Small Area Estimation in R with Application to Mexican Income Data

Ann-Kristin Kreutzmann (ann-kristin.kreutzmann@fu-berlin.de)[1], Sören Pannier (soeren.pannier@fu-berlin.de), Natalia Rojas-Perilla (natalia.rojas@fu-berlin.de), Timo Schmid (timo.schmid@fu-berlin.de), Matthias Templ (templ@statistik.tuwien.ac.at)[2] and Nikos Tzavidis (N.TZAVIDIS@soton.ac.uk)[3]

Keywords: SAE, small area, EBP, poverty, inequality

1. Introduction

In the last decades policy decisions are often based on statistical measures. The more detailed this information is, the better is the basis for targeting policies and evaluating policy programs. For instance, the United Nations suggest more disaggregation of statistical indicators for monitoring their Sustainable Development Goals and also the number of National Statistical Institutes (NSIs) that notice the need of more disaggregated statistics is increasing. Dimensions for disaggregation can be characteristics of the individuals or households like sex, age or ethnicity, economic activity or spatial dimensions like metropolitan areas or districts. Primary data sources for variables that are used to estimate statistical indicators are national household surveys. However, sample sizes are usually small or even zero at disaggregated levels. Therefore, direct estimators based only on survey data can be unreliable or not available for small domains. While the option of more specific surveys is costly, model-based methodologies for dealing with small sample sizes can help to obtain reliable estimates for small domains. The so-called *Small Area Estimation* (SAE) methods [1,2] link survey data that is only available for a proportion of households with administrative or census data available for all households in the area of interest. Even though a wide range of SAE methods is proposed by academic researchers, these are, so far, applied only by a small number of NSIs or other practitioners like the World Bank. This gap between theoretical possibilities and practical application can have several reasons. One reason can be the lack of suitable statistical software. The free software environment R helps to counteract this issue since researchers can make their codes available to the public via packages. Thus, new methods can reach the practitioner faster than with non-free software. The next two sections summarize which packages are already available and what could be improved in the future.

2. Review of R packages for SAE

Traditional SAE methods lead to predictions for simple indicators like means and totals that are called empirical best linear unbiased predictors (EBLUP). For the estimation of these indicators several software packages exist in R. Under the heading SAE in the Comprehensive R Archive Network (CRAN) Task View: Official Statistics & Survey Methodology [3], three packages are mentioned besides the packages **nlme** [4] and **lme4** [5] for mixed effects models. The **rsae** package by Schoch [6] provides functions for robust small area estimation using basic unit- or area-level models for predicting area means. Also based on unit- and area-level models are the functions in the package **hbsae** [7]. The model can be fit either by using restricted maximum likelihood or using hierarchical Bayes methods. The package **JoSAE** [8] covers functions for generalized regression estimators and the unit-level EBLUP. Other packages like **sae2** [9] and **saery** [10] focus on functions

[1] Freie Universität Berlin, Germany

[2] Technische Universität Wien, Austria

[3] University of Southampton, United Kingdom

for estimators based on an area-level model with time effects following Rao and Yu [11]. Another aspect is considered in package **saeSim** [12]. Instead of estimation methods, it provides tools for simulating data that is required in SAE.

A widespread application of SAE is the estimation of poverty and inequality indicators for small domains and their presentation on geographical maps [13]. For this application the estimation of means and totals is not sufficient any more, since most poverty indicators are complex, non-linear indicators like the Foster-Greer-Thorbecke indicators [14] or the Gini coefficient [15]. But also for the estimation of quantiles more sophisticated methods are needed. Popular SAE approaches for the estimation of complex indicators are the Empirical Best Predictor (EBP) estimation method [16], the World Bank method [17] and the M-Quantile approach [18]. While the M-Quantile approach and the World Bank method are not yet implemented in R, the R package **sae** provides the EBP approach among a wide range of other commonly used SAE methods [19].

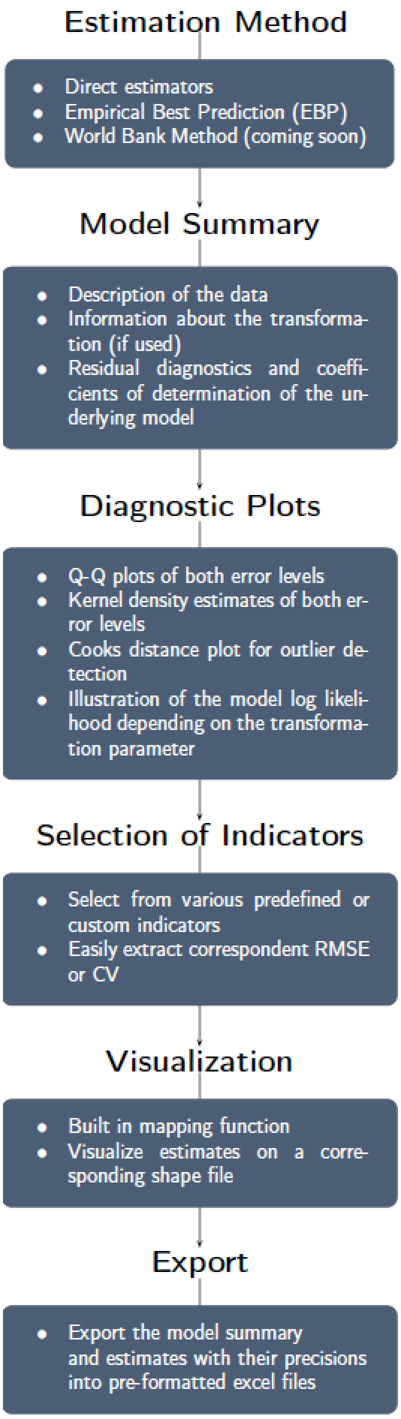
Estimation Method
Direct estimators
Empirical Best Prediction (EBP)
World Bank Method (coming soon)
Model Summary
Description of the data
Information about the transformation (if used)
Residual diagnostics and coefficients of determination of the underlying model
Diagnostic Plots
Q-Q plots of both error levels
Kernel density estimates of both error levels
Cooks distance plot for outlier detection
Illustration of the model log likelihood depending on the transformation parameter
Selection of Indicators
Select from various predefined or custom indicators
Easily extract correspondent RMSE or CV
Visualization
Built in mapping function
Visualize estimates on a corresponding shape file
Export
Export the model summary and estimates with their precisions into pre-formatted excel files

Figure 1: User support in package emdi

3. THE R PACKAGE EMDI

All R packages mentioned in Section 2 together enable the estimation of indicators using different SAE methods. However, the packages do not support the user beyond the estimation. For instance, as shown in Molina and Marhuenda [19], the estimation results using the **sae** package are sufficient for producing summary statistics, diagnostic plots and data frames with desired predictions. However, the creation cannot be obtained by a single command and must be conducted manually and no further information, e.g. about data, transformations and model fit is given. Thus, the presentation of results is not illustrative for users at a first glance. An example how the user-comfort for the application of SAE methods in R can be improved is the package **emdi** [20]. The greatest benefit of this package is the support of the user beyond the estimation i.e. in the investigation and presentation of obtained results. This is achieved by using the S3 object system [21]. By assigning classes to returns it is defined how the objects of this class look like [22]. Furthermore, methods can be defined that work differently depending on the class of the object. The estimation methods in package **emdi** comprise direct estimation and the EBP approach for predefined indicators with the possibility to add custom indicators. The resulting objects are of class `"emdi","direct"` and `"emdi","model"`, respectively. Figure 1 shows what the methods in package **emdi** provide for the user. For instance, the `summary` of an object of class `"emdi", "model"` gives a description of the data, residual diagnostics and the coefficients of determination of the underlying models as well as information of the transformation if one is used. Graphical diagnostics checks can be received using the `plot` function. In Figure 2 five plots are shown that are provided if the Box-Cox transformation is selected. The data used in this example is data from municipalities of the Mexican state Estado de México (EDOMEX). It is provided by the national institute of statistics and geography (INEGI) that is responsible for performing the national population and housing census every ten years and the household income and expenditure survey around every two years.

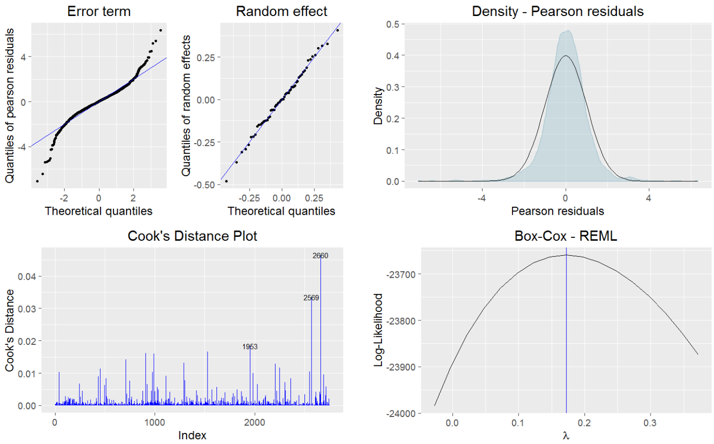

Figure 2: Diagnostic plots obtained by function `plot`: QQ-plots and Kernel density estimations of both error levels, a Cooks distance plot for outlier detection and an illustration of the model log likelihood depending on the transformation parameter.

Since the estimation automatically returns a set of predefined indicators a function called `estimators` helps to select single indicators or groups of indicators. It also returns the root mean squared error and the coefficient of variation (CV) if selected. Furthermore, a mapping tool for plotting the estimated indicators on their geographic regions enables the creation of high quality visualizations as shown for the Head Count Ratio in EDOMEX in Figure 3. The user also has the opportunity to export an informative output to excel.

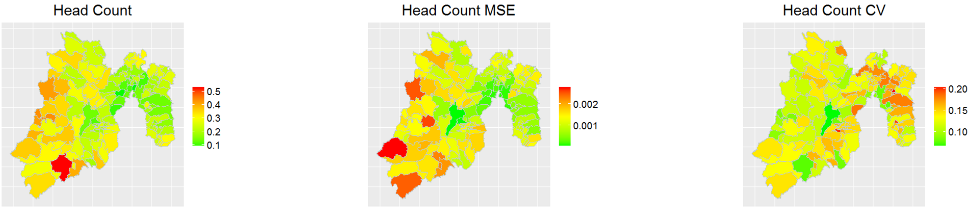

Figure 3: Maps for the point estimates, the mean squared error and the coefficient of variation of the Head Count Ratio.

4. Conclusions

The functionalities of package **emdi** demonstrate how the application of SAE can be supported using R packages and the possibilities of R object systems. Thus, also users with only basic knowledge of the methods can apply them in practice and easily check if the method is suitable. For the future, it is desirable that more SAE methods can be applied easily using R. Thus, the gap between theoretical possibilities for the estimation of disaggregated indicators and the practical application may become smaller.

REFERENCES

[1] J.N.K. Rao and I. Molina, Small Area Estimation, John Wiley & Sons (2015).

[2] D. Pfeffermann, New important developments in small area estimation, Statistical Science 28(1) (2013), 40-68.

[3] M. Templ, CRAN Task View: Official Statistics & Survey Methodology, Version 2016-12-01 [accessed: 06.01.2017],
URL: *https://cran.r-project.org/web/views/OfficialStatistics.html*

[4] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar and R Core Team, nlme: Linear and nonlinear mixed effects models, R package version 3.1.-122.
URL: *https://cran.r-project.org/web/packages/nlme/index.html*

[5] D. Bates, M. Mächler, B. Bolker and S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67(1) (2015), 1-48.

[6] H. Boonstra, hbsae: Hierarchical Bayesian Small Area Estimation, R package version 1.0.
URL: *https://CRAN.R-project.org/package=hbsae*

[7] T. Schoch, rsae: Robust small area estimation, R package version 0.1-5.
URL: *https://cran.r-project.org/web/packages/rsae/index.html*

[8] J. Breidenbach, JoSAE: Functions for some Unit-Leve Small Area Estimators and their Variances.
URL: *https://CRAN.R-project.org/package=JoSAE*

[9] R.E. Fay and M. Diallo, sae2: Small Area Estimation: Time-series Models, R package version 0.1-1.
URL: *https://CRAN.R-project.org/package=sae2*

[10] M. Esteban Lefler, D. Morales Gonzalez and A. Perez Martin, saery: Small Area Estimation for Rao and Yu model, R package version 1.0.
URL: *https://CRAN.R-project.org/package=saery*

[11] J.N.K. Rao and M. Yu, Small-area estimation by combining time-series and cross-sectional data, The Canadian Journal of Statistics 22(4) (1994), 511-528.

[12] S. Warnholz and T. Schmid, saeSim: Simulation Tools for Small Area Estimation, R package version 0.7.0.
URL: *https://CRAN.R-project.org/package=saeSim*

[13] The World Bank, More than a pretty picture: using poverty maps to design better policies and interventions, Report, The International Bank for Reconstruction and Development – World Bank (2007).

[14] J. Foster, J. Greer and E. Thorbecke, A class of decomposable poverty measures, Econometrica 52(3) (1984), 761-766.

[15] C. Gini, Variabilità e mutabilità: Contributo allo studio e delle distribuzioni e relazioni statistiche, Studi Economico-Giuridici della R, Universitádi Cagliari.

[16] I. Molina and J.N.K. Rao, Small area estimation of poverty indicators, The Canadian Journal of Statistics 38(3) (2010), 369-385.

[17] C. Elbers, J. Lanjouw and P. Lanjouw, Micro-level estimation of poverty and inequality, Econometrica 71(1), 355-364.

[18] R. Chambers and N. Tzavidis, M-quantile models for small area estimation, Biometrika 93(2) (2006), 255-268.

[19] I. Molina and Y. Marhuenda, sae: An R package for Small Area Estimation, The R Journal 7(1) (2015), 81-98.
URL: *http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf*

[20] S. Pannier, A.-K. Kreutzmann, N. Rojas-Perilla, T. Schmid, M. Templ and N. Tzavidis, emdi: Estimating and Mapping Disaggregated Indicators, R package version 1.0.0.
URL: *https://CRAN.R-project.org/package=emdi*

[21] J.M. Chambers, Statistical models in S, Champman &Hall (1993).

[22] F. Leisch, Creating R packages: A tutorial, in P. Brito: Compstat 2008 – Proceedings in Computational Statistics, Physica Verlag (2008).