Progress report on simulation and synthetic data generation for EU-SILC data

A. Alfons, S. Kraft, P. Filzmoser, and M. Templ

Forschungsbericht CS-2009-6
November 2009

Kontakt: P.Filzmoser@tuwien.ac.at
Progress Report on Simulation and Synthetic Data Generation for EU-SILC Data.*

Andreas Alfons
(Vienna University of Technology)

Stefan Kraft
(Vienna University of Technology)

Peter Filzmoser
(Vienna University of Technology)

Matthias Templ
(Vienna University of Technology & Statistics Austria)

March, 2009

1 Particular Objectives

The aim is to investigate and compare estimation methods in survey statistics in a close-to-reality environment based on realistic datasets from selected European countries.

2 Introduction

Since the aim of the AMELI project is the development and improvement of different methods for the estimation of the Laeken Indicators, a crucial point is the evaluation of these methods. An adequate environment is needed that is as close to reality as possible and that also allows for variations regarding outliers or missing values. Therefore, Monte Carlo simulation studies seem to be most suitable for the investigation of the methodology. As a basis for these simulation studies, populations are needed from which samples can be drawn according to the true sampling design.

*This work was funded by the European Commission within the FP7-project AMELI (project number 217322)
3 First Results

3.1 Simulation of populations

Only in a few cases population data is available. Even if so, such data sets usually cannot be used due to confidentiality reasons. Hence, adequate populations have to be simulated on the basis of current survey data.

Münich and Schürle [2003b] describe a general simulation model for the simulation of populations with only discrete variables. This model was first applied to the German Microcensus data within the DACSEIS project (www.dacseis.de). The model considers the stratification of the survey as well as the household structure within the data. Although some simplifications have been made regarding the correlation structure, Münich and Schürle [2003b] show that the simulated populations reflect most of the dependencies found in the sample data.

For the application to household and budgets surveys, Münich and Schürle [2003a] extended their model by the generation of continuous variables. They presented two different methods for the simulation of income and expenditure variables. Within their first approach, Münich and Schürle [2003a] generated observations for these variables by drawing from lognormal distributions. The parameters of these distributions were estimated from the sample data. The second approach was to classify the continuous variables so that the previously used methods can be applied for simulation. A uniform partition was chosen to retransform the discrete values into continuous ones.

The work of Münich and Schürle [2003a,b] has been the initial point for all developments concerning the simulation of populations. The previous simulation methods have been investigated on the basis of the Austrian SILC data. Due to the great number of different variables - both categorical and continuous - that are important for the computation of the Laeken Indicators and should therefore be included in the simulation, these methods were found to be inadequate. Consequently, the previous simulation methods have been adapted and new methods have been developed. Populations containing six categorical variables together with personal net income and the personal income components have been simulated so far. As a next step, the focus is put on the simulation of the equivalised household income and the corresponding household income components.

The developed methods and the underlying theory are described in a diploma thesis written by Stefan Kraft [Kraft, 2009] and are implemented in R [R Development Core Team, 2009]. Some parts of this work were already presented at the Wiesbaden meeting in November 2008 and the newest developments will be presented at the Olten meeting in June 2009.

3.2 Simulation studies

Using Monte Carlo simulation studies in survey statistics for the evaluation of the quality of estimators has a long tradition. A comprehensive report including a schematic overview of the simulation design used in the DASEIS project can be found in Münich et al. [2003]. They describe the simulations carried out for the DACSEIS project, but without special focus on evaluation of the influence of outliers on the estimation of
indicators. Hulliger and Männich [2006] consider a simulation study to estimate the influence of outliers on the Gini coefficient and the quintile share ratio. In this work they consider representative outliers in the population.

For carrying out simulation studies in the AMELI project, the R package simFrame [Alfons, 2009] is being developed. It implements an object-oriented framework for statistical simulation using 54 classes. The object-oriented implementation gives maximum control over input and output, thus allowing for trustworthy computations [see Chambers, 2008], while at the same time providing clear interfaces for extensions by user-defined classes and methods. The current version is focused on design-based, close-to-reality simulation studies in official statistics. Control classes allow certain proportions of the data to be contaminated or set to missing values. Thereby, different contamination or missing data models are represented by different control classes and the existing framework may easily be extended.

Sampling is another important issue in design-based simulation, and many methods are already implemented in R. The package sampling [Tillé and Matei, 2008] is focused on functions for drawing and calibrating samples. Numerous methods for survey sampling are implemented, including probability proportional to size (PPS) sampling, stratified sampling and balanced sampling. However, the implementation of these methods is very slow in many cases. Some functions for PPS sampling are also available in the package pps [Gambino, 2005]. The package sampling includes the Sampford algorithm [Sampford, 1967] to obtain a sample without replacement and with unequal probabilities.

In simFrame, stratified sampling and sampling of whole groups (e.g., households for EU-SILC data) with equal and unequal probabilities are implemented. More complex sampling methods can be provided by the user. Drawing repeated samples before the actual simulations are carried out reduces computation time dramatically, as computations like stratification are performed only once. Furthermore, the samples can be saved permanently before running the simulation study, which simplifies the reproduction of simulation results and is beneficial if computations fail.

Plot methods with sensible defaults are implemented and an appropriate plot method is selected automatically depending on the structure of the simulation results.

4 Future Tasks

Future tasks include extending the package simFrame to model-based simulation designs and mixtures of model- and design-based simulation.

References


