Survey Methodology and Simulation in R. Possibilities and New Methods

Matthias Templ

Technische Universität Wien / Statistik Austria

For a short time ago only few methods for complex survey sampling and estimation were available in R. Fortunately, this is no longer true, and much more sophisticated methods as in any other software are (freely) available nowadays. This is also clearly reflected by the CRAN Task View on Official Statistics and Survey Methodology
(http://cran.r-project.org/web/views/OfficialStatistics.htm).

Various methods and packages have been developed during the FP7-project AMELI (Advanced Methodology for European Laeken Indicators) by the "Viennese project team", where the following corresponding R-packages are already available at the Comprehensive R Archive Network:

- **simFrame**: a framework for simulations in Official Statistics to compare point and variance estimators under different survey designs as well as different conditions regarding missing values, representative and non-representative outliers. It allows to compare (user-defined) point and variance estimators in a simulation environment.

- **simPopulation**: a package to simulate synthetic, confidential, close-to-reality populations for surveys based on sample data. Such population data can then be used for extensive simulation studies in Official Statistics, using simFrame for example.

- **laeken**: a package which provide functions to estimate certain Laeken indicators (at-risk-of-poverty rate, quintile share ratio, relative median risk-of-poverty gap, Gini coefficient) including their variance for domains and stratas based on bootstrap resampling.

- **VIM**: a framework for visualization of the structure of missing values using suitable plot methods. It also provides EM-based multiple imputation using robust methods handling data consisting of continuous, semi-continuous, binary, categorical and/or count variables. The package also comes with a graphical user interface.

After considering practical problems related to complex survey data (finite populations, sampling design, sampling weights, missing values, outliers), the structure, the relationships and the application of these packages are demonstrated using EU-SILC, a complex and popular data set from Official Statistics. In addition, simulations will show the benefits of, e.g., robust semi-parametric methods for estimations related to the income distribution.