The usage of R programme for official statistics

Servete Muriqi, Kosovo Agency of Statistics, servete.muriqi@rks-gov.net

Drita Sylejmani, Kosovo Agency of Statistics, drita.gj.sylejmani@rks-gov.net

Hydai Morina, Kosovo Agency of Statistics, hydai.morina@rks-gov.net

Abstract

The Kosovo Agency of Statistics (KAS) has started to use R programme for different phases of survey which has helped overall the sample process, designing and extracting the sample, calculation of sampling errors, calculation of weights and the calibration process which allow us to improve the accuracy of the estimates.

The R contains different packages with the set of tools for selecting the sample and calibration of weights which is a technique that uses auxiliary information to weight the statistical units. Calibration allows us to reduce sampling variance or non-sampling errors such as nonresponse bias by taking advantage of the auxiliary information given by the auxiliary variables who's the total is known in the population. KAS is using R programme for estimation and calibration of weights in all household surveys.

The paper describes the sample design of EU SILC (Statistics on Income and Living Conditions), calculation of weights and the calibration process using R packages. For the first time the calibration process was applied at SILC survey and overall process of weights was made easier and has improved the accuracy of estimates of the data. Then the R has started to use also for other household surveys for different stages of survey.

Keywords: R, sample design, weight, calibration, SILC

1. Introduction

Kosovo Agency of Statistics (KAS), respectively the Department of Social Statistics (DSS) with the support of the World Bank, for the first time in 2018 conducted the survey on Statistics on Income and Living Conditions (SILC) known as EU-SILC. The purpose of publishing the results of the Statistics on Income and Living Survey (SILC), is to provide statistical data on the living conditions of households in Kosovo and other similar issues related to living standards. The data from SILC aims to alleviate the lack of information in the field of living standards of households in Kosovo. These data will serve as a useful reference base for all users of statistical data.

The Statistics on Income and Living was funded by a World Bank (WB) grant, and professional support was provided by experts under IPA 2015, as well as the EU-SIDA project, through the Swedish Statistical Office and experts from Bulgaria and Malta. The Department of

Methodology and Information Technology has made a valuable contribution to the development of the questionnaire in electronic form in the Survey Solutions application in the CAPI / tablet method (computer-assisted personal interviewing for data collection), sample preparation, maps and monitoring of work in the field that has been done by the regional offices for the collection data on income and living conditions.

The sampling unit has given a great contribution to the sampling design part and extracting the sample. This was done using the R packages. Also R programme was used to calculate the weights and calibration part.

2. The methodology of Statistics on Income and Living Conditions (EU-SILC)

The European Union (EU) Statistics on Income and Living Conditions (EU-SILC) is an instrument that aims to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income distribution, poverty and social exclusion. It also covers various related EU living conditions and poverty policies, such as child poverty, access to healthcare and other services, housing, over-indebtedness and quality of life. It is also the main source of data for microsimulation purposes and flash estimates of income distribution and poverty rates. This instrument is anchored in the European Statistical System (ESS).

The survey provides two types of data:

- Cross-sectional data refer to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions;

- Longitudinal data refer to individual/household changes over time, observed periodically over a four-year period (or more years if a longer duration panel is used).

Information on social exclusion and housing conditions is collected mainly at household level, while labour, education and health information is obtained for persons aged 16 and over. The core of the instrument, income at detailed component level, is collected both at personal and household level.

EU-SILC has been used to provide data on the structural indicators of social cohesion (at-risk-of-poverty rate, S80/S20) and in the context of the two Open Methods of Coordination in the field of social inclusion and pensions. Since 2010, at the outset of the Europe 2020 strategy, EU-SILC data are being used for monitoring poverty and social inclusion in the EU.

2.1 Statistics on Income and Living Conditions in Kosovo

The SILC survey is based on concepts, definitions and methodological recommendations of the Eurostat “DOCSILC065 operation 2021” for SILC.

The data pertain to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions which belongs to cross sectional data.

Longitudinal data pertain to individual-level changes over time, observed periodically over a four-year period. Social exclusion and housing condition information is collected mainly at household level while labour, education and health information is obtained for persons aged 16 and over. The core of the instrument, income at very detailed component level, is mainly collected at personal level.

- The main objectives of the SILC are in the first place to provide basic data required for policy making at national level and for different sectors.
- SILC is a tool for providing data on income distribution, level and structure of poverty and social exclusion in a timely and comparable manner.
- SILC provides four basic files containing target variables based on common concepts and definitions.
- Annual data contains the following components:
- Household register, Personal register, Household data, Personal data of people aged 16 and more.
- Every year additional data on household and household members on specific topics is collected via ad-hoc modules. Indicators on poverty and social exclusion are calculated on the basis of SILC data using common methodology approved by Eurostat for the data collection, for the obtaining of target variables and for the calculation of common indicators. The poverty rate is defined as 60% of the median equivalised disposable income.

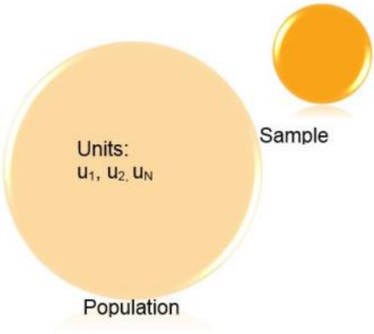
Target population consists of all persons living in private households. Persons living in collective households and in institutions are generally excluded from the target population.

2.2 Sample

In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations.

Typically, the population is very large, making a census or a complete enumeration of all the values in the population either impractical or impossible. The sample usually represents a subset of manageable size.

Figure 1: Sample selection



Statistics are calculated on the basis of the samples collected so that inferences can be drawn or extrapolations made from the sample to the population. The data sample may be drawn from a population without replacement, in which case it is a subset of a population; or with replacement, in which case it is a multi-subset

2.3 Sample Design of SILC Kosovo

The data set is based on sample survey. The sampling frame was based on the data and cartography from last Kosovo Census and frame was updated time to time. For the purposes of the census enumeration, Kosovo was subdivided into enumeration areas (EAs), which are relatively small operational segments defined for the census enumeration. A total of 4,626 EAs were defined for Kosovo, and these were used as the primary sampling units (PSUs) selected at the first sampling stage for the SILC. The sample design is two stage. On the first stage EA were selected and in second stage the household was selected within EA. Rotational (integrated) design refers to sample selection based on four subsamples or replications that are all similar in size and design and representative of the whole population. The subsample incudes 95 EA while within each sample EA, 12 sample households were selected at the second stage, for a total sample size of 1140 households in each subsample. From one year to the next, some replications are retained, while others are dropped and replaced by new replications.

The survey is stratified into 14 strata (by 7 regions and urban/rural).

Figure 2. shows what the chosen sample should look like in the case of four rotational groups. As mentioned above, every effort should be made to ensure that each sub-sample ‘captures’ the characteristics of the population observed.

Figure 2: Rotational design

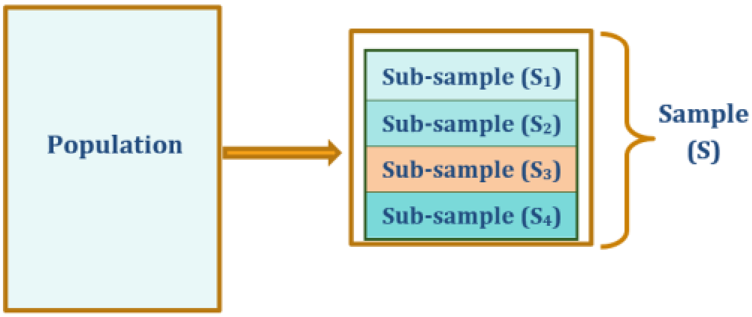
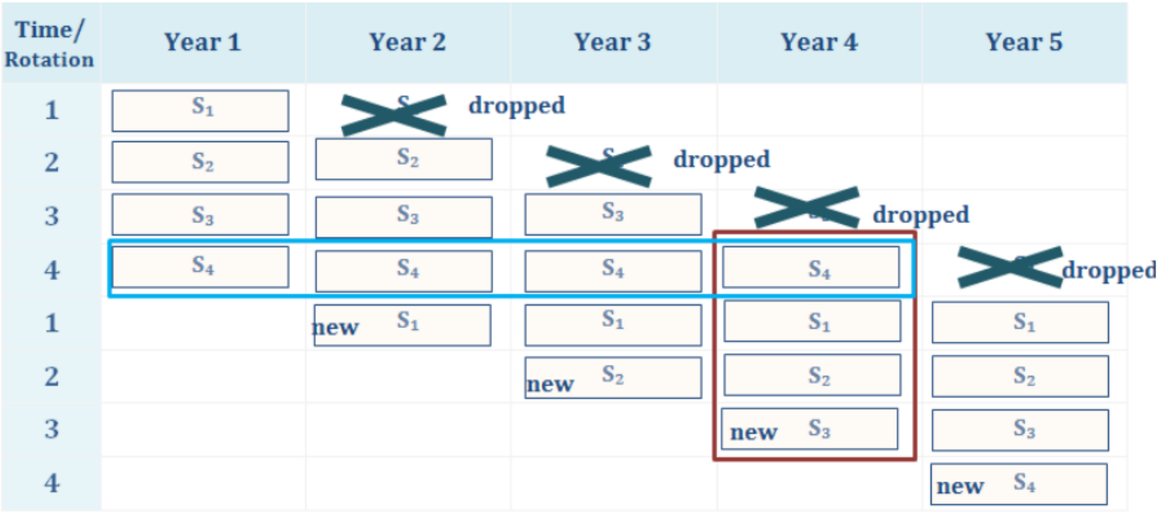


Figure 3. shows what happens with four rotational groups over five years. The blue rectangle presents one example of the longitudinal data, while the red rectangle presents one example of the cross-sectional data. The sample for any one year consists of four replications, which have been in the survey for between one and four years. Any particular replication remains in the survey for four years; each year, one of the four replications from the previous year is dropped and a new one is added. Between year T and T+1 the sample overlap is 75% (in the absence of attrition); the overlap between year T and year T+2 is 50%; and it falls to 25% from year T to year T+3, and to zero for longer intervals.

Figure 3: Illustration of a sample rotational design



Time/ Rotation	Year 1	Year 2	Year 3	Year 4	Year 5
1	S_1	dropped			
2	S_2	S_2	dropped		
3	S_3	S_3	S_3	dropped	
4	S_4	S_4	S_4	S_4	dropped
1		new S_1	S_1	S_1	S_1
2			new S_2	S_2	S_2
3				new S_3	S_3
4					new S_4

At the beginning, a cross-sectional representative sample of households is selected. It is divided into four sub-samples, each on its own representative of the entire population and similar in structure to the entire sample. One sub-sample is purely cross-sectional and is not followed up after the first round. Respondents in the second sub-sample are asked to participate in the panel for two years, in the third sub-sample for three years, and in the fourth for four years. From year 2 onwards, one new panel is introduced each year; the respondents are asked to participate for four years. In any one year, the sample consists of four sub-samples, which together constitute the cross-sectional sample. In the first year they are all new samples, while in all subsequent years only one is a new sample. In year 2, there are three panels; in year 3, one is a panel from the second year and two from the third year; in subsequent years, one is a panel from the second year, one from the third year, and one from the fourth (final) year

2.4 Data collection

Data collection in *Statistics on Income and Living Conditions* is done by CAPI method. The CAPI method insures control and monitoring the data in time, has reduced the errors during the data collection, the validation rules are applied and with this application the enumerators are monitored every time. CAPI facilitates logic checks, skip patterns, and validations during the interview. This makes the survey more efficient and helps assure higher quality data. It also saves later efforts on data cleaning and data entry.

It can automatically record each interview's start time, end time and GPS location, making it easy for supervisors to check whether an enumerator indeed conducted a given interview or not by comparing its time and GPS data to that of other interviews during which a supervisor was present.

3. Weights in SILC

The sample design is used for calculation the weights for SILC data.

The weights based on the construction and design are:

1. Design weight (for households and for a selected respondents)
2. Cross-sectional weight
3. Base weight
4. Longitudinal weight

3.1 Design weight for household

The design weights are defined for all selected units. The household design weights are aim to draw inference on the household population from the household samples.

The design weights of the households is the inverse of the inclusion probabilities.

Design weight variable DB080 is computed as follows:

DB080h = 1 / (probability of selection of h)

The term *probability of selection of h*, the sampling probability for the i-th sample PSU in the h-th stratum, is the product of the probabilities of selection at every stage in each sampling stratum

The design weights have to be inflated by the inverse of the response propensities in order to compensate for the loss of units in the sample. A classical procedure consists of modifying the design weights by a factor inversely proportional to the response rate within each 'homogeneous group', in which the response probabilities are assumed to be equal:

$$\mathrm{DB080}_{\mathrm{h}}^{(\mathrm{N})} = \mathrm{DB080}_{\mathrm{h}} \cdot \frac{1}{\mathrm{R}_{\mathrm{K}}}$$

Where R_K denotes s the (weighted) response rate in the group k the household h belongs to:

$$\mathrm{R}_{\mathrm{K}} = \frac{\text{sum of design weights of responding units in cell k}}{\text{sum of design weights of selected units in cell k}}$$

3.2 Adjustment to external sources (calibration): SILC target variables DB090

Cross-sectional household weight (DB090)

After adjustments for non-response and to external sources (calibration) of the household design weight, the cross-sectional household weight (DB090) is calculated. DB090 is used to weigh household data and indicators produced at household level.

Age and sex are the natural ancillary variables used in a human population survey. The distribution of the human population by age and sex are taken from other statistical sources. Modifying the survey weights in the right way ensures that the sample exactly reproduces the population structure. As regards variables in the survey that are correlated with the ancillary information, the precision of estimates is usually improved by applying the new calibrated weights.

More precisely, suppose that there are ‘J’ auxiliary variables x1…xj…xJ, called calibration variables, with known population totals (for the numerical variables) or marginal counts (for the categorical variables). Without loss of generality, we can assume that all the calibration variables are numerical (otherwise, we consider the 0/1 variables for each category).

New household weights (DB090) are ‘as close as possible’ (as determined by a certain distance function) to the initial weights DB080(N)

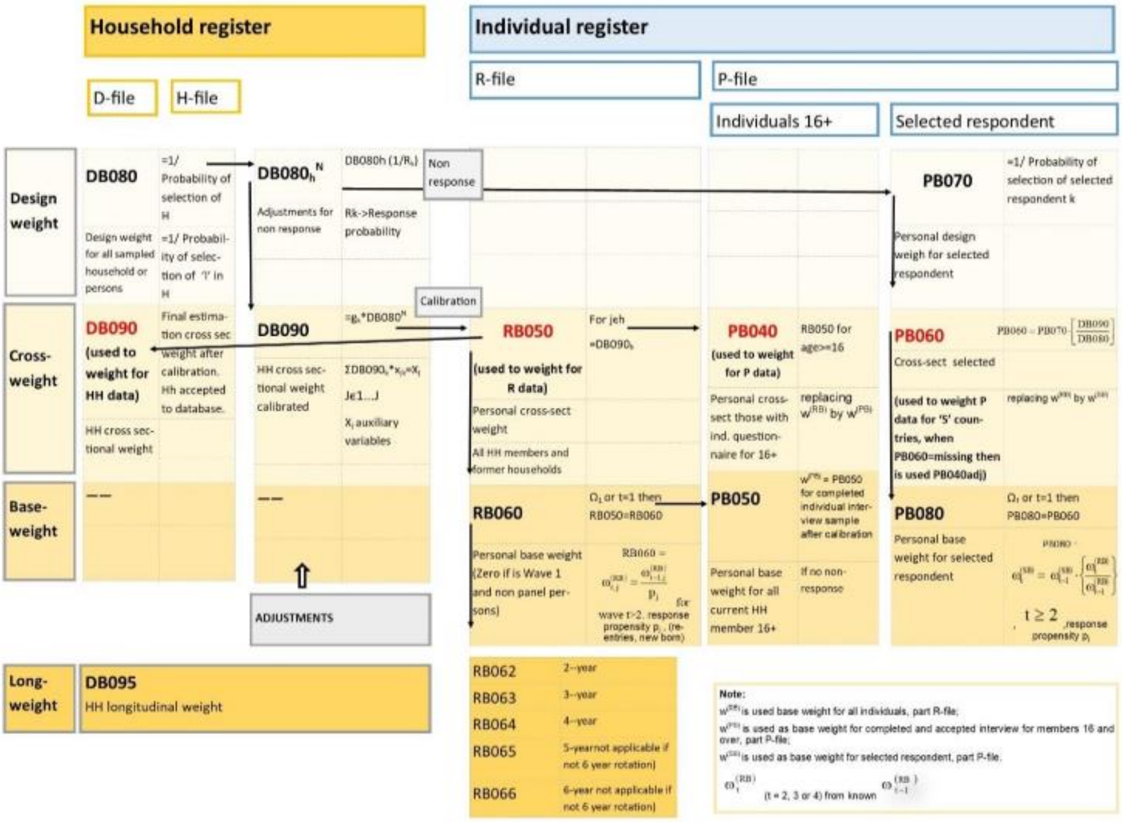
These new weights are calibrated on the totals Xj of the ‘J’ auxiliary variables; in other words they verify the calibration equations:

$$\forall\, j = 1 \ldots J \qquad \sum_{k \in s} \mathrm{DB090}_k \cdot x_{jk} = X_j \, .$$

$$\text{Where } \mathrm{DB090}_k = g_k \times \mathrm{DB080}^{(N)}$$

In R language package survey is used for calibration process.

Figure 4: Scheme of the weights


Household register
D-file
H-file
Individual register
R-file
P-file
Individuals 16+
Selected respondent
Design weight
Cross-weight
Base-weight
Long-weight
DB080
=1/ Probability of selection of H
Design weight for all sampled household or persons
=1/ Probability of selection of 'I' in H
DB080$_h^N$
DB080h (1/R$_k$)
Non response
Adjustments for non response
Rk->Response probability
Calibration
DB090
(used to weight for HH data)
Final estimation cross sec weight after calibration. Hh accepted to database.
HH cross sectional weight
DB090
=g$_k$*DB080^N
HH cross sectional weight calibrated
ΣDB090$_k$*x$_{jk}$=X$_j$
Je1…J
X$_j$ auxiliary variables
——
——
ADJUSTMENTS
DB095
HH longitudinal weight
RB050
(used to weight for R data)
For jeh
=DB090$_h$
Personal cross-sect weight
All HH members and former households
PB040
(used to weight for P data)
RB050 for age>=16
Personal cross-sect those with ind. questionnaire for 16+
replacing w$^{(RB)}$ by w$^{(PB)}$
PB070
=1/ Probability of selection of selected respondent k
Personal design weigh for selected respondent
PB060
$\mathrm{PB060} = \mathrm{PB070} \cdot \left[\frac{\mathrm{DB090}}{\mathrm{DB080}}\right]$
Cross-sect selected
(used to weight P data for 'S' countries, when PB060=missing then is used PB040adj)
replacing w$^{(RB)}$ by w$^{(SB)}$
RB060
Ω$_1$ or t=1 then RB050=RB060
Personal base weight (Zero if is Wave 1 and non panel persons)
RB060 =
$\omega_{t,i}^{(RB)} = \frac{\omega_{t-1,i}^{(RB)}}{p_i}$ for wave t>2, response propensity p$_i$, (re-entries, new born)
PB050
w$^{(PB)}$ = PB050 for completed individual interview sample after calibration
Personal base weight for all current HH member 16+
If no non-response
PB080
Ω$_1$ or t=1 then PB080=PB060
Personal base weight for selected respondent
PB080 =
$\omega_t^{(SB)} = \omega_{t-1}^{(SB)} \cdot \left\{\frac{\omega_t^{(RB)}}{\omega_{t-1}^{(RB)}}\right\}$
, $t \geq 2$,response propensity p$_i$
RB062 2--year
RB063 3--year
RB064 4--year
RB065 5-yearnot applicable if not 6 year rotation)
RB066 6-year not applicable if not 6 year rotation)

Note:

w$^{(RB)}$ is used base weight for all individuals, part R-file;

w$^{(PB)}$ is used as base weight for completed and accepted interview for members 16 and over, part P-file;

w$^{(SB)}$ is used as base weight for selected respondent, part P-file;

$\omega_t^{(RB)}$ (t = 2, 3 or 4) from known $\omega_{t-1}^{(RB)}$

4. Conclusions

KAS in the last years have been working on improvement of the quality in the overall process of statistical production.

KAS is in continuing process of adoption of methodology and best practices align of EU requirements.

The standardization of data collection across the different sectors within KAS has ensured the consistency and reliability of data collection practices leading to more accuracy and comparable statistics. KAS has an agreement with the World Bank to carry out surveys using tablets and the Survey Solutions software.

Web-based data collection from enterprises preserves confidentiality, reduce the form filling burden, and increase the efficiency of the statistical processes. The use of the internet could be further enhanced if all business data collection were processed through a common web site. All these measures have increased the cost-effectiveness of data collection and the production of higher quality statistics.

Increasing the knowledge and capacity of staff to use the new technologies and the usage of different software and packages for calculation and compiling the statistics is improving the efficiency and timeliness of data production and dissemination processes.

5. Referencies

https://circabc.europa.eu/sd/a/f8853fb3-58b3-43ce-b4c6-a81fe68f2e50/Methodological%20guidelines%202021%20operation%20v4%2009.12.2020.pdf

https://askapi.rks-gov.net/Custom/76fddec3-ec60-4377-9f73-874f9f4e0392.pdf

https://askapi.rks-gov.net/Custom/9e74a2dc-870a-490a-9e88-064fc6d73297.pdf