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

S. Warnholz T. Schmid

We are very grateful to the Editor and the two reviewers for their constructive comments. These have been very helpful in preparing the revised version of this paper. We have done our best to incorporate them in the revision. The following notes explain how we addressed each comment.

Replies to reviewer A

The manuscript introduces the R package saeSim, which simplifies writing code for simulation studies in small area statistics. The manuscript is enjoyable to read, but contains some shortcomings as well as several typos and other mistakes. The software seems useful for small area statisticians and could be a valuable contribution to the community. I particularly like the notion of treating simulation as a data manipulation process.

There are only minor issues that need to be addressed.

- Page 2: The authors cite Kolb (2013) for synthetic data generation. They should also mention what is available in R, for instance the approach by Alfons, Kraft, Templ & Filzmoser (2011, Statistical Methods & Applications) that is implemented in package simPop.

  Reply: We now reference Alfons, Kraft, Templ & Filzmoser (2011, Statistical Methods & Applications) together with Kolb (2013). The R-package simPop is now mentioned in the paragraph where we also cite simFrame, also on page 2.

- Page 2: The authors mention NUTS levels without explaining what they are. Please add a short explanation, including what the abbreviation NUTS stands for.

  Reply: We deleted the reference to the NUTS levels. After the revision the paragraph doesn't lose relevant information to follow the common theme. The revised version is now as follows:

  Survey statistics are used, for example, in order to deliver specific indicators as a basis for economic and political decision processes. Of special interest here are regional or group specific comparisons (cf. Schmid and Mnnich 2014). Surveys which are utilized to report these regional indicators, however, are generally designed for larger areas. Hence sample information on more detailed levels, e.g. municipalities, is hardly available so that classical estimation methods (direct estimators) may lead to high variances of the estimates (cf. Ghosh and Rao 1994). In this case small *area estimation methods may reveal highly improved results for the target estimates. Small area estimation has become more and more attractive over the last decade:*

- Page 5, last paragraph: I'm not sure which line in the simple example for the pipe operator is more readable. The line "sum(1:10)" translates into "sum of 1 to 10". The line "1:10 %>% sum" tanslates into "generate 1 to 10, then take the sum". I'm sure that the authors can find a simple example that is better suited to illustrate their point.

  Reply: We acknowledge that we have to be more specific here. We changed the example to: `colMeans(matrix(rnorm(10), ncol = 2))` which translates to `rnorm(10) %>% matrix(ncol = 2) %>% colMeans`.

- Page 6 and 7, code examples: The authors load package saeSim in each of the two code examples. Yet there is another code chunk before those examples which use functionality from the package. If readers try to execute all code in the order it appears in the paper, they will get error messages. I suggest to load the package in the beginning of Section 4 (in a separate chunk), and to remove this line from the two examples to focus on the relevant code.

  Reply: We load now saeSim (and magrittr) in the beginning of section 4. We checked the code again, so that all expressions are valid when executed in the order in which they appear in the article. We also made the seed available to the reader. Furthermore the package sae produced naming conflicts as it loads a lot of dependencies. We do not load the package but reference to it with the '::' operator.

- Page 8, last line: The authors write "... we store these data tables...". Are those data frames or data tables? In R, this can be confusing due to the data.table package.

  Reply: To avoid any confusion we changed it into 'data frame'. The paragraph is quoted as a reply to you next remark.

- Page 9: The authors state that they store some information as attributes to the main data frame. Attributes can be a dangerous thing in R as there are some functions for manipulating data that may remove any attributes. Please give a discussion on this issue and provide more motivation for using attributes.

  Reply: We agree that relying on attributes may be dangerous in R. The problem is that in some situation we have to process a single data frame plus some meta data. An alternative is to make that meta data available by taking advantage of R's lexical scoping rules. However, this may be a dangerous path as the functions, which need the meta data, are no longer self contained. They depend on the specific state of the objects instead of their arguments. We want to avoid this situation. Another possibility is to make meta data available only to the functions in which they are needed by using closures. This ensures that the functions are self contained but adds unnecessary complexity because advanced functional programming techniques are not "mainstream" for R users. We also could design an own *data class* which is more formal to ensure type consistency (for example using a S4-class). However, we tried to design saeSim such that it can be used with as many existing tools as possible. We think this is possible by sticking to the "data.frame in, data.frame out" philosophy. Thus we decided to support functions which preserve attributes. All present functions in the package will preserve attributes of the main data frame.

We added the following paragraph to section 4.2:

Before we begin to construct the simulation setup, we store these data frames as attributes to the population data. This allows us to process meta data alongside the main data frame. It is important to note that not all functions for manipulating data frames in R preserve these attributes. Users of saeSim have to keep this in mind when they implement new functions. Defining the interfaces between components differently is one possibility to avoid the usage of attributes. This can be done, for example, by using generic vectors or S4 classes instead of data frames. However this will add complexity to the process of data manipulation underlying the package which we try to avoid by following the paradigm: data frame in, data frame out. Thus all functions in saeSim preserve the attributes of the main data frame.

- Page 13: To the best of my knowledge, the preferred format for CRAN links is http://CRAN.R-project.org/package=saeSim (also note the capitalization of CRAN and R).

  Reply: This has been corrected.

- Page 13: For non-representative outliers, the authors give the explanation ”outliers are part of the sample but not the population”. This may be confusing to some readers. I suggest that they add ”e.g., incorrectly recorded values”.

  Reply: This has been added.

Comments regarding language are given in the following. There may be more mistakes, hence I recommend that the authors ask a colleague to proofread the manuscript (preferably an English native speaker).
Reply: We apologise for the mistakes in the article. We did a careful double check during the revision. In addition, we followed your suggestion to check the manuscript by an English native speaker.

- Page 1, abstract and first paragraph: ”has been substantially grown” to ”has grown” – done

- Page 3, line -8: ”is build on a sampling model” to ”is built on a sampling model” – done

- Page 4, line -2: ”different point of views” to ”different points of view” – done

- Page 6, lines 1-2: ”In saeSim, we rely on this operator, although all functions can be used without, we strongly recommend to use it.” to ”In saeSim, we rely on this operator. Although all functions can be used without it, we strongly recommend its usage.” – done

- Page 6, second paragraph: ”... which has a data.frame as input as well as return value.” The correct ”... which has a data.frame as input as well as as return value.” is cumbersome since as appears twice in a row, so I suggest to change this to ”... which has a data.frame both as input and as return value.” – done

- Page 10, first paragraph: ”This focus on the definition of each component and on meeting the convention of the defined interface is the intended approach.” I’m confused by this sentence, please rephrase.

  Reply: We acknowledge that the sentence may create some confusion. Therefore we have deleted the sentence to improve the readability of the paragraph. In the revised version of the paper we say the following:

  In the next step, we define components which add the estimates of interest to the data. Here we compute the direct estimator of the mean income in each domain and the EBLUP under the BHF model. Although this could be done in one step, we separate the two computations to illustrate how to combine several estimations and how to define each component independently of the others. This automatically organises the simulation and each component is arranged using the simulation framework. Hence we define two functions, one for adding the direct estimates and one for adding the EBLUP.

- Page 13, second paragraph: ”version-control” to ”version control”. – done

- Page 13, second paragraph: ”it’s purpose” to ”its purpose”. – done

Replies to reviewer C

The package appears to be useful for most statisticians working with small area methods. It is a good idea to build the framework using the idea of process of data manipulation. This makes the package easy to use. The manuscript is interesting and useful, and I find it acceptable for publishing after minor revision.

- The authors should mention - probably in Section 4 - that in order to repeat a simulation with identical results, the seed used in the original simulation has to be known.

  Reply: We agree with this comment. We have added this information right before we start illustrating the model-based simulation. We have also added the seed in the code. In the first version, the seed was set in a hidden chunk. In Section 4.1 in the revised version of the paper we say the following:

  In this case the base-component is a data frame with an id variable named idD and constructed with the function base_id. Any random number generator in R can be used. However, we have normally distributed variates, for which some predefined functions are available in the package. For the reproducibility of the following results we also set the seed to 1. The seed is not part of a simulation setup in saeSim but needs to be defined by the researcher.

There are some grammatical errors:
Reply: We apologise for the errors in the article. We did a careful double check of the paper. In addition, the manuscript was checked by an English native speaker.

(1) abstract: remove 'been' from "...has been substantially grown" – done

(2) Introduction: remove 'been' from "...has been substantially..."– done

(3) Introduction (last two lines on page 1): rephrase "...reproducible research aims that the full output of the academic research,..." using "...aims at the availability of the full... " – done

(4) Replace 'neighboured' by 'neighbouring' – done

(5) At the end of page 1, remove '-' in "source-code" – done

(6) In the beginning of section 2 (second line), replace 'less' by 'few' in "...for domains where less or no sampled..." – done

(7) Below equation (1), replace 'build' by 'built' in "...introduced by Fay and Herriot (1979) is build..." – done