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 R Core Team (2014) and LaTeX, and also packages like knitr (Yihui 2013) and Sweave (Leisch 2002) and more recently rmarkdown (Allaire, McPherson, Xie, Wickham, Cheng, and Allen 2014), 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. This means, 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.* (Knuth 1992, p.99)

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

The objective of small area estimation is to produce reliable statistics (means, quantiles, proportions, etc.) for domains where little or no sampled units are available. Groups may be areas or other entities, for example defined by socio-economic characteristics. The demand for such estimators is rising as they are used for fund allocation, educational and health programs (Pfeffermann 2013). As direct estimation of such statistics are considered to be unreliable, methods in small area estimation try to improve the domain predictions by borrowing strength from neighboured or *similar* domains. This can be achieved by using additional information from census data to assist the prediction for non-sampled or domains with little information. For the purpose of this article we will introduce two basic models frequently used in small area estimation, the unit-level model introduced by Battese, Harter, and Fuller (1988) and the area-level model introduced by Fay and Herriot (1979).

As applications in small area estimations rely often on sensitive information unit-level information is not always available. Instead only aggregates, or rather the direct estimators are. In such situations area-level models are used. The area-level model introduced by Fay and Herriot (1979) is build on a sampling model:

$$y_i = \mu_i + e_i,$$

where y_i is a direct estimator of a statistic of interest μ_i for an area i with $i = 1, \ldots, D$ and D being the total number of areas. The sampling error e_i is assumed to be independent and normally distributed with known variances $\sigma^2_{e,i}$, i.e. $e_i|\mu_i \sim N(0, \sigma^2_{e,i})$. The model is modified with the linking model by assuming a linear relationship between the true area statistic μ_i and some auxiliary variables x_i:

$$\mu_i = x_i^\top \beta + v_i,$$

with $i = 1, \ldots, D$. Note that x_i is a vector containing area-level (aggregated) information for P variables and β is a vector $(1 \times P)$ of regression coefficients describing the (linear) relationship. The model errors v_i are assumed to be independent and normally distributed, i.e. $v_i \sim N(0, \sigma^2_v)$. Furthermore e_i and v_i are assumed to be independent. Combining the sampling and linking model leads to:

$$y_i = x_i^\top \beta + v_i + e_i. \tag{1}$$

Model 1 is effectively a random-intercept model where the distribution of the error term e_i is heterogeneous and known.

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:

asd asd

asd asd

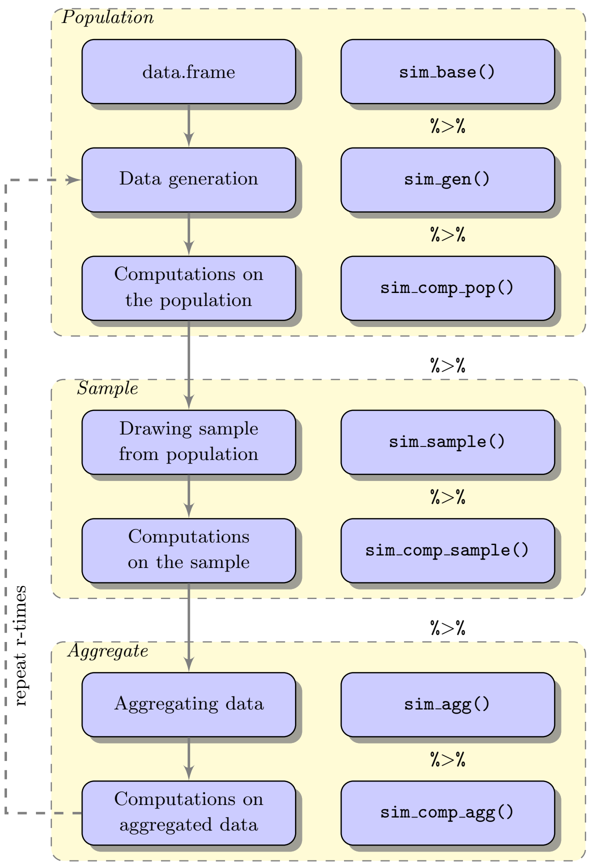

Figure 1: Process of simulation.

asd asd

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. Case study

5. 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

Allaire J, McPherson J, Xie Y, Wickham H, Cheng J, Allen J (2014). *rmarkdown: Dynamic Documents for R*. R package version 0.3.3, URL http://CRAN.R-project.org/package=rmarkdown.

Battese GE, Harter RM, Fuller WA (1988). "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association*, **83**(401), 28–36.

Fay R, Herriot R (1979). "Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association*, **74 (366)**, 269–277.

Knuth DE (1992). *Literate Programming*. CSLI, Stanford.

Leisch F (2002). "Sweave, Part I: Mixing R and LaTeX." *R News*, **2**(3), 28–31. URL http://CRAN.R-project.org/doc/Rnews/.

Pfeffermann D (2013). "New Important Developments in Small Area Estimation." *Statistical Science*, **28**(1), 40–68. doi:10.1214/12-STS395. URL http://dx.doi.org/10.1214/12-STS395.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Yihui X (2013). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC. ISBN 978-1482203530, URL http://yihui.name/knitr/.

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

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/

Volume VV *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd