Improving LFS by Using Different Data Sources

Liliana Roze

1. Introduction

Various data sources are utilized in producing the Labour Force Survey (LFS) data in Latvia. Introduction of this is a lengthy journey where a specific introduction date cannot be set since new external and internal sources and databases were added gradually. Moreover, regular data updates should take place and consider the influence of socio-economic events, therefore use of external data sources require collaboration with data owners and holders.

Over the years, apart from the used data sources themselves, the utilisation thereof has changed as well. Nowadays, source data are used at several data processing stages (incl., when building target sampling frames), which allows improving quality of the data collection, and optimizing weight calculation.

2. Latvian LFS

Latvian LFS has two-stage sample design. The sample units used to build the first-stage sample frame are small territories – survey polygons formed by valid occupied dwellings. In the second stage, dwellings are selected. The LFS is negatively coordinated in the first-stage sample.

To minimize survey costs and non-response, the persons to be blocked are defined, and then the sampling units used in the previous samples are blocked for approximately four years. Blocked persons are identified by using address identification information.

3. Sampling frame

To make a list of valid dwellings and persons within them, a sampling frame is built. It is done by merging data from several sources, and the frame is updated monthly.

Register of Natural Persons and Address Register are the primary sources used to build the sampling frame (Figure 1 illustrates the key information in each register). The data are merged by using personal identity numbers and identifiers in the Address Register.

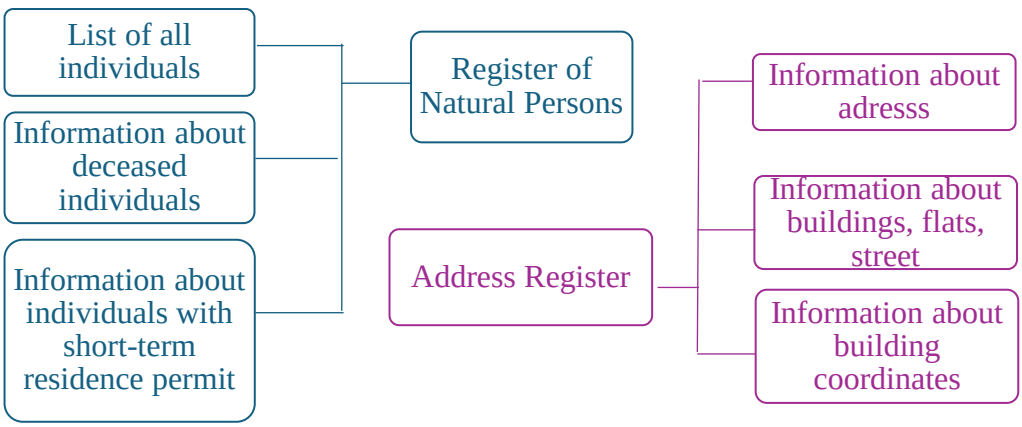

Figure 1. Information from Register of Natural Persons and Address Register

It should be noted that the Register of Natural Persons contains information on all inhabitants, including prisoners and residents of nursing homes, and that not all addresses in the Address Register are suitable for statistical needs. Therefore, additional information (e.g., on residence permits, nursing homes, electricity consumption, buildings unsuitable for habitation, etc.) is used. The information is mainly used to identify people who are not usual residents of Latvia and to differentiate dwellings that are not resided by private households. Consequently, the respective persons and dwellings are excluded from the sampling frame.

Based on a common requirement, the survey results should be broken down by administrative region. This aggregation in Latvia is based on the Classification of Administrative Territories and Territorial Units. Such an approach enables acquisition of precise and detailed information thereby improving the quality of selection.

It is worth mentioning that, in the sample design weighting computation, blocked persons are also considered, however, during the sampling, only the persons and dwellings not blocked are selected.

As mentioned prior, the frame is updated with the latest information on a monthly basis, while the LFS is conducted quarterly, and the sample frame thereof is drawn semi-annually. Namely, the samples of the first two quarters are drawn from one common sampling frame while samples of the last two quarters are drawn from another common sampling frame, i.e., the latest sampling frame available at the time of the sampling is used.

Aiming to improve quality of the data collection and ensure that information about the sampling units is the most up-to-date before the data collection, an updated information about addresses and telephone numbers (obtained from different administrative sources as well as statistical surveys) is also integrated.

4. Sample weighting

Final weights are computed with the help of calibration. Initially, weighting frame is acquired as the primary calibration source. Similarly to the sampling frame, primary information is drawn from the Register of Natural Persons and the Address Register, removing the inapplicable information (collective dwelling, residence permits, deceased individuals). The Classification of Administrative Territories and Territorial Units for aggregation weighting frame is also applied.

The weighting frame unit is a person, therefore estimation of unemployment and employment rates may be facilitated by the State Revenue Service and the State Employment Agency data. The State Revenue Service has information on the registered employed persons and the State Employment Agency has unemployed-related data. The auxiliary variables and calibration sums related to unemployment and employment are computed based on a gender and an age group (there are 5 and 7 age groups).

Afterwards, based on the number of persons registered in a dwelling, the dwelling size is calculated and applied to the weights (i.e., each person gets a value equal to or less than 1, e.g., if one person is registered in the dwelling – it gets 1, if two persons – each gets 0.5). The calibration sums are calculated by applying the territorial division.

Then, external data about distribution of the population in private households are incorporated and categorized into levels of detail in accordance with the LFS regulation. It is important to mention that external data do not match the sampling frame data. External data are needed to adjust the calibration sums. Consequently, for calibration sums and auxiliary information three different levels of detail are established. The smallest level of detail is divided

into four territorial divisions (Riga, State cities, other towns, and rural territories) as well as gender and age groups (each divided into 19 segments). The second level of detail involves wider territorial division, categorized by regions of residence and five age groups. The third level of detail focuses on the capital and the State cities.

Next, the calibration sums are calculated. Even though initial data processing takes place on a weekly basis, the calibration sums are determined for the period of the corresponding weight calculation. As a result, weighting frame with 232 auxiliary information variables from several sources are obtained. The variables are used in calibration and to generate calibration sums. Ultimately, three outputs are obtained: weighting frame, auxiliary information, and calibration sums. The main data sources and necessary outputs are shown in the Figure 2.

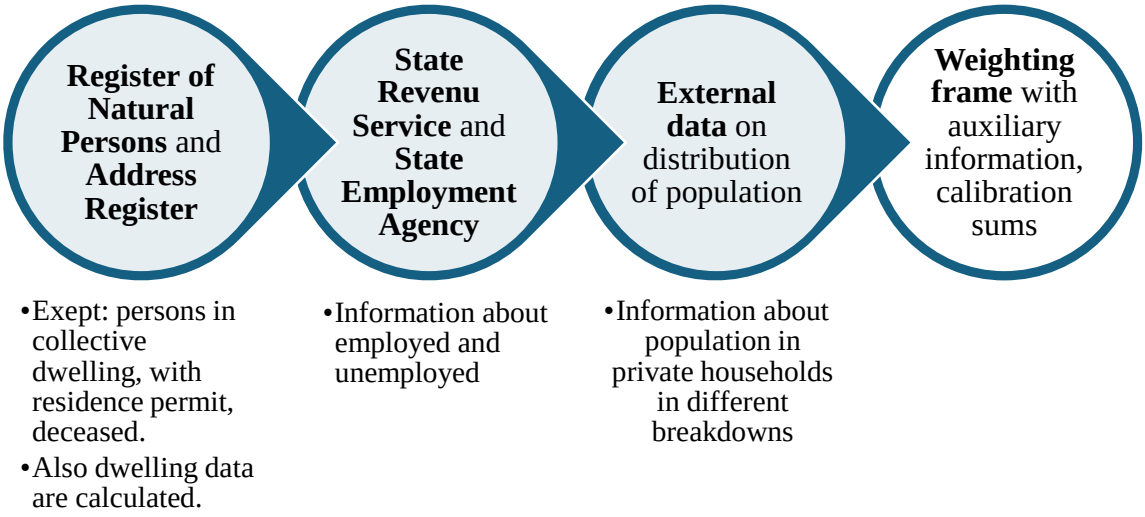

Figure 2. Data sources used to compute weighting frame, auxiliary information, and calibration sums

Thus, by using personal identification, the information from the weighting frame is merged with the information on valid LFS respondents. It is worth mentioning that also persons not having information about them in administrative sources may be interviewed in the LFS. In this case, the missing data are imputed from collected LFS survey data. This may happen at a rate of approximately 0.02 % quarterly.

Finally, the calculated calibration weights are applied. The weights comply with and depend on administrative information, thereby enhancing representation of the target population in the sample.

5. Conclusions

There are several setbacks to consider.

1. Use of some registers requires preparation or revision of legislative framework, which can be time-consuming.
2. As regards data sources, the data must have high quality and correct integration of the data calls for thoughtful verification and analysis.
3. The monthly updating of the LFS sampling frame benefits from a skilful merging of the key administrative sources on the data maintainer side.

Unfortunately, we lack specific results from comparing outcomes with and without administrative data (as obtaining this data took a long time and during the process sample design methodology was changed as well). However, leveraging various administrative sources across several data processing stages has been beneficial for quality of the data collection process, as well as optimization of the weight calculation. Moreover, it helps to understand and explain population change and demographic fluctuations better and allows building more representative sampling frame.