Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, 1–xx.
http://www.ajs.or.at/

Simulation Tools for Small Area Estimation: Introducing the R-package saeSim

Sebastian Warnholz
Freie Universität Berlin

Timo Schmid
Freie Universität Berlin

Abstract

The abstract of the article in English

Keywords: package, small area estimation, reproducible research, simulation, R.

1. Introduction

Reproducible Research has become a widely discussed topic inside science and also the field of statistics. Thanks to the many mostly open-source tools like the R-language and LaTeX, and also packages like knitr and Sweave, the integration of source code and text is possible and as a problem solved. In that sense the demand for Literate Programming can be if wanted incorporated in the work flow of a scientist. Not only are tools available to make research reproducible, also the demand of making the analysis of articles reproducible is rising. What this means is, that the source-code and data is published alongside an article. However, the requirements in style and clarity of source code are different from the written words in the article itself. This demand *human readable source-code* has already been expressed by Knuth (1984):

> *Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.*

Small Area Estimation is a growing field inside the field of statistics, where simulation studies play an important role. New statistical models are applied in model-based and design-based simulation studies. Considering the demands in reproducibility we want to propose a framework for simulation studies. This framework addresses three demands. First, making tools for data generation available and reusable. Second, unify the process behind simulation studies inside the field of Small Area Estimation. And third, making source-code of simulation studies available, such that it supports the conducted research in a transparent manner. In this article we want to introduce a new package for the R-language addressing these demands. We will show the importance of simulation studies in the field. Introduce the simulation framework. And demonstrate how to use the package to map the simulation as a process to R.

2. Small Area Estimation

What is small area estimation? 2 Models to introduce. FH + BHF. Why are simulation studies important in the field.

3. A simulation framework

Design- vs. model-based simulation, we do not promote one or the other. Figure 1 illustrates the steps in a simulation. Based on a data table, which may be a real or synthetic population or a set of ID variables representing the hierarchy of the data, you start a simulation.

Introducing a simulation study as a data manipulation process. From data generation to estimating models. How can this process be mapped into the R-language. Present the simulation phases as flow-diagram and map the function names to the problem domain. Defining interfaces between phases: data.frame in, data.frame out. What is the difference between a design-based (fixed-population) and model-based (random-population) approach.

Why is this package helping in terms of literate programming (code is written for humans not machines) and reproducible research? Package naming conventions. Use the %>% operator to compose simulation set-ups. Reuse defined scenarios to compose new scenarios: What is a contamination scenario? A standard scenario plus contamination:

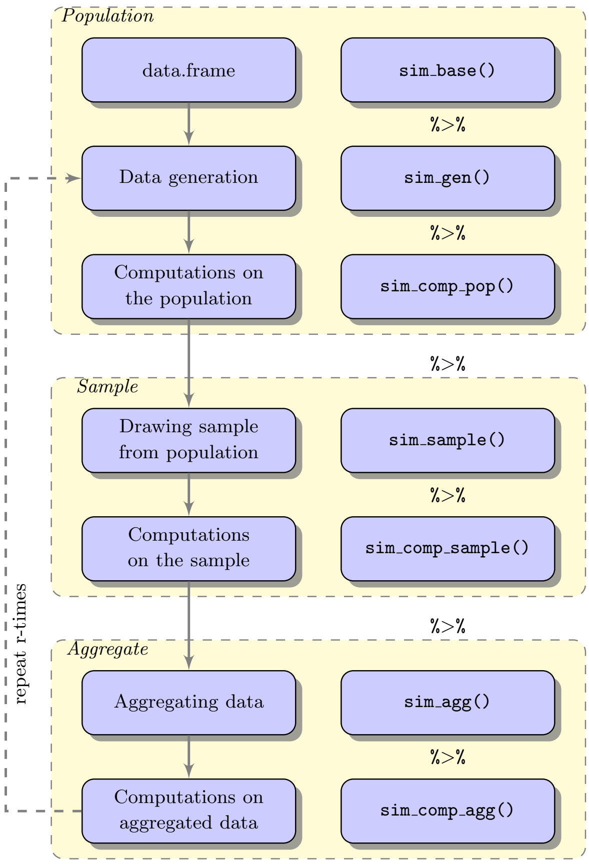

Figure 1: Process of simulation.

asd asd

```
> contaminatedSetup <- standardSetup %>% sim_gen_cont()
```

First develop a scenario outside any looping structures, then run it R-times:

```
> simulationResults <- simSetup %>% sim(R = 500)
```

There is an easy back-end to run simulations in parallel (in Windows with special care):

```
> simulationResults <- simSetup %>% sim(R = 500, parallel = TRUE)
```

4. Data generation

This is not a data generation tool. However, it supports some useful functions to add random variates to the data. Definition of outliers, see Rao (?) – how is it addressed in the package. More than linear models, define any response with:

```
> setup %>% sim_resp_eq(y = g(2 * x) + e)
```

5. Adding steps to a simulation

Is present in unit-level simulation studies. Draw with simple random sampling from the whole population or within cluster/domains. Specify the sample size as integer or fraction and add weights if necessary. Wrappers around dplyr::sample_n and dplyr::sample_frac. Adding user specified functions to the simulation process is what separates this package from a mere data generation tool. The interface is simple: Add functions with one argument which is the data at that moment and which return the modified data. I am not sure if this and the sampling section is necessary – maybe it should just be summarized in the overview section...

6. Case study

7. Outlook

Use this package to share and publish simulation studies alongside papers. Contribute to the package to make your ideas available. Contribute to the package and make your whole simulation study available.

References

Affiliation:

Sebastian Warnholz
Department of Economics
Freie Universität Berlin
D-14195 Berlin, Germany
E-mail: Sebastian.Warnholz@fu-berlin.de
URL: http://www.wiwiss.fu-berlin.de/fachbereich/vwl/Schmid/Team/Warnholz.html