determination of optimal sample allocation for two-stage stratified samples. The R2BEAT version used in this paper includes three functions:

- beat.1st that computes multivariate optimal allocation for different domains in one-stage stratified sample design;
- beat.2st that computes multivariate optimal allocation for different domains in two-stage stratified sample design, considering the design effect;
- beat.cv that, given a multivariate optimal allocation, calculates the coefficient of variation.

In our specific case, because all the considered surveys will implement a two-stage sample design, only the function beat.2st was used. Before using this function, a preliminary step was needed. In fact, it has been necessary to define some parameters for the function (see the package documentation for the input formats)[7]. In particular:

1. population size of each stratum in terms of households or individuals depending on the aim of the survey (N);
2. strata to be censused (CENS);
3. cost of interviews per stratum (COST);
4. number of PSUs to be selected in each stratum (MINCS);
5. average dimension of SSUs in each stratum (DELTA, Δ);
6. minimum number of SSUs to be interviewed in each selected PSU (MINIMUM);
7. mean of each target variables in each strata (M, $\hat{p}$ or $\hat{\bar{x}}$);

7 The latest version available on R-CRAN includes two functions that, starting from survey data, provide the inputs as the other functions require.

8. estimated population standard deviation of each target variable in each stratum (S, $\hat{S}$);
9. intra-class correlation coefficients for each target variable in each stratum (RHO, $\hat{p}$);
10. estimator effect for each target variable in each stratum (EFFST, $\widehat{effst}$).

To better understand the optimal allocation, a few words on points 7-10 are necessary. In particular, the mean can be estimated using the function svystatTM of the package ReGenesees. For the estimated population standard deviation, a distinction between a binary and a quantitative variable is needed. In the case of binary variables, the estimated population standard deviation of variable *i* is equal to

$$\hat{S} = \sqrt{\hat{p}_i \times (1 - \hat{p}_i)} \tag{1}$$

where $\hat{p}_i$ is an estimate of the proportion of individuals with a given characteristic in the population strata, which is the mean of a binary variable. While, in the case of a quantitative variable, it is equal to

$$\hat{S} = \sqrt{\hat{M}_i^2 - \hat{\bar{Y}}_i^2} \tag{2}$$

where $\hat{M}_i^2 = \frac{\sum_{k \in s} y_{ik}^2 \, w_k}{N}$ and $\hat{\bar{Y}}_i = \frac{\sum_{k \in s} y_{ik} \, w_k}{N}$ are the quadratic mean and the arithmetic mean estimated on sample data of the previous occasion(s) of the survey, respectively, with y_{ik} the value of the target variable i observed on the unit *k*, w_k the corresponding sampling weight and *N* the population size.

A crucial statistic for the optimal sample allocation in a two-stage sampling design is the intra-class correlation coefficient, $\hat{p}$. Due to logistics, in household surveys, the sample is not wide-spread in the whole country, but often "clusterised" in PSUs and, then, in SSUs (clusters of family members). If the units of clusters are too similar to each other, it is not efficient to collect too many units from the same cluster. Anyway, sometimes a small loss in efficiency of the estimates is accepted because it is paid off by a gain in the organisation of the survey and, moreover, by cost reduction.

Therefore, to optimise the sample size, it is important to compute the intra-class correlation coefficient, because it provides a measure of data clustering

in PSUs and SSUs. In general, if the value of $\hat{p}$ is close to 1 the clustering is high and it is convenient to collect only a few units in the cluster. On the contrary, if $\hat{p}$ is close to 0, the collection of units from the same cluster does not affect the efficiency of the estimates.

To compute $\hat{p}$, another important statistic must be introduced: the design effect (DEFF, *deff*). The design effect measures how much the sampling variance under the adopted sampling design is inflated with respect to a simple random sample (*srs*), with the same sample size. In formula:

$$deff(\hat{\bar{Y}}_i) = \frac{\text{var}(\hat{\bar{Y}}_i)_{des}}{\text{var}(\hat{\bar{Y}}_i)_{srs}} \quad (3)$$

$$= 1 + \hat{\rho}_i\,(b-1)) \quad (4)$$

where *b* is the average cluster (*i.e.* PSU) size in terms of the final sampling units and $\hat{p}_i$ the intra-class correlation within the cluster (PSU) for the variable *i*.

The package R2BEAT takes into account a more general expression of *deff*. This expression refers to a typical situation for household surveys in which PSUs are assigned to Self-Representing (SR) strata, that is they are included for sure in the sample, or to Not-Self-Representing (NSR) strata, where they are selected by chance. In practice, this assignment is usually performed by comparing their measure of size (MOS) with respect to the threshold:

$$\lambda = \frac{\bar{m}\,\Delta}{f} \quad (5)$$

where $\bar{m}$ is the minimum number of SSUs to be interviewed in each selected PSU (MINIMUM), $f = n/N$ is the sampling fraction and Δ (DELTA) is the average dimension of the SSU in terms of elementary survey units. Then DELTA must be set equal to 1 if, for the survey, the selection units are the same as the elementary units (that is, household-household or individuals-individuals), whereas it must be set equal to the average dimension of the households if the elementary units are individuals, while the selection units are the households.

PSUs with MOS exceeding the threshold are identified as SR, while the remaining PSUs are NSR.

Then the extended expression of *deff* is

$$deff(\hat{\bar{Y}}_i) = \frac{N_{SR}^2}{n_{SR}}(1 + (\hat{\rho}_{i,SR}\,(b_{SR} - 1)) + \frac{N_{NSR}^2}{n_{NSR}}(1 + (\hat{\rho}_{i,NSR}\,(b_{NSR} - 1)) \quad (6)$$

where, for *SR* and *NSR* strata,

- N_{SR} and N_{NSR} are the population sizes;
- n_{SR} and n_{NSR} are the sample sizes;
- $\hat{p}_{i,SR}$ and $\hat{p}_{i,NSR}$ the intra-class correlation coefficient for the variable *i*;
- b_{SR} and b_{NSR} are the average PSU size in terms of the final sampling units.

Of course, if there are no *SR* strata the expression (4) recurs. The design effect is equal to 1 under the *srs* design and increases for each additional stage of selection, due to intra-class correlation coefficient which is, usually, positive. It can be computed using the function svystatTM from ReGenesees, setting up the parameter deff=TRUE.

The intra-class correlation coefficient for *NSR* can be derived from the expression of *deff*, that is

$$\hat{\rho}_{i,NSR} = \frac{deff_i - 1}{b_{SNR} - 1}. \quad (7)$$

While it is not necessary to compute the intra-class correlation coefficient for SR strata because just one PSU is selected and the intra-class correlation is 1 by definition.

The last statistic to be defined is the estimator effect (EFFST, *effst*), that is, how much the sampling variance of the applied estimator under the adopted design is inflated or deflated with respect to the HT estimator, on the same sample. The *effst* is equal to

$$effst(\hat{\bar{Y}}_i)_{st} = \frac{\text{var}(\hat{\bar{Y}}_i)_{st}}{\text{var}(\hat{\bar{Y}}_i)_{HT}} \quad (8)$$

It is an optional parameter for R2BEAT, but it useful to take into account, from the allocation phase, of the impact on the estimates of a different estimator other than the Horvitz-Thompson estimator (*HT*). Indeed, the most applied estimator for the households survey is the calibrated estimator that, through the use of auxiliary variables, provides better estimates than HT.

Then, a reduction of the final sample size can be expected when applying the calibration estimator in place of the HT.

The optimal allocation is defined by R2BEAT solving the minimum optimisation problem

$$\begin{cases} C = \min \\ \text{CV}\left(\hat{\bar{Y}}_{i,h}\right) \leq \delta\left(\hat{\bar{Y}}_{i,h}\right) & \begin{matrix} i = 1,\ldots,J \\ h = 1,\ldots,L \end{matrix} \end{cases},$$

where C is the global cost of the survey (if COST is equal to 1 in all the strata, $C = n$) and $\text{CV}\left(\hat{\bar{Y}}_{i,h}\right)$ is the relative error we expect to observe estimating the Y_i variable in the stratum h (expected errors), that must be less than or equal to the precision constraints defined by the user. The solution is obtained with an iterative algorithm. Then, $\hat{S}$ (see, *e.g.*, expression (1) and (2)) is multiplied by the new design effect and the estimator effect and a new allocation is computed. The algorithm stops when the difference between two consecutive iterations is lower than a predefined threshold. At each step, an allocation is provided and the design effect is updated following the expression (6).

The package R2BEAT provides several outputs that help the evaluation of the allocation. For further details see the manual of the package available on-line.

It is important to point out that, as stated at the beginning of this section, the n is usually given. Then, to find the optimal allocation it is necessary to tune the precision constraints until the desired sample size is matched.

2.3 First stage selection (FS4)

FS4 is an open-source generalised software developed by Istat for stratification and selection of the PSUs in a two (or more) stages sampling design[8].

The function carries out the stratification of the PSUs, for each estimation domain, according to a size threshold (5). PSUs with a measure of size exceeding the threshold are identified as Self-Representing units, SR (see also previous section). The remaining Not Self-Representing units PSUs, NSR,

8 https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/fs4.

are ordered by measure of their size and divided into strata whose sizes are approximately equal to the threshold multiplied by the number of PSUs to be selected in each stratum (MINCS). In this way, strata are composed of PSUs having as homogeneous as possible size. Next, the selection of a fixed number of PSUs per stratum is carried out using Sampford's method (unequal probabilities, without replacement, fixed sample size), implemented by the UPsampford function of the R package sampling.

The package FS4, through the function StratSel, stratifies and selects the PSUs to be sampled. Furthermore, the output provides the first order inclusion probability for each PSU, that is, for the first selection stage. The inclusion probability for the second stage can be easily obtained by dividing the number of SSUs to be selected in the PSU by the measure of size of the PSU. Then, the design weights for each unit in the sample are equal to the inverse of the product of the first-order and the second stage inclusion probabilities.

The inclusion probabilities of the first and second stage of the sampling scheme are such that the overall inclusion probabilities are almost constant for each unit in the sample, resulting in a self-weighting samples.

3. Sampling design of selected surveys

3.1 Labour Force Survey

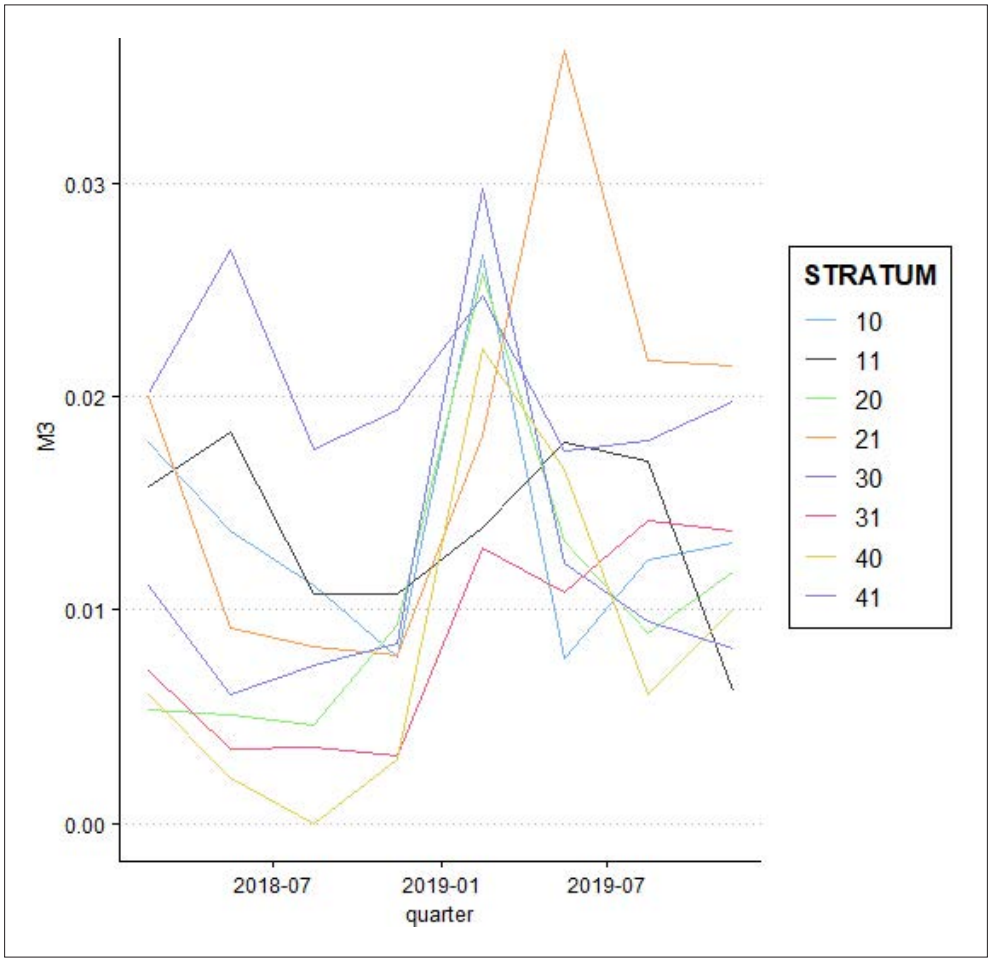
The new allocation of the Labour Force Survey (LFS) was determined on the basis of the main target variables, as established by NBS:

1. non-active (inactive);
2. employed (oc);
3. unemployed (som).

As already described in Section 2.2, the information from the 2018-2019 survey waves is useful for optimising the allocation. In detail, the estimates of the ratios of the target variables: namely non-active, employed, and unemployed population and their sampling variances, as well as the design effects (3) and the estimator effects (8) for the target estimators in each of the 8 available waves, was evaluated to provide the necessary input to R2BEAT, see Section 2.2. The R package Regeneeses was employed for this preliminary activity (Section 2.1).

To determine the sample allocation, different scenarios were analysed with different input information; indeed, having 2 years of LFS survey data available, one can use estimates from the last year only (*i.e.* the most recent information) or otherwise an average of the estimates from all surveys in the period 2018-2019. The latter alternative was chosen because it was more conservative in terms of variance. This was because in 2019 there was an increase in the unemployment rate in some areas, which therefore resulted in a lower variance in the latest period. Figure 2 depicts the unemployment rate for all the strata in the 8 quarters.

Figure 2 - Unemployment rate 18-19



Source: Our processing

Another important feature characterising the different alternative scenarios was the minimum number of interviews in each PSU. This parameter has an impact on the allocation via the design effect (see Equation 4).

Taking into consideration the differences in the rates in the two years, as mentioned above, the constraints on the required coefficients of variation of the estimators are "averaging" the current values. For the determination of sample size and its allocation within strata, the R package R2BEAT (see Section 2.2) was employed.

Let us see more in detail the main steps for determining the key input element for the allocation with R2BEAT. First, it was necessary to estimate the mean (rates) of the target variables, Following Section 2.1 to proceed in evaluating these quantities, we have to define the applied sampling design for the available data in the data frame lfs:

```
library(Regenesees)
des_19 <- e.svydesign(data = lfs, ids = ~ centr + ident_HH,
strata = ~ strat_new1,
weights = ~ w_corr, check.data = TRUE)
```

where ids is the identifier of the selection units, centr the PSU and ident_HH the household identifier, the sampling weights in the function are w_corr, the already corrected weights for non-response. Finally, strata represents the stratification of the sampling design. LFS weights are then calibrated to known totals for regions, urban or rural area, and age-sex groups. The calibration step in ReGenesees is easily carried out with the following function:

```
cal_18 <- e.calibrate(design=des_19, df.population=popfill,
calmodel= ~ reg + urb +gr_vr - 1,
calfun= "linear", bounds =   c(0.1,3))
```

Once the calibrated weights are calculated, the estimates and sampling errors can then be simply evaluated with the following function:

```
est_2018_reg_urb <- svystatTM(cal_18,y=~total+inact+oc+som +activ,
by = ~ reg:urb, estimator="Mean",
vartype=c("se"), deff=TRUE)
```

where by=~reg:urb provides the estimates for the most detailed domain, that is the strata for which the allocation is determined. Estimates can be evaluated at each desired level, specifying the value for by, appropriately.

The previous steps were applied to all 8 available waves and then the results were averaged to provide the input to R2BEAT.

The design effects in (4) of the two-stage design was evaluated on a new stratification composed of three classes:

- rural;

- urban;
- Chisinau.

This stratification was chosen based on the sampling design and the different features of rural and urban areas.

Moreover, only waves relating to 2019 were considered for design effect calculation, since the sampling design in terms of primary sampling units is different from the previous year and assuming the PSUs of the new sample design will have the most recent size.

Function svystat, similarly to svystatTM, evaluates the estimates and the design effects, in this case the option by requires the new specifically defined stratification stra1.

```
deff <- svystat(des_19,
kind ="TM",
estimator="Mean",
y= ~ inact+oc+som+activ,
by=~stra1,
deff=TRUE,
forGVF=TRUE)
```

Similarly, the estimator effect is the ratio between the variance of the calibrator estimator and the variance of the HT estimator under the applied sampling design. This component helps to determine the sampling size, taking into account that a calibration estimator is used instead of the HT estimator. As mentioned in Section 2.2, effst is an optional parameter for the software r2beat that can be used to consider the actual calibration estimator instead of the HT in defining a more efficient allocation.

Similarly to those seen above, the following instructions provide the ratio between $var(CAL)_{des}$ and $var(CAL)_{srs}$ in variable DEFF of the data frame effst:

```
effst <- svystat(cal_18,
kind ="TM",
estimator="Mean",
y= ~ inact+oc+som+activ,
by = ~ stra1,
deff=TRUE,
forGVF=TRUE )
```

The estimator effect can then be obtained by removing the design effect from the previous result.

```
effst$EFFST <- effst$DEFF/deff$DEFF
```

Other relevant quantities needed for allocation are the bNAR, average dimension of the PSU and b_{AR}, the average size of the families.

Finally, the input for the LFS allocation is:

```
strat=data.frame(STRATUM=as.numeric(N[,1]),
N=N[,2],M1=M1[,2],M2=M2[,2],M3=M3[,2])
strat$S1=sqrt(strat$M1*(1-strat$M1))
strat$S2=sqrt(strat$M2*(1-strat$M2))
strat$S3=sqrt(strat$M3*(1-strat$M3))
strat$CENS=0
strat$COST=1
strat$DOM1=1
strat$DOM2=substr(strat$STRATUM,1,1)
strat$DOM3=substr(strat$STRATUM,2,2)
strat$DOM4 <- strat$STRATUM
```

where N is the size of the stratum in terms of individuals, M1, M2, and M3 the average of values over the 8 waves of the three target variables and S1, S2, S3 their population variances (see also (1) for the applied expression in their evaluation), *DOM 1* the national domain, *DOM 2* the region, *DOM 3* the urban-rural classification and *DOM 4* the finer domain, *i.e.* the stratum. The design features are described in the object des:

```
des<-data.frame(STRATUM=des[,"STRATUM"], STRAT_MOS=des[,"PSU_MOS"],
DELTA=des[,"b_ar"],
MINIMUM=c(rep(36,6),rep(36,2)))
```

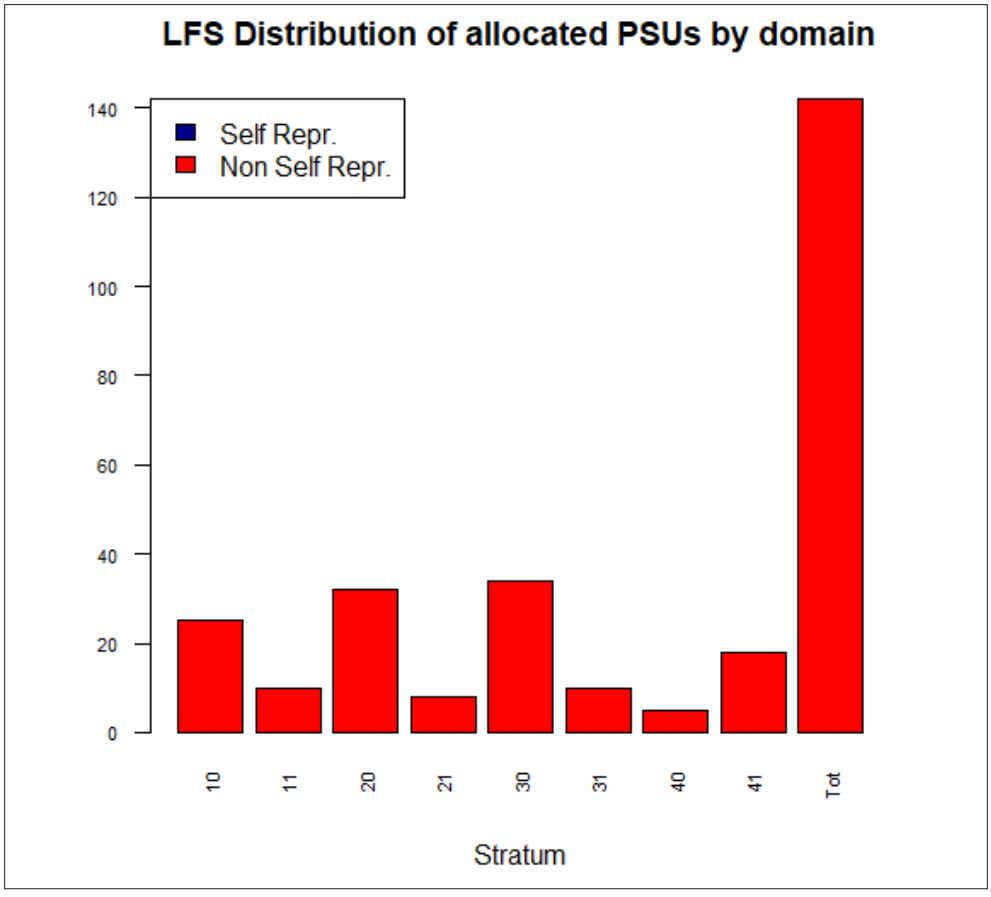
where STRAT_MOS is the PSU size in terms of individuals, DELTA is the average size of the secondary units (households) in terms of individuals, !MINIMUM! is, as recalled above, the min- imum number of households to be interviewed in each PSU. The sample size and its allocation among strata are reported in Table 1. An illustration of PSUs allocation is in Figure 3.

Table 1 - LFS allocation

Region	Urban	EA	Individuals	HH	Oversampling
1	0	25	2282	910	978
1	1	10	836	364	434
2	0	32	2912	1157	1252
2	1	8	653	284	344
3	0	34	3041	1325	1207
3	1	10	816	355	430
4	0	5	444	176	253
4	1	18	1433	646	1006
		142	12417	5101	6022

Source: Our processing

Figure 3 - LFS allocation among strata



Source: Our processing

Table 2 reports the expected errors of the three target variables for the different domains with the proposed allocation.

The corresponding yearly errors can be approximated by multiplying the previous errors by the following quantity:

$$\frac{1}{4}(1+\frac{3}{4}\rho_1+\frac{1}{8}\rho_3)$$

Table 2 - LFS Expected CVs

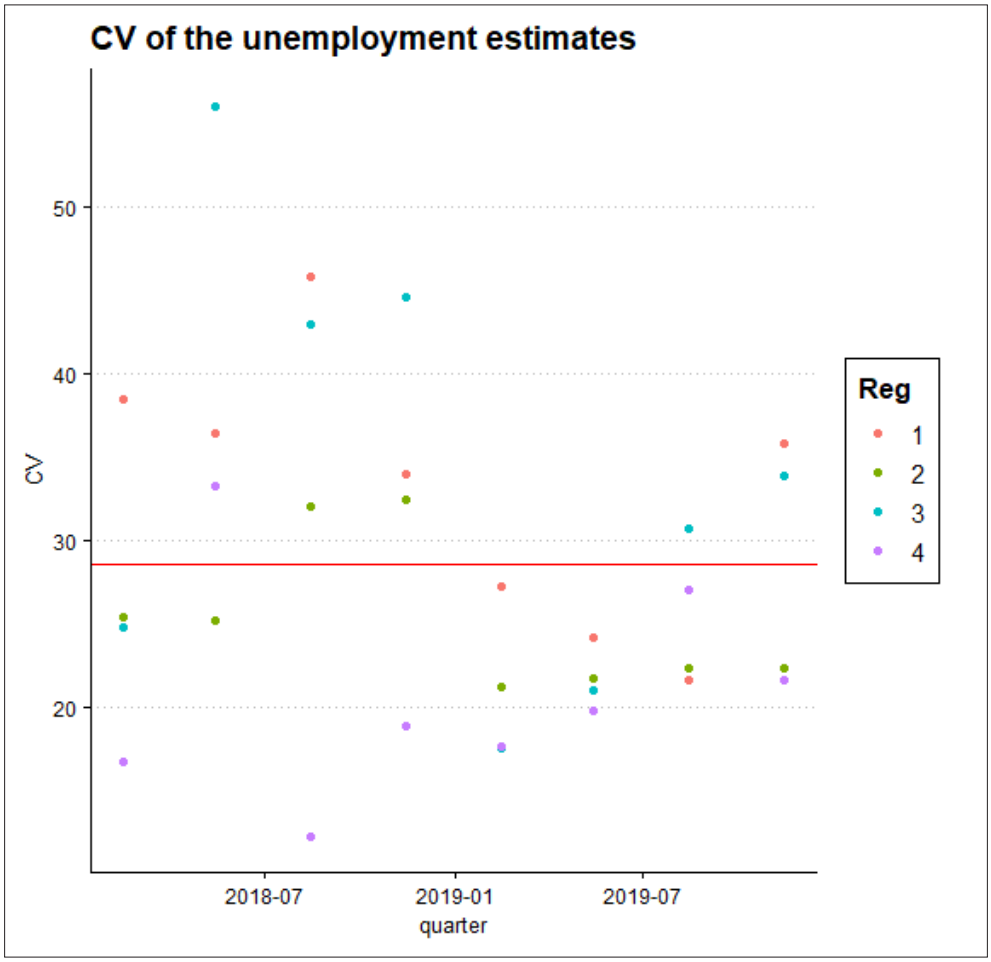
Type	Dom	Active	Empl	Unemp
National	1	0.0127	0.0326	0.1420
Region	1	0.0278	0.0620	0.2795
Region	2	0.0228	0.0745	0.2860
Region	3	0.0206	0.0673	0.2860
Region	4	0.0264	0.0437	0.2571
Urban	1	0.0170	0.0496	0.1965
Urban	0	0.0165	0.0325	0.2013

Source: Our processing

where ρ_1 is the correlation between quarters at one lag and ρ_3 is the correlation between quarters at three lags.

Figure 4 reports the observed errors of the unemployment rate estimates in the 8 waves of 2018/2019, while the horizontal red line represents the input constraints for the CV.

Figure 4 - LFS CV at regional level in 2018/2019 and required errors



Source: Our processing

Note that the new Eurostat regulation on social surveys would require tighter constraints, however the needed sample size would be much higher than the current one and for this reason, it would be difficult to apply.

3.1.1 A Proposal of gradual introduction of the new LFS sample

Instead of replacing the current sample with the new LFS sample all at once, we propose a gradual introduction that takes into account the LFS (2-2-2) panel scheme. In fact, since each quarter a panel (1/4 of the total sample) is replaced, we suggest that the new sample enters only for a portion, as a replacement of the exiting panel.

Let us denote the new panels from the new allocation with letters a-f and the panels of the current running panels with n-q, and with 1-4 the waves, *e.g.* for quarter T1-21 *a*1 is the entering panel selected with the new LFS allocation, whereas *n*4 is panel *n* at its 4th survey occasion, *o*3 panel *o* at its 3rd survey occasion, and so on. Similarly, for the following quarters. At quarter T2 22, all four panels are selected with the new allocation and the transition from the old to the new sample is complete, as follows:

T1	21	a1						n4	o3			r2
T2	21	a2	b1						o4	p3		
T3	21		b2	c1						p4	q3	
T4	21			c2	d1						q4	r3
T1	22	a3			d2	e1						r4
T2	22	a4	b3			e2	f1					

The proposed gradual introduction of the new LFS allocation has the main advantage of keeping the composition of the sample in each quarter in the 4 panels of households at different survey occasion (first, second, third, and fourth interview), thus preventing that for some quarters the composition of the sample is altered.

In fact, if we introduce all the new LFS sample in T1 21, for that quarter it will be composed of all households at their first interviews.

A well-known problem of panel surveys is differences in responses according to the number of occasions the units have been already interviewed, therefore a time-series break might occur if the composition is altered.

On average, this scheme does not require a much higher sample size during the transition between the old and new sample, since the two allocations are

similar in terms of total sample size. However, there might be some issues due to different stratification of the current and new allocation and the global allocation can then be different from both. Moreover, as the panels come from different samples, the number of PSUs may be higher depending on how many PSUs should be kept from the previous allocation to maintain the panels on the last survey occasions.

Let us recall the rotation mechanism of PSU in different strata:

a. Chisinau: for the current design there is a rotation of the PSUs in Chisinau. The proposed scheme could be easily applied to this stratum. In fact, in Chisinau, after 6 consecutive quarters (*i.e.* completing the life-span of a panel) the PSU is removed from the survey. Each quarter 1 fourth of all PSUs is renewed when a new panel is entering the sample (*i.e.* 1/4 of the quarter sample), in this case, the new panel (and its corresponding PSUs) can be easily selected according to the new survey introducing the new allocation.

   For Chisinau, the number of allocated PSUs will be moderately larger depending only on the sampling stratification: following the previous scheme the entering panels *a*1, *b*1 and so on, will be selected in PSUs replacing the current ones.

b. Remaining strata - no rotation of PSUs is planned unless the PSU is exhausted. For these strata, extra PSUs would be needed for all the first year to account both for the closing panels of the current survey and the entering panel of the new allocation. To avoid this additional number of PSUs, the new PSU sample could be positively coordinated with the current one. However, as many PSUs are currently already almost fully investigated, this option could be difficult to apply in practice.

Note that the transition by itself does not increase the risk of exhaustion of the current PSUs. The households have been already selected, *e.g.* panel p4 in T3 21 has been already selected in quarter T2 20, similarly, all the panels n-r are to be selected by the end of 20 for the conduction of the current LFS scheme.

Finally, estimates from the two different sampling schemes (the current and the one) running at the same time with different size can be obtained

either (a) by combining estimates from the different sources, for example, weighting the estimates with their sampling variance, or (b) by pooling data at the micro-level, for example by simply re-weighting the different sample weights to the known totals.

3.2 Household Budget Survey

Target variables of the Household Budget Survey have been defined by NBS as:

1. total pro-capite expenses (TE);
2. food pro-capite expenses (FE);
3. total pro-capite revenues (TR).

And the related precision constraints:

Domain	CV(TE)	CV(FE)	CV(TR)
National	0.02	0.01	0.04
Regional	0.04	0.02	0.06
Urban/Rural	0.02	0.02	0.04

The 8 rounds of the survey in 2018 and 2019 were available. Due to the high rise in non-response rates occurred in 2019, only this year has been retained, as we wanted to be conservative by considering the worst situation.

For each quarter of 2019, estimates of the three target variables have been calculated by using ReGenesees.

```
library(Regenesees)
des_19rev_2 <- e.svydesign(data = hb19rev_2, ids = ~ cod_ter + id_hh,
strata = ~ strat_new1,
weights = ~ wcor, check.data = TRUE)
```

Note that initial weights are the ones already treated to handle non-response.

Calibration estimates have been obtained by using known totals (total population by region and rural/urban, number of children and number of retired males and females):

```
cal_19rev_2 <- e.calibrate(design=des_19rev_2, df.population= popfill,
calmodel= ~ reg + urb_rur + n_copii + n_pens_f + n_pens_m - 1,
calfun= "linear", bounds =   c(0.001,10))
```

Then, inputs for two-stage allocation:

- the variability of target variables in strata;
- the design effect (*deff*);
- the estimator effect (*effst*);
- the intra-cluster correlation coefficient ($\hat{rho}$)

have been prepared in the same way already described for the Labour Force Survey, with a key difference, due to the different nature of the target variables, *i.e.* categorical in LFS and continuous in HBS: the variance in strata for HBS target variables has been calculated with the method of moments, as reported in (2).

The allocation and PSUs selection have been performed in two separate executions, one regarding strata not including Chisinau (urban and rural), and the other one only for the two strata of Chisinau. The reason for this procedure is because the PSU selection (function StratSel in FS4 package) does not allow to define different minimum numbers of SSUs for each selected PSU, while after different trials we realised it was necessary to increase the number of PSUs only in Chisinau strata by fixing a lower minimum number of SSUs per PSU (12, compared to the value 30 used for other strata). In this way, we were able to increase the number of PSUs and decrease the SSU allocation in Chisinau, otherwise exceedingly high.

The following steps were executed for strata not including Chisinau. First, the optimal allocation:

```
stratif_all <- stratif[stratif$STRATUM!=40 & stratif$STRATUM!=41,]
errors_all <- errors
des_file_all <- des_file[des_file$STRATUM!=40 & des_file$STRATUM!=41,]
des_file_all$MINIMUM <- 30
psu_file_all <- psu_file[psu_file$STRATUM!=40 & psu_file$STRATUM!=41,]
rho_all <- rho[rho$STRATUM != 41,]
effst_all <- effst[effst$STRATUM!=40 & effst$STRATUM!=41,]

alloc <- beat.2st(stratif_all,
errors_all,
des_file_all,
psu_file_all,
rho_all,
deft_start = NULL,
effst_all)
```

and then the selection of PSUs:

```
allocat <- alloc$alloc[-nrow(alloc$alloc),]

sample_2st <- StratSel(dataPop= psu_file_all,
idpsu= ~ PSU_ID,
dom= ~ STRATUM,
final_pop= ~ PSU_MOS,
size= ~ PSU_MOS,
PSUsamplestratum= 3,
min_sample= min,
min_sample_index= FALSE,
dataAll= allocat,
domAll= ~ factor(STRATUM),
f_sample= ~ ALLOC,
planned_min_sample= NULL,
launch= F)
```

Same for Chisinau strata, first the optimal allocation:

```
stratif_4 <- stratif[stratif$STRATUM==40 | stratif$STRATUM==41,]

errors_4 <- errors[2:3,]
errors_4$CV1 <- mean(errors_4$CV1)
errors_4$CV2 <- mean(errors_4$CV2)
errors_4$CV3 <- mean(errors_4$CV3)
errors_4$DOM <- c("DOM1","DOM2")

des_file_4 <- des_file[des_file$STRATUM==40 | des_file$STRATUM==41,]
des_file_4$MINIMUM <- min
psu_file_4 <- psu_file[psu_file$STRATUM==40 | psu_file$STRATUM==41,]
rho_4 <- rho[rho$STRATUM==40 | rho$STRATUM==41,]
effst_4 <- effst[effst$STRATUM==40 | effst$STRATUM==41,]

alloc_4 <- beat.2st(stratif_4,
errors_4,
des_file_4,
psu_file_4,
rho_4,
deft_start = NULL,
effst_4)
```

and then the PSUs selection:

```
allocat <- alloc_4$alloc[-nrow(alloc_4$alloc),]

sample_2st_4 <- StratSel(dataPop= psu_file_4,
idpsu= ~ PSU_ID,
dom= ~ STRATUM,
final_pop= ~ PSU_MOS,
size= ~ PSU_MOS,
PSUsamplestratum= 3,
min_sample= min,
min_sample_index= FALSE,
dataAll= allocat,
domAll= ~ factor(STRATUM),
f_sample= ~ ALLOC,
planned_min_sample= NULL,
launch= F)
```

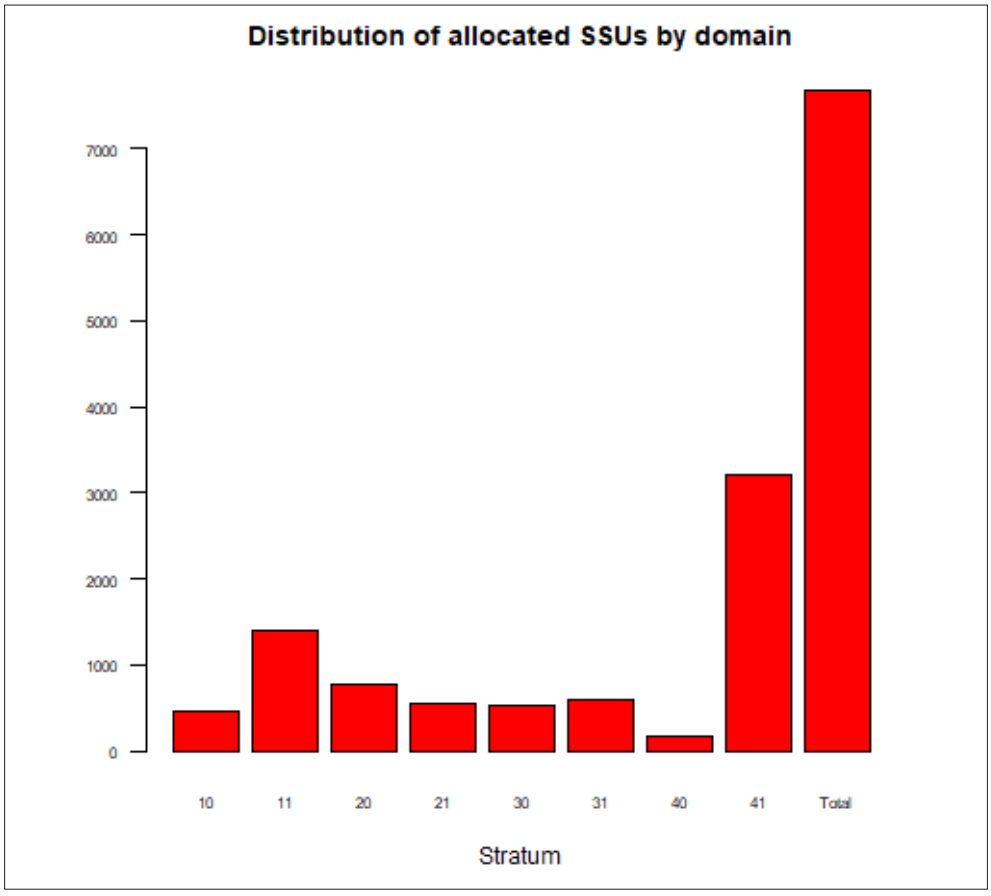
The outputs of the separate executions have been put together, obtaining the overall two- stage sample design reported in Table 3.

Table 3 - HBS allocation

Region	Urban	EA	HH	Oversampling
1	0	12	336	460
1	1	18	579	1396
2	0	15	465	781
2	1	9	297	548
3	0	12	324	518
3	1	12	339	603
4	0	3	48	162
4	1	57	684	3218
Total		138	3072	7886

Source: Our processing

Figure 5 - HBS optimal allocation



Source: Our processing

The oversampling is obtained considering the non-response rate observed in the previous occasion of the survey that are:

Region	Urban	Non-response rate
1	0	0.27
1	1	0.59
2	0	0.40
2	1	0.46
3	0	0.37
3	1	0.44
4	0	0.70
4	1	0.79

The expected errors are reported in Table 4.

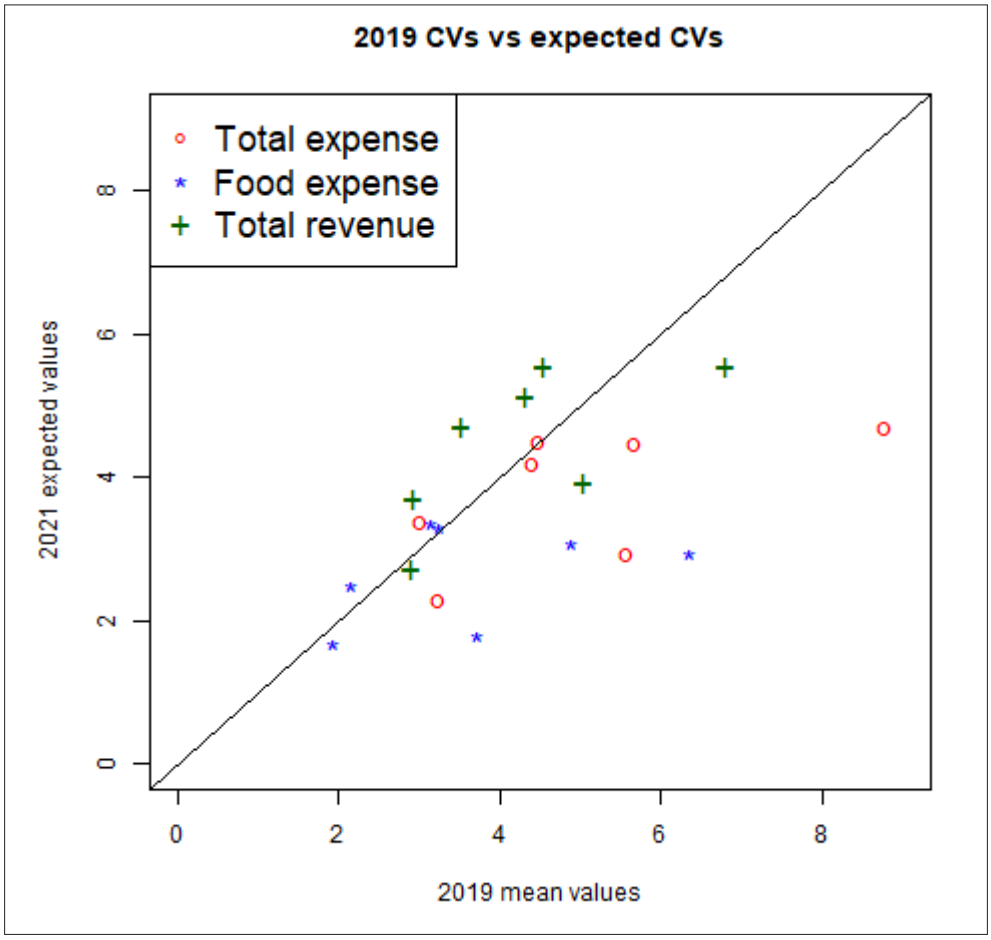
Table 4 - HBS expected CVs

Type	Dom	CV(TE)	CV(FE)	CV(TR)
National	1	0.0228	0.0166	0.0273
Region	1	0.0418	0.0330	0.0556
Region	2	0.0451	0.0334	0.0514
Region	3	0.0446	0.0306	0.0472
Region	4	0.0470	0.0294	0.0554
Urban	1	0.0337	0.0249	0.0394
Urban	0	0.0293	0.0179	0.0370

Source: Our processing

Figure 6 reports how these expected coefficients of variation are with respect to those observed in HBS 2019 rounds.

Figure 6 - HBS 2019 CVs vs expected CVs



Source: Our processing

Analysing this Figure, it can be seen that there is a general gain in terms of precision for variables “Total expense” and “Food expense”, while for “Total revenue” in some cases there are gains and losses in others.

3.3 Domestic Violence Survey

The new allocation of the Domestic Violence surveys has been defined on the basis of the three main target variables chosen by the NBS:

- psychological violence,
- physical violence and
- sexual violence.

The input values for the sample allocation assumes rates equal to the estimates obtained in the last survey in 2010.

Similarly to the process for determining the allocations of 3.1 and 3.2, first the target variables estimates, the sampling variances, design effects ((3)) and estimators effects ((8)) are obtained by means of the software ReGenesees (see Section 2.1) and then the allocation via the software R2BEAT.

In particular, the relevant features of the previous occasion of the Domestic Violence survey that shall be maintained are:

- the sample size approximately equal to 1,500 interviews and PSUs equal to 150,
- for each sample SSU (household) only a female within is selected,
- the calibration is obtained on the number of females for national level, region, urban/rural and age-sex classes.

Given these features, the MINIMUM parameter (minimum number of interviews per PSU) has been set approximately as the one of the last survey, *i.e.* 10 households/interviews per PSU. Note also that in this case the number of final units is equivalent to the number of SSUs (households).

The resulting allocation by domain is reported in Table 5, while Figure 7 illustrates the PSUs allocation in strata.

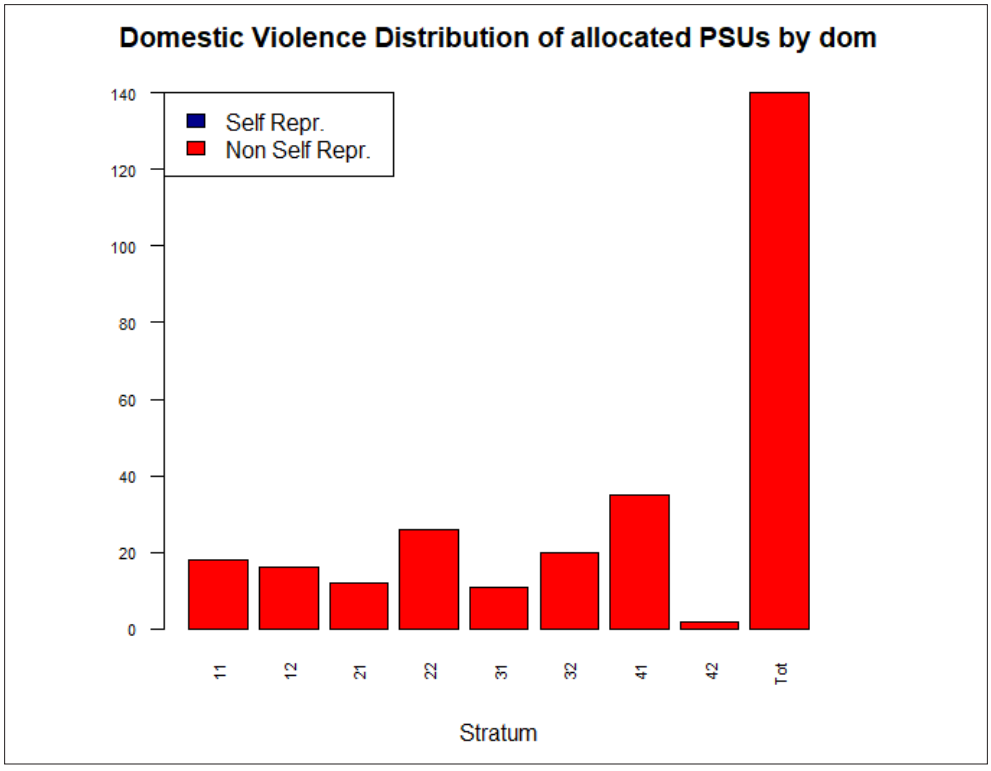
Table 5 - Domestic Violence survey allocation

Region	Urban	EA	Individuals	HH	Oversampling
1	0	18	180	180	237
1	1	20	199	199	311
2	0	28	280	280	368
2	1	14	134	134	209
3	0	22	221	221	291
3	1	12	121	121	189
4	0	2	20	20	26
4	1	38	380	380	594
Total		154	1535	1535	2225

Source: Our processing

To guarantee the required sample size, an oversampling is obtained on the basis of the response rate of the 2012 survey, that is: 0.76 and 0.64 in urban and rural areas, respectively.

Figure 7 - HBS 2019 CVs vs expected CVs



Source: Our processing

Finally, Table 6 reports the expected CV for the three target variables.

Table 6 - Domestic Violence survey expected CVs

Type	Dom	Ps-vio	Ph-vio	S-vio
National	1	0.0319	0.0471	0.0718
Region	1	0.0640	0.0889	0.1503
Region	2	0.0566	0.0792	0.1251
Region	3	0.0763	0.0992	0.1599
Region	4	0.0593	0.1223	0.1422
Urban	1	0.0373	0.0659	0.1078
Urban	0	0.0522	0.0661	0.0946

Source: Our processing

3.4 Energy Consumption Survey

Target variables of the Energy Consumption Survey have been defined as:

1. Coal, consumption (kg);
2. Natural gas, consumption (m^3);
3. Liquefied (petroleum) gases, consumption[9] (l);
4. Lighters and pellets, consumption (kg);
5. Firewood, consumption (m^3);

and the related precision constraints are reported in Table 7.

Table 7 - Energy Consumption, precision constraints CVs

Domain	Precision constraints, CV					
	Coal	Natural gas	Liquefied petrolium gas	Lighters and pellets	Firewood	Electricity
National	0.0872	0.0200	0.0600	0.9999	0.0700	0.0600
Regional	0.4022	0.1069	0.2667	0.9999	0.1731	0.0629
Urban/Rural	0.1635	0.0689	0.1271	0.9999	0.1144	0.0408

Source: Our processing

The data of the last occasion of the survey were available.

The estimates of the target variables have been calculated by using ReGenesees, as follows:

9 Because of its low level, it is used just as control variable.

```
library(Regenesees)
desHT <- e.svydesign(data=ene, ids=~ centr + id_cce,
strata=~strat,
weights=~w_cor,
fpc = NULL, self.rep.str = NULL,
check.data = TRUE)
```

Note that initial weights are the ones already treated to handle non-response.

Calibration estimates have been obtained by using known totals (consumption of natural gas (m^3) [ccf2321] and consumption of electricity (kWh) [ccf23131] by region and coal procurement at national level [ccf23101]):

Because the calibration was already performed by NBS, the function ext. calibrated is used:

```
des <- ext.calibrated(data=ene, ids=~ centr + id_cce,
strata=~strat,
weights=~w_cor,
fpc = NULL, self.rep.str = NULL,
check.data = TRUE,
weights.cal=~wfinal,
calmodel=~region:(ccf2321 + ccf23131) + ccf23101 - 1)
```

Then, inputs for two-stage allocation have been prepared:

- the variability of target variables in strata;
- the design effect (*deff*);
- the estimator effect (*effst*);
- the intra-cluster correlation coefficient ($\hat{p}$).

as already described.

Finally, the optimal allocation was obtained, using the function beat.2st of R2BEAT:

```
alloc <- beat.2st(stratif,
errors,
des_file,
psu_file,
rho,
deft_start = NULL,
effst,
epsilon1 = 5,
mmdiff_deft = 1,maxi = 15,
epsilon = 10^(-11), minnumstrat = 2, maxiter = 200, maxiter1 = 25)
```

and setting MINIMUM=36.

The selection of the PSUs was performed through the function StraSel of FS4:

```
sample_2st <- StratSel(dataPop= psu_file,
idpsu= ~ PSU_ID,
dom= ~ STRATUM,
final_pop= ~ PSU_MOS,
size= ~ PSU_MOS,
PSUsamplestratum= 1,
min_sample= 36,
min_sample_index= FALSE,
dataAll= allocat,
domAll= ~ STRATUM,
f_sample= ~ ALLOC,
planned_min_sample= NULL,
launch= F)
```

and setting MINCS=1.

An overview of the allocation of the two-stage sample design is provided in Table 8.

The oversampling was obtained considering the non-response rates observed in the previous occasion of the survey and shown in Table 9.

Table 8 - Energy Consumption allocation

Region	Urban	EA	HH	Oversampling
1	0	26	928	1006
1	1	11	396	515
2	0	13	462	501
2	1	3	106	143
3	0	12	447	495
3	1	4	141	169
4	0	2	55	64
4	1	27	955	1441
		98	3490	4334

Source: Our processing

Table 9 - Energy Consumption, non-response rate observed in the previous occasion of the survey

Region	Urban	Non-response rate
1	0	0.08
1	1	0.23
2	0	0.08
2	1	0.26
3	0	0.09
3	1	0.16
4	0	0.14
4	1	0.36

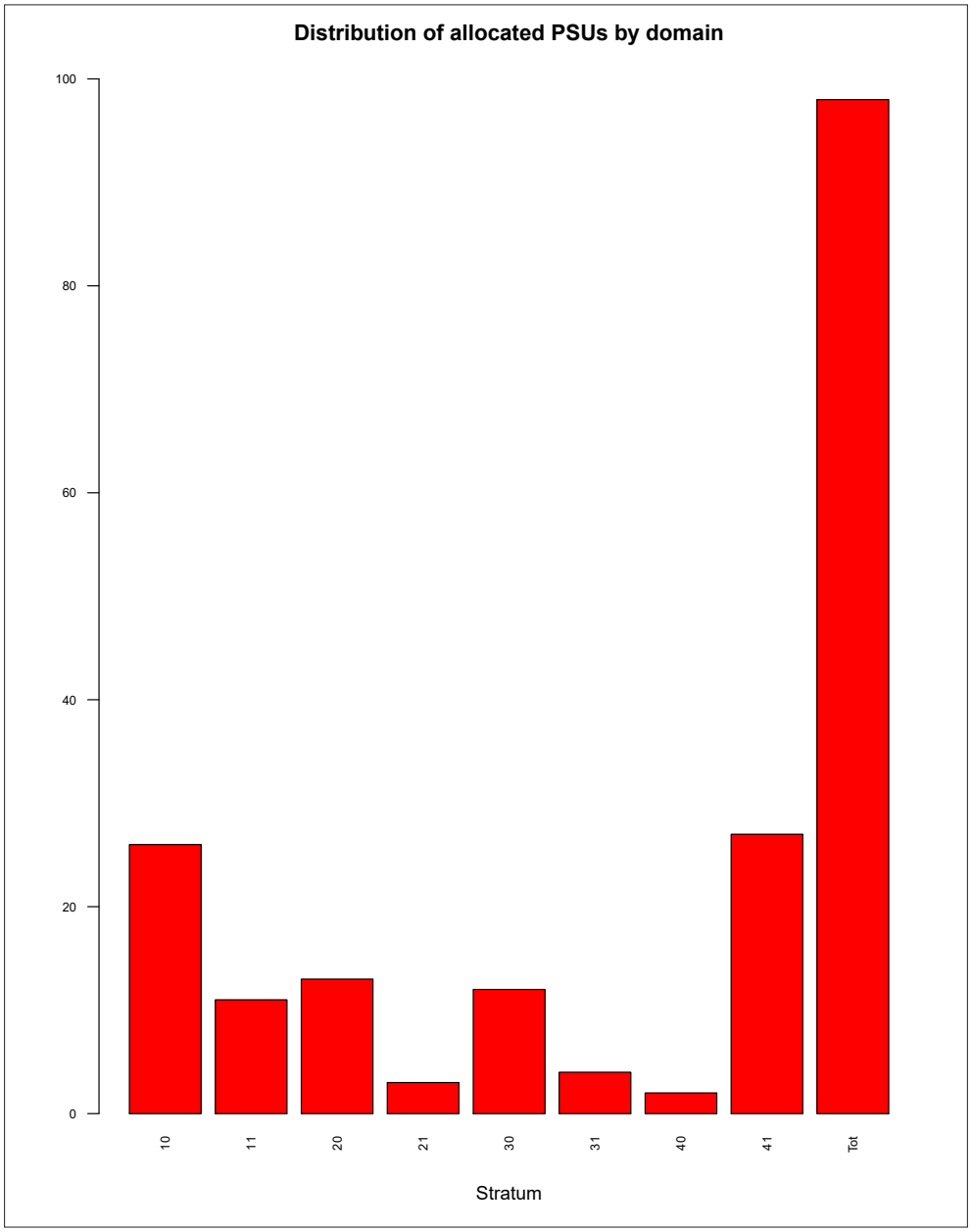
Source: Our processing

Table 10 - Energy Consumption, expected CVs

Type	Dom	Coal	Natural gas	Liquefied petrolium gas	*Lighters and pellets*	Firewood	Electricity
national	1	0.0872	0.0356	0.0354	0.1977	0.0354	0.0263
regional	1	0.0946	0.0797	0.0443	0.3650	0.0452	0.0329
regional	2	0.2553	0.1069	0.0718	1.0533	0.0680	0.0565
regional	3	0.2004	0.0635	0.0663	0.2651	0.0578	0.0629
regional	4	0.4022	0.0487	0.2667	0.5734	0.1731	0.0532
urban	1	0.1004	0.0689	0.0360	0.2307	0.0368	0.0333
urban	0	0.1635	0.0413	0.1271	0.3775	0.1144	0.0408

Source: Our processing

Figure 8 - Optimal allocation of Energy Consumption survey, PSUs by stratum

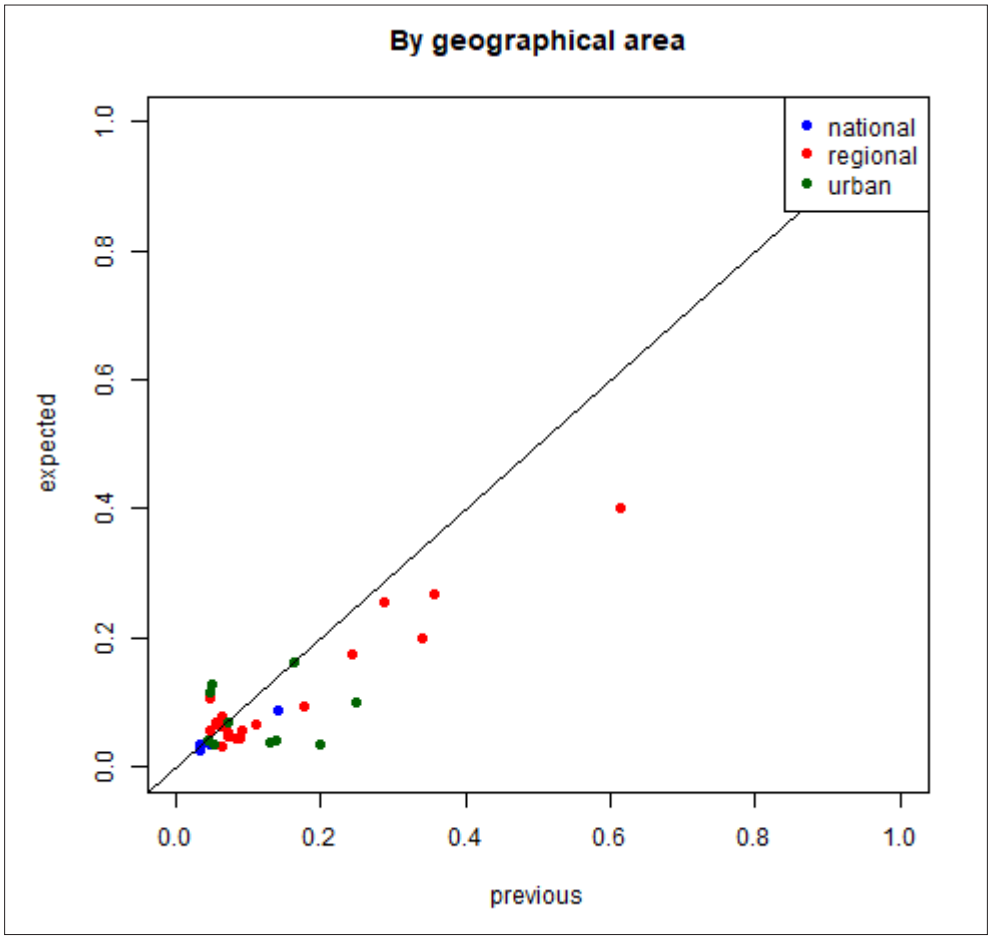


Source: Our processing

The expected errors for the Energy Consumption survey are in Table 10.

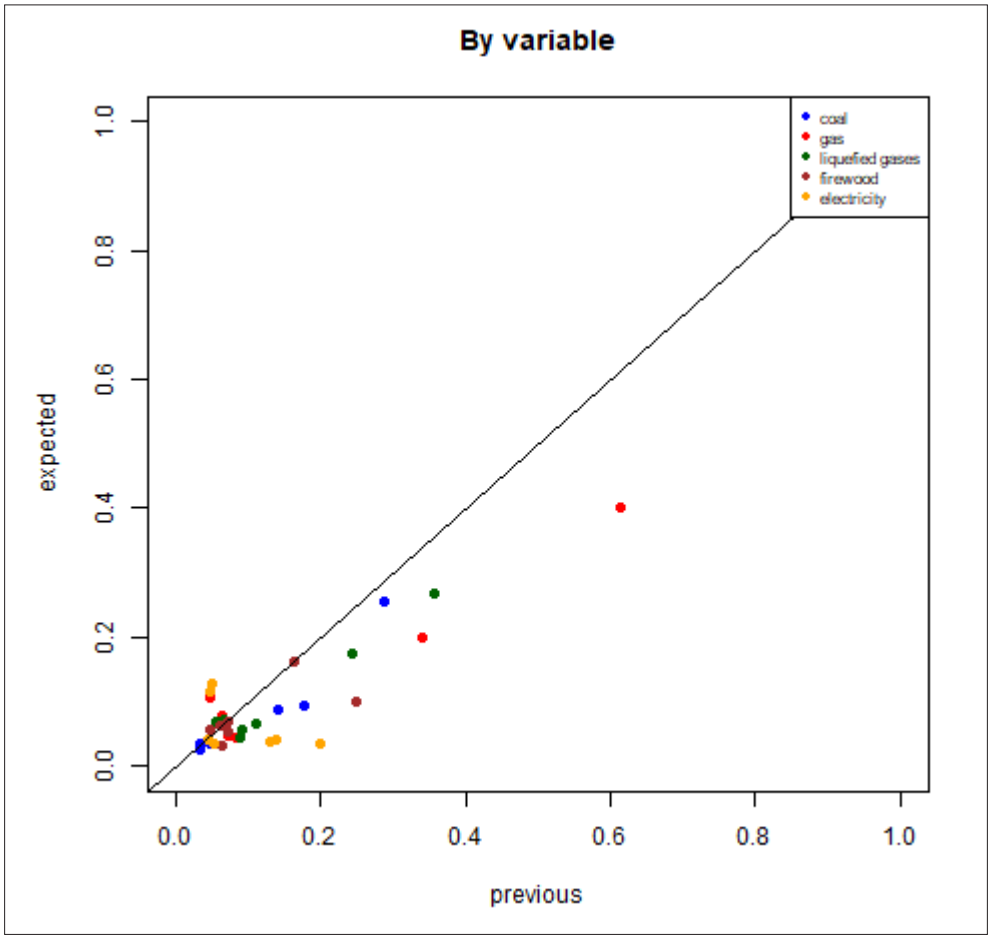
How these expected coefficients of variation are concerning those observed in previous occasions? The comparison with respect to interest variables and geographical area is reported in Figures 9 and 10, respectively.

Figure 9 - Energy Consumption previous CVs vs expected CVs, by geographical areas



Source: Our processing

Figure 10 - Energy Consumption previous CVs vs expected CVs, by interest variables



Source: Our processing

Analysing these figures, it can be seen that a general gain in terms of precision for variables is expected using the proposed allocation.

4. Survey coordination

The structure of the master sample has been built through the coordination of the four illustrated surveys.

In the literature, several sample coordination methods have been developed to obtain a maximal or minimal overlapping between samples drawn at different occasions (see, *e.g.* Matei and Skinner, 2009 and reference therein).

It is possible to distinguish, with respect to a survey taken as reference (in our case LFS), between negative and positive coordination. In the former, the aim is that PSUs with a high probability to be included in the sample of the reference survey have a low probability to be included in the sample of the other survey. The overlapping between the two samples is expected to be minimum and the sample of the two surveys is expected to be spread among different PSUs. On the contrary, for positive coordination, PSUs with a high probability to be included in the sample of the reference survey have a high probability to be included in the sample of the other survey. The overlapping between the two samples is expected to be maximum and the samples of the two surveys are expected to be concentrated in fewer PSUs.

It is important to point out that the two samples are still probabilistic. The choice between negative and positive coordination depends on the need of spreading as widely as possible the sample across PSUs to avoid exhausting households in PSUs and reduce the design effects for the master sample estimates (that is, the estimates obtained joining the same variables collected in the four surveys), and the need to concentrate the interviews due to the constraints established by the organisation of the data collection network. In our case, the main aim was to coordinate different surveys to minimise as much as possible the number of PSUs of the master sample, to match the number of available interviewers.

In the following sections, the scenario of no coordination and the scenario of maximum coordination are illustrated in detail.

4.1 No-coordination

The no coordination scenario was obtained by drawing independently the PSUs for the four surveys. The results we obtained represent the upper bound in terms of PSUs included in the master sample. In the following tables the number of PSUs for the surveys taken two by two, in case of no coordination, are reported:

	LFS		
HBS	0	1	
0	3421	138	3559
1	134	4	138
	3555	142	3697

	LFS		
Violence	0	1	
0	3414	129	3543
1	141	13	154
	3559	142	3697

	HBS		
Violence	0	1	
0	3414	129	3543
1	145	9	154
	3559	138	3697

	LFS		
Energy	0	1	
0	3465	134	3599
1	90	8	98
	3555	142	3697

	HBS		
Energy	0	1	
0	3467	132	3599
1	92	6	98
	3559	138	3697

All the details on this scenario are in Section 5.1.1.

4.2 Maximum coordination

Positive sample coordination has been obtained by considering the LFS PSUs sample a pivot, and coordinating the HBS, Domestic Violence and Energy Consumption samples with it, favouring the PSUs with a higher number of households. Anyway, positive coordination has never been applied for PSUs in the Chisinau urban stratum, due to the very low average number of households.

To obtain maximum coordination, a deterministic approach has been used. The surveys (HBS, Domestic Violence, Energy Consumption) were considered separately. Looking at each stratum, the number of PSUs already selected for the LFS sample and those required for each survey are counted. Three were the possible situations for the sample of PSUs of the coordinated surveys:

1. if the needed number of PSUs for one of the other surveys was equal to the number of PSUs selected for LFS, then the latter were all included;
2. if the needed number of PSU for one of the other surveys was lower than the number of PSUs selected for LFS, then only the LFS PSUs with higher MOS were included;
3. if the needed number of PSU for one of the other surveys was higher than the number of PSUs selected for LFS, then the PSUs selected for LFS were supplemented with the PSUs originally selected for the surveys in case of independent selection.

In the following tables the number of PSUs for the surveys taken two by two, in the case of maximum coordination, are reported:

	LFS		
HBS	0	1	
0	3487	72	3559
1	68	70	138
	3555	142	3697

	LFS		
Violence	0	1	
0	3473	70	3543
1	82	72	154
	3569	142	3697

	HBS		
Violence	0	1	
0	3452	91	3543
1	107	47	154
	3559	138	3697

	LFS		
Energy	0	1	
0	3506	93	3599
1	49	49	98
	3555	142	3697

	HBS		
Energy	0	1	
0	3511	88	3599
1	48	50	98
	3559	138	3697

All the details on this scenario are in Section 5.1.2.

5. Master sample

Two alternative versions of the master sample have been evaluated, depending on the sample coordination of the PSUs for the four different surveys:

- master sample with no coordination;
- master sample with maximum coordination of PSUs.

The master sample with no coordination is characterised by a higher number of PSUs, both in terms of the number of selected ones and in terms of their joint use in each quarter. On the contrary, the master sample obtained by positive sample coordination of the PSUs is characterised by a lower number of initial PSUs, with a higher rate of exhaustion. For both scenarios, a simulation has been carried out to evaluate the exhaustion rate of the PSUs and the need for renewal.

The advantage of the no-coordination solution is that it requires almost no renewal of PSUs outside Chisinau urban. The drawback is in the high number of PSUs simultaneously used in each quarter (270). Conversely, the maximum-coordination scenario requires an additional (limited) renewal of exhausted PSUs out of Chisinau urban (19), but the number of PSUs to be accessed each quarter is lower: 210 instead of 270.

5.1 Master sample simulation of use

The use of the master sample during its life span (from 2021 to 2025, for a total of 20 quarters) has been simulated. The simulation was carried out differently for “Chisinau urban” and the remaining strata.

In other than “Chisinau urban” strata:

1. in quarter t=1 (first quarter of 2021) the number of households to be interviewed are assigned to all LFS and HBS PSUs, the relative PSUs’ household counter is consequently updated;
2. in quarter t+1, an attempt is made to assign the corresponding number of LFS or HBS households to be interviewed; it is successful only if not yet interviewed households are sufficient to meet the required number of households:

- if it is successful, the respective household counter is increased by the new amount;
- otherwise, the PSU is declared as "exhausted" and has to be renewed: the household counter is set to zero and is valorised by adding the current number of LFS and/or HBS households to be interviewed;

3. step 2 is repeated until the end of the period (quarter t=20);
4. without indicating the quarter, step 1 is performed for PSUs belonging to Energy Consumption and/or Domestic Violence.

In "Chisinau urban" the flow of operations is similar, but with two important differences:

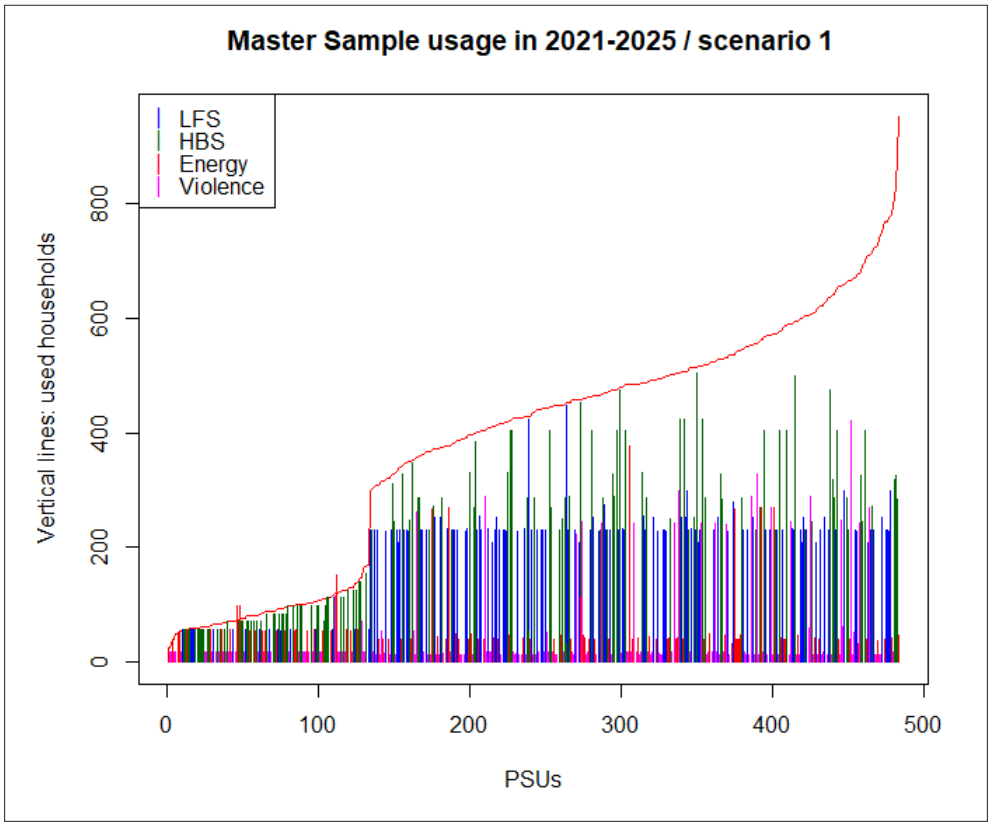
- each PSU can be used by only one survey;
- in the case of PSUs used by LFS, a different exhaustion criterion is adopted: a single LFS PSU is used only for a rotation group, and after 6 consecutive quarters the PSU is removed from the survey.

5.1.1 Simulation of use of the master sample with no coordination of PSUs

In Figure 11 the use of the master sample with no coordination of PSUs is displayed. Alongside the x-axis, PSUs are ordered by size (total number of households), that is indicated by the red line. Each vertical bar indicates the use of households by the four surveys (characterised by different colours). Each time the vertical bar exceeds the red line, then the corresponding PSU has to be substituted. In the no-coordination scenario we have the following situation:

- number of PSUs in master sample: 486 (135 in Chisinau urban)
- number of PSUs contemporarily used in each quarter: 270 (75 in Chisinau urban)
- exhaustion/renewal of available PSUs out of Chisinau urban: 0
- exhaustion/renewal of available PSUs in Chisinau urban: 193

Figure 11 - Simulation of use of the master sample with no coordination of PSUs (scenario 1)



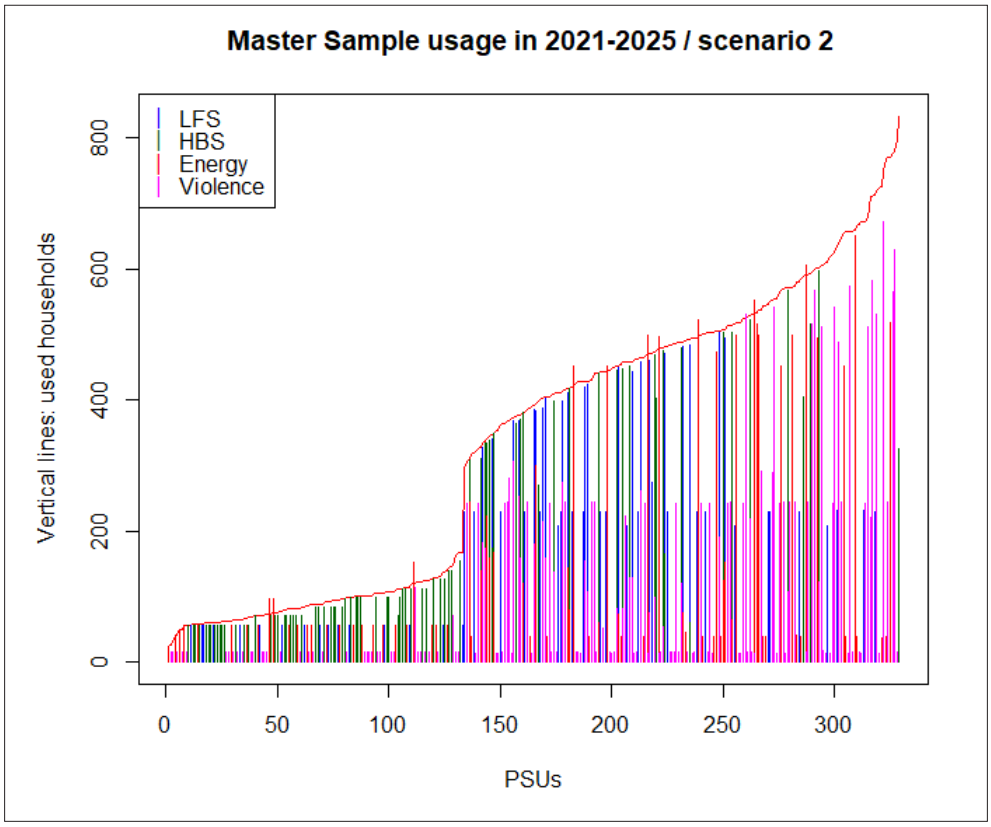
Source: Our processing

5.1.2 Simulation of use of the master sample with coordination of PSUs

In Figure 12 the use of the master sample with maximum coordination of PSUs is displayed. In the maximum coordination scenario we have the following situation:

- Number of PSUs in master sample: 332 (135 in Chisinau urban)
- Number of PSUs contemporarily used in each quarter: 210 (75 in Chisinau urban)
- Exhaustion/renewal of available PSUs out of Chisinau urban: 19
- Exhaustion/renewal of available PSUs in Chisinau urban: 193

Figure 12 - Simulation of use of the master sample with coordination of PSUs (scenario 2)



Source: Our processing

5.2 Selection of additional PSUs to substitute exhausted ones

As illustrated in the simulation, whichever solution is chosen, during the use of the master sample some PSUs will get exhausted, and an additional list of replacement PSUs will be required. It is fundamental to substitute each exhausted PSU with a new one using the following criteria. The new PSU has to be selected.

- In the same stratum utilised at the time of initial selection of the PSUs, *i.e.* in the stratum defined during the execution of the FS4 function StraSel;

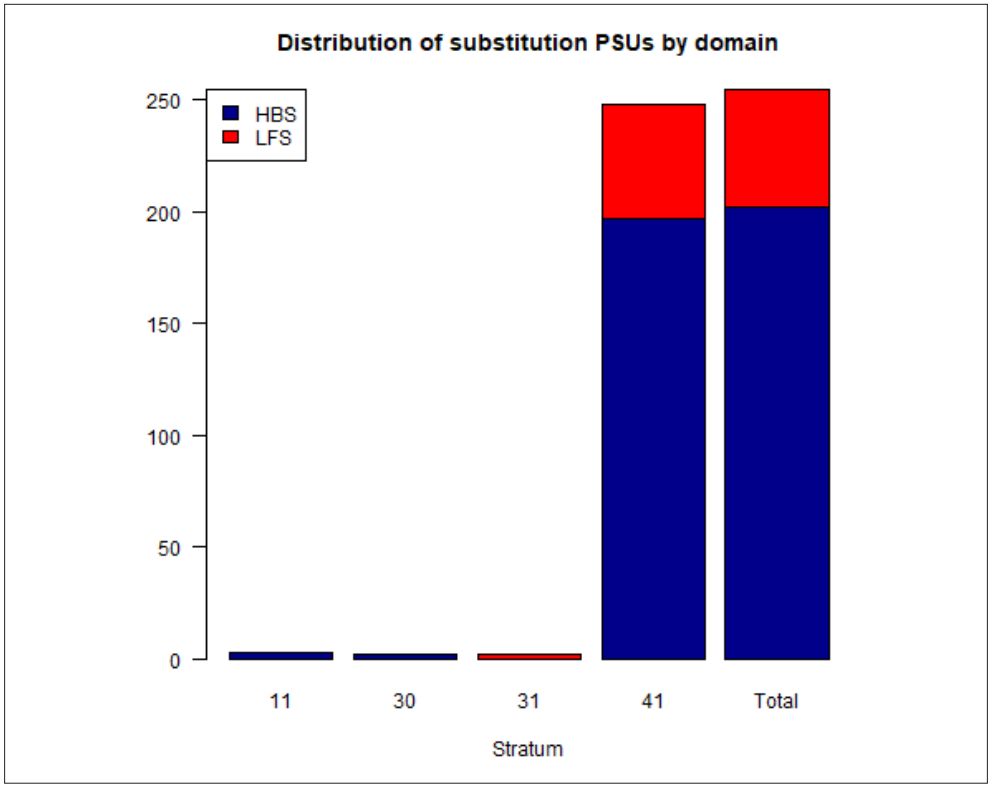
- with approximately the same size of the exhausted one, in order to preserve the first stage inclusion probability.

Having this in mind, two different lists of substitution PSUs have been produced, one for the no-coordination solution and another one for the maximum-coordination solution.

In the case of the no-coordination solution, a list is available with a total of 255 PSU: 202 for HBS and 53 for LFS. In the case of the maximum-coordination solution, the list includes 291 PSU, of which 219 for HBS and 72 for LFS.

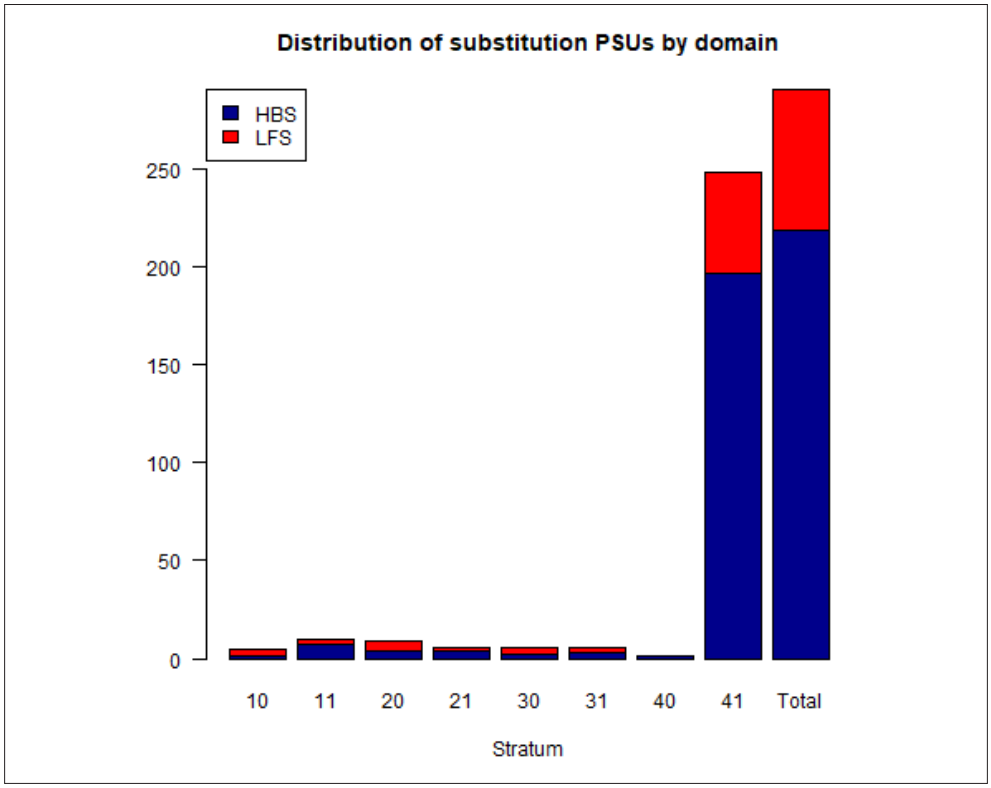
The distribution of additional PSUs in strata under the two scenarios of implementation of the master sample is visualised in Figure 13.

Figure 13 - Distribution of additional PSUs in strata
(a) No-coordination scenario



Source: Our processing

Figure 13 - Distribution of additional PSUs in strata
(b) Maximum-coordination scenario



Source: Our processing

By analysing these distributions, it is clear that there is a high concentration of substitution PSUs in "Chisinau urban", even more evident in the case of the first scenario.

6. Conclusions

Based on the sample designs of the four surveys that will make use of the new master sample during its life span (2021-2025), two possible configurations of the master sample have been defined, depending on the degree of sample coordination of the PSUs required by each survey.

The main advantage of the no-coordination solution is that it requires a very limited renewal of exhausted PSUs outside Chisinau urban. However, its drawback is the high number of PSUs that are simultaneously used in each quarter. Conversely, the maximum-coordination scenario requires a non-negligible renewal of exhausted PSUs out of Chisinau urban, but the number of PSUs to be accessed each quarter is lower.

This last feature (number of PSUs to be accessed each quarter) is crucial, as it has a strong impact on the feasibility and sustainability of the solution. Indeed, the current data collection network organisation is based on the availability of only 150 interviewers, where each interviewer covers only one PSU. If the maximum coordination solution is adopted, there are 135 PSUs out of Chisinau urban, and 75 in Chisinau urban. If the constraint in Chisinau urban is relaxed, and if the number of enumerators in Chisinau urban is increased, then this solution can be considered feasible.

The overall approach followed for the design of the master sample is such that it can be easily generalised in other similar situations. In particular, based on this experience, one of the most burdensome task of the whole procedure, namely the preparation of the input datasets required by the R2BEAT optimisation functions for the two-stage sample design, has been completely automatised in a new release of this package. This new version (1.0.2) also includes the functions of the FS4 package, thus simplifying the overall workflow. A complete example, from the input preparation to the two-stage allocation, ending with the PSU selection, is contained in a dedicated vignette, "*Two-stage sampling design workflow*"[10].

10 https://urlsand.esvalabs.com/?u=https%3A%2F%2Fbarcaroli.github.io%2FR2BEAT%2Farticles%2FR2BEAT_workflow.html&e=3cfb7ead&h=098d22c7&f=y&p=y.

References

Barcaroli, G., M. Ballin, H. Odendaal, D. Pagliuca, E. Willighagen, and D. Zardetto. 2020. "SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys". In *The Comprehensive R Archive Network – CRAN*. Wien, Austria: Institute for Statistics and Mathematics of Wirtschafts Universität. https://cran.r-project.org/package=SamplingStrata.

Bethel, J. 1989. "Sample allocation in multivariate surveys". *Survey methodology*, Volume 15, N. 1: 47-57. Ottawa, Ontario, Canada: Statistics Canada.

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration estimation in survey sampling". *Journal of the American Statistical Association*, Volume 87, N. 418: 376-382.

Fasulo, A., G. Barcaroli, R. Cianchetta, S. Falorsi, A. Guandalini, D. Pagliuca, and M.D. Terribili. 2020. "R2BEAT: Multistage sample design". In *The Comprehensive R Archive Network – CRAN*. Wien, Austria: Institute for Statistics and Mathematics of Wirtschafts Universität. https://cran.r-project.org/package=R2BEAT.

Kish, L. 1995. "Methods for design effects". *Journal of Official Statistics - JOS*, Volume 11, N. 1: 55-77. Stockholm, Sweden: Statistics Sweden.

Matei, A., and C.J. Skinner. 2009. "Optimal sample coordination using controlled selection". *Journal of Statistical Planning and Inference*, Volume 139, N. 9: 3112-3121.

Neyman, J. 1934. "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". *Journal of the Royal Statistical Society*, Volume 97, N. 4: 558-625.

Tschuprov, A.A. 1923. "On the mathematical expectation of the moments of frequency distributions in the case of correlated observations". *Metron*, 2: 461-493; 646-683.

Särndal, C.-E. 2007. "The calibration approach in survey theory and practice". *Survey methodology*, Volume 33, N. 2: 99-119. Ottawa, Ontario, Canada: Statistics Canada.

Woodruff, R.S. 1971. “A Simple Method for Approximating the Variance of a Complicated Estimate”. *Journal of the American Statistical Association*, Volume 66, N. 334: 411-414.

Zardetto, D. 2015. “ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys”. *Journal of Official Statistics* - *JOS*, Volume 31, N. 2: 177-203. Stockholm, Sweden: Statistics Sweden.