A Framework for Producing Small Area Estimates Based on Area-Level Models in R

Replies to the comments raised by the review process

Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Schmidt, Nicola Salvati, Timo Schmid

First we like to thank you for your time, and constructive suggestions that you have made. These have been very helpful in preparing the revised version of this paper. We have done our best to respond to all comments. Below we respond to each of your comments. Your comments are indented, followed by our responses.

Replies to the editor

> Overview:
> This article is a welcome addition to the small estimation literature, describing the facilities
> of R package emdi for estimating area-level models. Package emdi is one of the more versatile
> R packages for small area estimation, with a slight focus on poverty mapping. The package
> has been extended with support for area-level models, which are an important class of models
> that are widely applicable since they do not rely on access to microdata. The article extends
> the documentation of the package, which is already quite good and comprehensive.

Reply: Thank you for your positive comments.

> Below I make some suggestions for improvements to the article and package code.
>
> Article:

> p.2, Table 1 and 2nd bullet point: Regarding 'adjusted variance estimation' it would be fair
> to add that packages employing a Bayesian approach do not need this as any point estimate
> of the variance parameter (e.g. posterior mean or median) will be strictly positive in that
> framework.

Reply: Thanks for your constructive suggestion. We added this clarification to the 2nd bullet point.

> p.4: it is noted that the Slud-Maiti bias correction is not applicable to out-of-sample domains
> because of its dependence on the shrinkage factor. From the equation given I do not understand
> this, as the shrinkage factor (gamma) simply goes to zero in the limit of infinite sampling
> uncertainty, and the equation still seems to be well-defined in that limit.

Reply: Thank you for raising this important issue. We acknowledge that we have to be more precise and that your finding is completely correct. Also Slud and Maiti (2006) state on page 241 that the shrinkage factor equals zero if $n_j = 0$ (out-of-sample domain). In that case the equation

$$\hat{\theta}_i^{\text{FH, Slud-Maiti}} = \exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\hat{\sigma}_u^2\left(1 - \hat{\gamma}_i^{*\text{log}}\right)\right\}$$

reduces to

$$\hat{\theta}_i^{\text{FH, Slud-Maiti}} = \exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\hat{\sigma}_u^2\right\}.$$

We discussed, if we should implement this option in emdi. For all of the other included methods it is possible to estimate the point (EBLUP) and mean squared error (MSE) estimator. Therefore

we also a had a look at the corresponding MSE estimator suggested by Slud and Maiti (2006) (p. 248, eq. 23). The estimator and its derivation are very extensive. The transfer of the equation for out-of-sample domains goes beyond the scope of our paper. Even if we proceed as in the case of the point estimator, i.e. setting the shrinkage factor to zero, a lot of parts of the equation cancel out and we do not know the properties of the remaining component of the estimator. All of the other included estimation methods in emdi are based on publications which clearly describe the properties of the estimators and/or test the estimators with the help of simulation studies or real data applications. For this reason and to stay in consistency with the rest of the package, we decided not to incorporate the out-of-sample option for the estimator suggested by Slud and Maiti (2006). Another reason is that emdi already contains the crude back-transformation option with which it is possible to receive estimates for out-of-sample domains when applying a log transformation.
But your finding is correct and we have rephrased the sentence in the paper.

> p.14: the 'Wald test statistic' is only approximately chi-squared distributed, as the variances
> of the direct and model-based estimates are dependent (more so if there are many observations
> per domain)

Reply: Thank you for pointing this out. We added 'approximately' to the sentence.

> I think that the data example provided is a bit unfortunate. It has a very large R-squared
> of 0.92 (see p.12, and why is the adjusted R-squared even higher?), which is not typical of
> most (area-level) SAE applications. Relatedly the model-based estimates are quite close to
> the direct estimates, and it is hard to see any differences in the map plot of the point estimates
> on p.16.

Reply: Thanks for this interesting comment. The data example is indeed a very pleasant example. However, we think that the data is sufficient for example data in the R package because of following reasons:

- The only purpose of the data is showing the functionality of the functions. The idea of the example data is not to show a real application.
- Even though the coefficient of determination is very high, it can happen in real applications. To give an example, we refer to Kreutzmann et al. (2021). They estimate household net wealth at two different regional levels in Germany with a FH R^2 of 92% for the one model, and of 89% for the second model.

With regards to the questions of the higher adjusted R^2, we agree that the naming is confusing. The R^2 in the summary output is actually the adjusted R^2 of a standard linear model. The adjusted R^2 in the summary output is the FH R^2 following Lahiri and Suntornchost (2015). According to the paper, the standard adjusted R^2 underestimates the true adjusted R^2 and is thus the more conservative measure compared to the proposed modified R^2 for FH models. We addressed your important finding by changing the naming of the different two options of the R^2 to $AdjR2$ and FH_R2.

> Package R code:
>
> The package is very well documented, and arguments are checked thoroughly with informative
> messages in case of unexpected inputs.

Reply: Thank you for your positive comment.

> The output objects seem to contain both raw residuals and standardized residuals. In order
> to keep the output objects smaller, it might be better to only store the raw residuals and add
> a 'get'-function for the standardized residuals.

Reply: Thank you for the thorough review of the returned objects. The argument you are raising is very justified, and your solution would reduce the consumed memory. However, as often it is a trade of between computation and memory. As we are using the standardized residuals in multiple post-estimation analysis and also assume the std. residuals to be used frequently by researchers. From our point of view the excess memory consumption is justified. Especially in the context of area level models the residual vectors are rather short and thus only consume very little additional memory, e.g., a numerical vector of length 1000 consumes about 8kB, but even a much larger numerical vector of length 1 million would only contribute approximately 8MB to the memory usage.

> Logical operators '|' and '||' seem to be used interchangeably in conditional statements. Gen-
> erally '||' should be preferred in such cases. The same holds true for '`&`' and '`&&`'.

Reply: Thanks for the careful observation. We changed the logical operators where reasonable for higher code consistency. The changes can be seen via the linked pull request:
https://github.com/SoerenPannier/emdi/pull/44

> Often, out-of-sample domains are omitted by indexing such as
> `eblup_data$Direct[framework$obs_dom == TRUE].`
> Here the '`== TRUE`' is redundant as obs_dom is already a logical vector.

Reply: Thank you for your profound checking of these comparisons. The point you are making is absolutely valid, and also something that was discussed extensively in our developing team. The point you are making is touching the conflict between less code and better readable code. Your suggested change would reduce the amount of code to read and manage, and would also slightly improve the speed of the software, but would reduce readability of the code be developers new to the package. As the community of contributors to emdi keeps growing we decided to prioritize readability and understandability of the code. Hence, we prefer the explicit comparison to `TRUE`.

> The matrix algebra code in the package could be improved in many places, at least performance-
> wise. For example, dense diagonal matrices are constructed explicitly, even identity matrices,
> whereas R contains many shortcuts to avoid the need for such explicit dense diagonal ma-
> trices (e.g. '`diag←`' or broadcasting using vector * matrix or matrix * vector, etc.). There
> are many places where a matrix is explicitly inverted as in `solve(A)`, whereas generally it is
> recommended to either use `solve(A, rhs)`, or use explicit matrix decompositions, e.g. using
> `chol()` in the case of symmetric matrices. This can also improve numerical robustness.
> The code is now much like one would write the equations, which obviously has the advantage
> of being easier to read. However, the main disadvantage is that the package will not scale well
> to large input data with thousands of domains, say.
> If it is to be applicable to larger data sizes, the matrix algebra code deserves more attention
> in the long-term development of the package.

Reply: Thank you for raising this important point. Your concerns with the momentary over-usage of the "solve"-function are fully justified. The issues with the scalability of our model estimations is an issue we are aware of and aim to improve in the future. For better transparency we added the points you are raising into the issues of the github-repository
(https://github.com/SoerenPannier/emdi/issues/48).

References

Kreutzmann, A., P. Marek, M. Runge, N. Salvati, and T. Schmid (2021). The Fay–Herriot model for multiply imputed data with an application to regional wealth estimation in germany. *Journal of Applied Statistics 49*(13), 3278–3299.

Lahiri, P. and J. Suntornchost (2015). Variable selection for linear mixed models with applications in small area estimation. *The Indian Journal of Statistics 77-B*(2), 312–320.

Slud, E. and T. Maiti (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society B 68*(2), 239–257.