National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

19-21

Estimating the RMSE of Small Area Estimates Without the Tears

Diane Hindmarsh and David Steel

Copyright © 2021 by the National Institute for Applied Statistics Research Australia, UOW. Work in progress, no part of this paper may be reproduced without permission from the Institute.

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

stats

MDPI

Research Paper

Estimating the RMSE of small area estimates without the tears

Diane Hindmarsh [1,2,*], David Steel [1]

1 Senior Professor, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, Faculty of Engineering and Information Sciences, University of Wollongong; dsteel@uow.edu.au
2 Honorary Fellow, National Institute for Applied Statistical Research, Australia, University of Wollongong; Lead Analyst, NSW Patient Survey Program, Bureau of Health Information, St Leonards, diane.hindmarsh@uow.edu.au
* Correspondence: diane.hindmarsh@uow.edu.au

Abstract:

Small area estimation (SAE) methods can provide information that conventional direct survey estimation methods cannot, but the use of SAEs is still very limited. One way to increase the use of SAE methods is to reduce the complexity of estimating the root mean square error (RMSE) of the estimates. This paper investigates the parametric bootstrap and the estimated RMSE produced as part of the output from SAS procedures as options. After showing that the estimated RMSE from the MIXED procedure in SAS and an estimated RMSE using a parametric bootstrap are similar to the published estimated RMSEs for the corn data in Battese, Harter & Fuller [9], the parametric bootstrap is used to estimate the RMSEs of SAEs for current smoking in males for local government areas (LGAs) obtained from the NSW Population Health Survey in Australia using EBLUP and EBP methods. For the EBLUP, the estimated RMSEs for the parametric bootstrap are similar to the estimated RMSEs created by the SAS MIXED procedure, except when the sample size for the small area is less than 5. Below this level the parametric bootstrap estimate gives a better estimate of RMSEs. For the EBP, the estimated bootstrap RMSEs were only noticeably different to RMSE estimates produced by the GLIMMIX procedure for LGAs with no respondents in the sample. Estimates of RMSE from the GLIMMIX procedure can be used for small area estimation, with parametric bootstrap producing better estimates of RMSEs for out-of-sample areas.

Keywords: Small Area Estimation; SAE; Empirical Best Predictor; MSE, parametric bootstrap, model-based survey estimation

Citation: Hindmarsh, D; Steel, D.; Estimating the MSE of small area estimates without the tears. *Stats* **2021**, *4*, Firstpage–Lastpage. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Local estimates of health risk factors are important for health promotion, service delivery, evaluation and planning purposes. The New South Wales Population Health Survey (NSWPHS) [1] is a large telephone survey that produces estimates of health risk factors on an annual basis for the Australian state of New South Wales (NSW) and for the broad health-administrative areas in NSW. Due to the small and sometimes non-existent sample sizes at the local government area (LGA) level, conventional direct survey estimates can only be used to create results with reasonable standard errors if data are aggregated over multiple years or to broader geographic regions. In order to create local estimates of health risk factors at the LGA level on an annual basis, small area estimation (SAE) methods are needed.

There are two major hurdles to implementing small area estimation methods as part of routine analysis of surveys such as the NSWPHS. One is determining the appropriate covariates to include in the model and the second is obtaining estimates of the reliability associated with the small area estimates, as measured by the root mean square error (RMSE).

A previous paper [2] provides an assessment of SAE methods based on the empirical best linear unbiased predictor (EBLUP) for linear mixed models and the best empirical predictor (EBP) for generalized linear mixed models that can be used to create LGA estimates for health risk factors from the NSWPHS. The paper also suggests how to determine the best set of covariates and obtain estimates of bias. This current paper provides background to the method of estimating the RMSE used in [2].

Providing estimates of RMSEs is important as the RMSE indicates the confidence users can have in the estimates themselves. Estimation of the RMSE of the EBLUP consists of three parts [3] and the estimation of the RMSE for a EBP is even more complex. It requires several assumptions and approximations that require either Monte Carlo methods or maximum penalised quasi-likelihood (MPQL) and REML methods [4]. Implementing these methods require high-level understanding of the underlying processes. Although statistical software packages may include options that can be used for small area estimation, the estimates of RMSE created in these statistical packages may be missing one or more components of the RMSE [3].

The potential impact of missed components in the estimated RMSE will depend to some extent on the "use case" of the SAE estimates. Annual LGA-level estimates of health risk factors from the NSWPHS would be used as one of many inputs in local evaluation and planning. Estimates based on SAE methods need to be accompanied by estimates of RMSEs to indicate the reliability of the results; if the estimates of RMSE could be obtained without resorting to computationally complex methods it is likely that SAE methods would be taken up more widely.

The parametric bootstrap provides an alternative to the computationally complex methods. An estimate of the RMSE of a small area estimator is obtained by assessing the variability between the replicates created by re-sampling and re-fitting the model to a large number of replicate samples. Resampling methods have been used in other small area estimation situations [5-7]. The non-parametric bootstrap and Jackknife estimation of RMSEs involve resampling from the actual observed values of the sample. When sample sizes are very small both of these methods do not provide good estimates of RMSE [8], so this paper focuses on using the parametric bootstrap.

An advantage of the parametric bootstrap approach is that people who do not have the mathematical background to implement or troubleshoot code that requires matrix method can usually understand and have confidence in a bootstrap method. It also gives an independent method against which to compare the estimated RMSE produced as part of the standard output by statistical packages such as SAS.

To assess the usefulness of the parametric bootstrap approach we first apply the parametric bootstrap to obtain the estimated RMSE for the corn data in Battese, Harter & Fuller (BHF) [9], comparing the results to those in the paper and the estimated RMSE obtained from the MIXED procedure in SAS. We then apply the parametric bootstrap method to obtain estimates of the RMSE for the EBLUP estimator for current smoking in males in 2006 using data from the NSWPHS and compare these to the estimates of RMSE obtained from the MIXED procedure. The response variable for the corn data is continuous for which the EBLUP, based as it is on a linear mixed model, is appropriate. The response variables from the NSWPHS are mainly categorical, in which case the EBP, based on the generalised mixed model is more appropriate. Therefore, the data used to fit the EBLUP are used to obtain EBP estimators using the GLIMMIX procedure in SAS, and the estimated RMSEs output from the GLIMMIX procedure were compared with those created using a parametric bootstrap.

The questions being asked are:

1. 1. When applying the EBLUP estimator to the BHF corn data how well do the estimated RMSEs using the parametric bootstrap compare with the estimated RMSEs

created by the software (in this case, the MIXED procedure in SAS) and the estimated RMSEs from the Battese et al [9]

2. 2. When applying the EBLUP estimator to the data from NSWPHS, how do the estimated RMSEs from a parametric bootstrap compare with the estimated RMSEs obtained from the MIXED procedure
3. 3. When fitting the more appropriate EBP to the data from the NSWPHS, how do estimated RMSEs from a parametric bootstrap compare with the RMSEs of predictions obtained using SAS GLIMMIX procedure
4. 4. Does the parametric bootstrap for estimating RMSEs for the EBP provide sufficient improvement over using the estimated RMSEs created by the GLIMMIX procedure to require the additional time required to run the parametric bootstrap when presenting results?

The aim of this work is to provide guidance for epidemiologists and other public health professionals to allow them to produce estimates of health risk factors and their associated estimated RMSEs for their LGAs even if they are not proficient at matrix algebra. In this situation, SAEs and the and level of error associated with them are used for guidance, policy development and evaluation.

Other options for calculating RMSEs are available, such as Hierarchical Bayes [10] and matrix methods [11] and may be more appropriate in certain situations.

2. Materials and Methods

For the first part of the evaluation, data in the seminal paper on the EBLUP model [9] were uploaded into SAS V9.2 [12]. The data were analyzed using the MIXED procedure in SAS based on the method used by Mukhopadhyay and McDowell [13] and compared to the estimated RMSEs in the original paper. The values of the area level variance component($\hat{\sigma}_v^2$), person level variance ($\hat{\sigma}_e^2$) and regression coefficients ($\hat{\beta}$) from the MIXED procedure were then used to estimate the RMSE using the parametric bootstrap process (see Appendix A for details).

For the second part of the evaluation, SAEs using the EBLUP were obtained for current smoking in males using unit record survey data from the 2006 NSWPHS together with area level covariates: proportions for age group, marital status, level of qualification [14] and private health cover [15]. Where possible, sex-specific data were included in the models. Unit record data for all these variables were available from the survey. The unit level covariate data used indicator variables for each level of each categorical covariate.

The estimates of $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ and $\hat{\beta}$ obtained from fitting the EBLUP were used in the parametric bootstrap process to estimate the RMSEs. The estimated RMSEs from the MIXED procedure were compared with the estimated RMSEs from the bootstrap process. See Appendix A for details of the parametric bootstrap for the EBLUP.

Finally, the third part of the evaluation involved fitting the EBP using the GLIMMIX procedure in SAS to the same data from NSWPHS as used for the EBLUP in the previous part of the evaluation. The GLIMMIX procedure provides estimated values of $\hat{\sigma}_v^2$ and $\hat{\beta}$ that were used in the parametric bootstrap process to estimate the RMSEs as outlined in Appendix B. The estimated RMSEs from the bootstrap process were compared with the estimated RMSEs produced by the GLIMMIX procedure, which are based on linearization.

3. Results

3.1. Validation using BHF data

The estimated RMSEs of predictions using the MIXED procedure output for the corn data are slightly higher than the estimated RMSEs presented in the original paper (Figure 1). In comparison, the bootstrapped estimates of RMSE are almost identical to the RMSEs presented in the original paper and slightly lower than the estimated RMSEs from the MIXED procedure.

As the estimated RMSEs using the MIXED procedure are slightly more conservative than the estimated RMSEs presented in the original paper, it is possible to estimate the RMSE of the predictions from a unit level model using a straightforward parametric bootstrap procedure, without using methods employing matrix algebra.

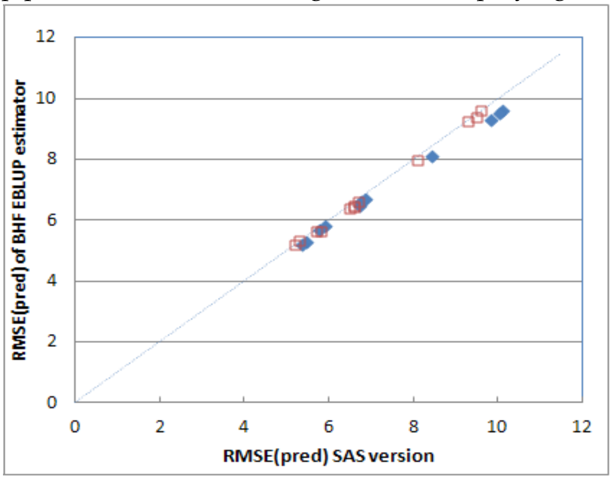

Figure 1: Estimated RMSE of predicted area of corn from parametric bootstrap (pink squares) and SAS MIXED procedure (blue diamonds) against estimated RMSE presented in [9]. 1:1 line presented for reference purposes.

Based on this comparison, use of the parametric bootstrap is shown to be an effective method of estimating RMSE for the EBLUP, and the estimated RMSE produced by the MIXED procedure is slightly lower but highly correlated.

3.2. Estimated RMSE using EBLUP

Figure 2 shows the estimated RMSEs for the EBLUPs using the parametric bootstrap and the MIXED procedure. The bootstrap estimates of RMSE (denoted RMSEbs) are, in general, slightly lower than, but in the same order of magnitude as the estimates of RMSE output from the MIXED procedure (denoted RMSEp). There is a cluster of areas where RMSEp ~ 0.032. All the areas involved have five or fewer responses. The value of 0.032 is very close to the theoretical maximum RMSE based on the g1g term, the dominant component of the Prasad-Rao estimator of MSE [16] ([3] page 137). The anomaly does not occur for RMSEbs (Figure 2), nor is it observed in a less complex model, for instance when the covariates only include age groups (Figure 3). There is also greater agreement between the bootstrap and the estimates of RMSE obtained from the MIXED procedure in the less complex model.

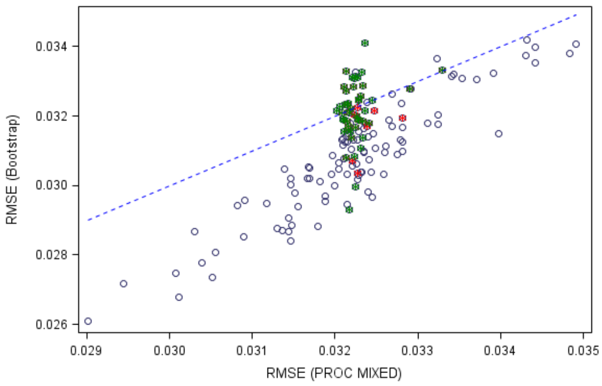

Figure 2: Bootstrap vs PROC MIXED RMSE, SMK, 2006, male, linear with age group, marital status, education level and private health status as covariates. Dotted line shows 1:1 correspondence.

Green markers denote sample size 5 or less; red markers denote areas where direct estimate is zero.

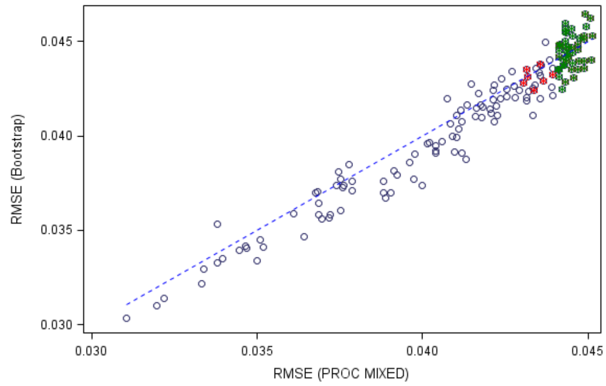

Figure 3: Bootstrap vs PROC MIXED RMSE SMK, 2006, male, linear, age group covariates only. Dotted line shows 1:1 correspondence. Green markers denote sample size 5 or less; red markers denote areas where direct estimate is zero.

3.2. Estimated RMSE using EBP

The minimum estimated RMSE from the bootstrap method is 2.7%, compared with 1.3% for the RMSE produced by the GLIMMIX procedure (Table 1). The maximum values are almost the same. Three LGAs have no responses for males in 2006. As the random effect term in the estimates for out-of-sample areas is zero, the estimates of RMSE produced by the GLIMMIX procedure for these out-of-sample areas will be those of the synthetic estimate. On the other hand, because of the variability in the random effect term factored into the parametric bootstrap, the bootstrap estimates of RMSE for these out-of-sample areas are far more appropriate. One of the three out-of-sample areas has an estimated bootstrap RMSE that is towards the maximum obtained from the parametric bootstrap process, but the other two have estimates that are only slightly higher than the mean RMSE (Figure 4). Because the in-sample areas have similar estimates between the bootstrap and the RMSE estimate produced as part of by the GLIMMIX procedure, if quick analysis is required then it may be more time-efficient to simply substitute the maximum value of the RMSE estimates produced as part of by the GLIMMIX procedure for these out-of-sample areas. This may over-estimate the RMSE for some of the out-of-sample areas, but it gives a margin of error of less than 10%, which, for an out-of-sample area is excellent.

Table 1. Summary statistics for estimated RMSE using output from the GLIMMIX procedure, parametric bootstrap

Source of estimate of RMSE	Minimum	Median	Mean	Maximum
GLIMMIX procedure	1.3%	3.5%	3.5%	4.6%
Bootstrap	2.7%	3.6%	3.6%	4.5%

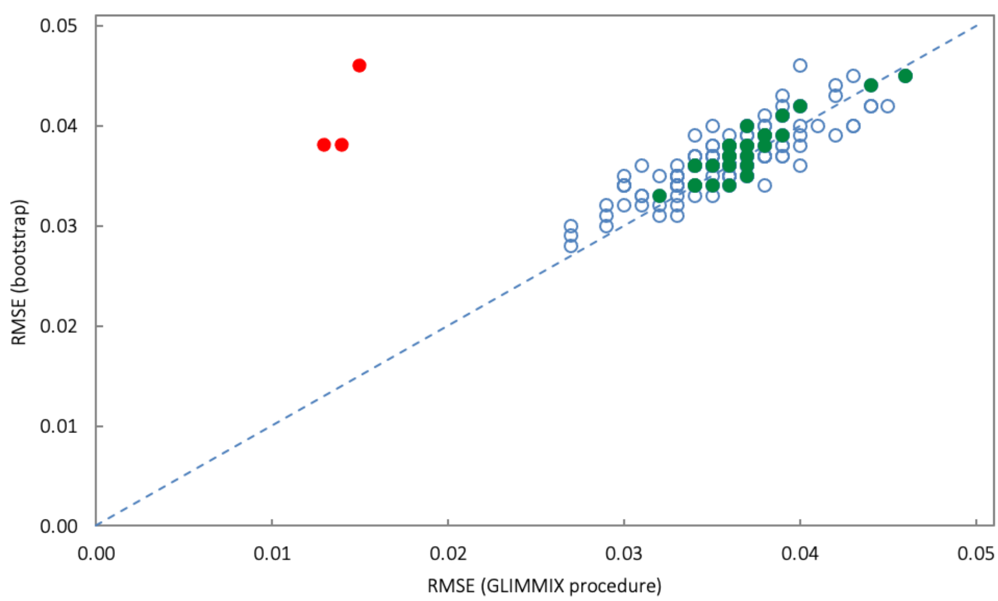

Figure 4. Bootstrap vs PROC GLIMMIX estimates of *RMSE* for EBP estimates, SMK, 2006, male. Dotted line shows 1:1 correspondence. Green markers denote sample size 5 or less; red markers denote out-of-sample areas.

4. Discussion

When compared with the estimated RMSEs for the corn data provided in the original BHF paper [9], the estimated RMSE based on the MIXED procedure created slightly lower estimates. Estimates obtained from the parametric bootstrap were almost identical to those in the paper. This suggests that the estimation method for RMSE in the MIXED procedure may indeed omit the third component of RMSE, which is of the order O(ng-1). The first two components are larger and therefore more important components. At times there may be a need for the estimate of RMSE to include the third component; at other times it may be sufficient to use an estimated RMSE that captures the most important components and is easier to calculate.

When applied to data from the NSWPHS, most estimates of RMSE were very similar between the parametric bootstrap and those obtained from the procedures themselves. For the MIXED procedure an anomaly was observed when the model was involved several covariates and the sample size for an area was less than 5. For the GLIMMIX procedure, the only time when the RMSEs created by the GLIMMIX procedure were underestimates was for out-of-sample LGAs.

There are three possible reasons for the anomalies in estimated RMSEs produced by the MIXED procedure and observed for areas with small sample size and complex models.

1. The linear model may be the issue here. The outcome variable is binary in nature, and this may provide evidence that the linear model is not appropriate when the sample size is very small.
2. It may indicate a breakdown of the SAS version of the Prasad Rao (PR) estimate of RMSE. SAS documentation [12] mentions that the behaviour of the Satterthwaite method of assessment has not been assessed fully when there are small sample sizes. The Kenward-Roger method of variance estimation in SAS uses a Satterthwaite-type method to estimate the degrees of freedom.
3. It could be that the Taylor series linearization approximation process used to estimate the PR estimator of RMSE in SAS is not particularly accurate with small sample sizes.

There is no such issue with the estimated RMSE for the EBP estimates, except for out-of-sample areas. Although for these areas, the ideal approach would involve either the parametric bootstrap or using matrix methods, another alternative is to substitute the maximum estimated RMSE from in-sample areas for areas that have no responses. This method is likely to overestimate the RMSE, making it a conservative option. From a time-saving point of view it avoids having to run the parametric bootstrap.

The results of this work were used to justify using output from the MIXED and GLIMMIX procedures to estimate RMSE in [17] and [2]. Out-of-sample areas were given the maximum RMSE for in-sample areas.

When using the parametric bootstrap, care must be taken to ensure that the model specification is correct. Any misspecification that shows up in model diagnostics applied to the original model, needs to be addressed prior to running the parametric bootstrap.

Estimates of the *precision* of the bootstrap estimate of RMSE can be obtained by running the parametric bootstrap a large number of times.

The ability to estimate the RMSE when creating small area estimates without having to run complex code and/or understand matrix algebra may lead to more opportunities for small area estimation to be routinely used in applied statistical areas. One such area would be for creating estimates of health risk factors for 150 LGAs in NSW from the NSWPHS despite this survey being developed to provide estimates of these health risk factors within the large health areas in the state. Like results from any other statistical analysis, small area estimates should be accompanied by details of the method of estimation of both the estimates themselves and the RMSE.

This paper has outlined what is quite a simplistic comparison of RMSE between the estimated RMSE from a paper with those produced by procedures in SAS and those produced with a parametric bootstrap. Definitive proof would require direct comparison of complex methods against those presented in this paper, however the level of sophistication presented was considered sufficient for the purposes to which the small area estimates were expected to be put.

If estimates at the LGA level have been requested using data from the NSWPHS, they have usually been based on 6 or 7 years of aggregated data and only provided where there are at least 200 responses, and are not publicly available. Small area estimates produced from other survey data sources may be created using synthetic estimates that avoid the need to consider complex matrix algebra, while others may present the estimates without including any measure of precision. Each of these alternatives has drawbacks: Estimates based on data aggregated over several years mean that it is not possible to create an annual time series to assess current trends, and LGAs will be unreported if they have fewer than 200 responses. The RMSE of synthetic estimates are known for under-estimating the bias, and providing estimates without any measure of quality makes it hard to know whether the results are fit-for-purpose. When compared to these alternatives, using the estimated RMSEs obtained from the GLIMIX procedure, with out-of-sample areas given the maximum RMSE for in-sample areas, or the RMSEs obtained from the parametric bootstrap provides a much better option.

6. Patents

None.

Supplementary Materials: none.

Author Contributions: Conceptualization, Diane Hindmarsh and David Steel; Data curation, Diane Hindmarsh; Formal analysis, Diane Hindmarsh; Funding acquisition, David Steel; Investigation, Diane Hindmarsh; Methodology, Diane Hindmarsh and David Steel; Supervision, David Steel; Writing – original draft, Diane Hindmarsh; Writing – review & editing, Diane Hindmarsh and David Steel. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Australian Research Council linkage grant, LP0776810ARC. The industry partners included Australian Bureau of Agriculture and Resource Economics, NSW Health, NZ Ministry of Health and Australian Bureau of Statistics.

Institutional Review Board Statement: Ethical review and approval were unnecessary for this study, due to the data being non-identifiable.

Informed Consent Statement: Written consent to publish this paper has been obtained from the data custodian for the NSW Population Health Survey program.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from NSW Population Health Survey program following submission of a formal data request and after permission was granted from NSW Health.

Acknowledgments: The authors wish to thank Professor Ray Chambers for valuable discussions regarding this work. The involvement of the linkage partners is acknowledged and sincerely appreciated as is the support of the NSW Health Chief Health Officer for providing access to the NSW Population Health Survey data through a data request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A – Steps to create parametric bootstrap estimate of MSE of EBLUP estimator

1. Fit the linear mixed model (Equation 1) to the data and save the values of $\hat{\sigma}_v^2$ $\hat{\sigma}_e^2$ and $\hat{\beta}$.

$$y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + v_g + e_{ig} \tag{1}$$

where y_{ig} is the response of the i^{th} person in the g^{th} small area,
n_g is the sample size in the g^{th} small area, $i = 1 \ldots n_g$, $g = 1 \ldots G$,
$\mathbf{x}'_{ig}$ is the vector of covariate values from the survey,
$\boldsymbol{\beta}$ is the vector of regression coefficients
v_g is a random effect reflecting area level effects, v_g are independent N(0, σ_v^2),
e_{ig} are the random errors independent N(0, σ_e^2)

For each simulation $(k = 1 \ldots K)$, continue as follows:

2. For each area (g) generate a random area term $v_g{}^*$ from the distribution $v_g{}^* \sim N\,(0, \hat{\sigma}_v^2)$
3. For each observation $i = 1 \ldots n$, generate a normal error term $e_i{}^*$ from the distribution $e_i{}^* \sim N(0, \hat{\sigma}_e^2)$
4. Use Equation 1 to simulate n values of $y_{ig,}$ $y_{ig}{}^*$ using $e_i{}^*$ and $v_g{}^*$ generated in steps 2 and 3, together with the covariates and vector of regression coefficients, and with n_g being the same as in the sample
5. Fit the linear mixed model to the simulated values and obtain the EBLUP estimates for each simulation ($\tilde{\tilde{Y}}_g{}^*$) for each area (Equation 2)

$$\tilde{\tilde{Y}}_g{}^* = \mathbf{X}'_{ig}\widehat{\boldsymbol{\beta}}^* + \hat{v}_g{}^* \tag{2}$$

where $\mathbf{X}'_{ig}$ is the vector of covariates associated with the i^{th} person in the g^{th} small area, obtained from census or other origin, and assumed known without error

6. Also, calculate the 'true' value of $\tilde{\tilde{Y}}_g$ (Equation 3) associated with the k^{th} simulation

$$\hat{\hat{Y}}_g = \mathbf{X}'_{ig}\widehat{\boldsymbol{\beta}} + v_g{}^* \tag{3}$$

7. Repeat steps 2 to 6, K times (at least 1000 is suggested)
8. Calculate the bootstrap estimate of the MSE of the EBLUP estimator for the g^{th} area using Equation 4

$$MSE_{b_g} = \frac{1}{K}\sum\left(\hat{\hat{Y}}_g - \tilde{\tilde{Y}}_g{}^*\right)^2 \tag{4}$$

Appendix B – Steps to create parametric bootstrap estimate of MSE of EBP estimator

1. Fit the non-linear mixed model (Equation 5) to the data using a logit link with the required covariates and save the values of $\hat{\sigma}_v^2$ and $\hat{\beta}$.

$$\text{logit}(y_{ig}) = \mathbf{x}'_{ig}\boldsymbol{\beta} + \nu_g \tag{5}$$

where y_{ig} is the response of the i^{th} person in the g^{th} small area, $g = 1...G$,
n_g is the sample size in the g^{th} small area, $i = 1 ... n_g$,
$\mathbf{x}'_{ig}$ is the vector of covariate values from the survey,
$\boldsymbol{\beta}$ is the vector of regression coefficients
ν_g is a random effect reflecting area level effects, ν_g are independent, $\text{N}(0, \sigma_v^2)$,

For each simulation $(k = 1 ... K)$, continue as follows:

2. For each area (g) generate a random area term $\nu_g{}^*$ from the distribution $\nu_g{}^* \sim N\,(0, \hat{\sigma}_v^2)$
3. Generate n simulated values of y_{ig}, with n_g being the same as in the sample, $y_{ig}{}^*$ This is a two-step process.
    a. First, use Equation 6 to calculate $E(p)$ substituting $\nu_g{}^*$ generated in step 2, together with the observed value of covariates and vector of estimated regression coefficients from step 1

$$E(p) = \frac{\exp(\boldsymbol{x}'_{ig}\widehat{\boldsymbol{\beta}} + \hat{\nu}_g{}^*)}{1 + \exp(\boldsymbol{x}'_{ig}\widehat{\boldsymbol{\beta}} + \hat{\nu}_g{}^*)} \tag{6}$$

4. Create a random variable between 0 and 1, for the n observations. For each observation, if the random variable is less than $E(p)$ then $y_{ig}{}^* = 1$ for the k^{th} similation, otherwise $y_{ig}{}^* = 0$.
5. Fit the same model as in step 1 to this set of simulated values, obtaining the estimated EBP for each area $\hat{P}_g^*$ (Equation 7) associated with the k^{th} simulation

$$\hat{P}_g^* = \frac{exp(\overline{\boldsymbol{X}}'_g\widehat{\boldsymbol{\beta}}^* + \hat{\nu}_g{}^*)}{1 + exp(\overline{\boldsymbol{X}}'_g\widehat{\boldsymbol{\beta}}^* + \hat{\nu}_g{}^*)} \tag{7}$$

6. For each replicate also calculate the 'true' value, given the simulated value of the random error variance term (Equation 8).

$$P_g^* = \frac{exp(\overline{\boldsymbol{X}}'_g\widehat{\boldsymbol{\beta}} + \hat{\nu}_g{}^*)}{1 + exp(\overline{\boldsymbol{X}}'_g\widehat{\boldsymbol{\beta}} + \hat{\nu}_g{}^*)} \tag{8}$$

7. Repeat steps 2 to 6, K times (at least 1000 is suggested)
8. Calculate the bootstrap estimate of the MSE of the EBP estimator for the g^{th} area using Equation 9

$$MSE_{b_g} = \frac{1}{K}\sum(\hat{P}_g^* - P_g^*)^2 \tag{9}$$

References

1. Williamson, M., Baker, D., Jorm, L. The NSW Health Survey Program Overview and Methods, 1996-2000. *New South Wales Public Health Bulletin Supplementary Series* **2001**, 12, 1-31. https://doi.org/10.1071/NB01S13
2. Hindmarsh, D.M.; Steel, D. Creating local estimates from a population health survey: practical application of small area estimation methods. *AIMS Public Health* **2020**, *7*, 403-424.
3. Rao, J. N. K., *Small Area Estimation (Methods and Applications)*, 1st ed.; John Wiley and Sons: Hoboken, NJ, USA, **2003**; p. 137.
4. Saei, A., Chambers, R. Small area estimation under linear and generalized linear mixed models with time and area effects. *S3RI Methodology Working Papers, M03/15* **2001**, available at: https://eprints.soton.ac.uk/8165/1/8165-01.pdf. Accessed 28 June 2021.

5. Hall, P., Maiti, T. On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **2006**, 68, 221–238.
6. Lahiri, S. N., Maiti, T., Katzoff, M., Parsons, V. Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, **2007**, 94, 469–485.
7. Pereira, L., Coelho, P. Assessing different uncertainty measures of EBLUP: a resampling-based approach. *Journal of Statistical Computation and Simulation*, **2010**, 80, 713–727
8. Herrador, M., Morales, D., Esteban, M., Sanchez, A., Santamaria, L., Marhuenda, Y., Perez, A. Sampling design variance estimation of small area estimators in the Spanish Labour Force Survey. *SORT-Statistics and Operations Research Transactions*, **2008**, 32, 177–197
9. Battese, G. E., Harter, R. M., Fuller, W. A. Error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **1988,** 83, 28–36
10. Gomez-Rubio, V., Best, N., Richardson, S., Li, G., Clarke, P. Bayesian Statistics for Small Area Estimation. Technical report (unpublished), Imperial College London. **2010**. Available from: http://eprints.ncrm.ac.uk/1686/1/BayesianSAE.pdf . Last accessed 28 June 2021.
11. Chambers, R., Chandra, H., Tzavidis, N. On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **2011**, 37, 153–170. Available at http://www.statcan.gc.ca/pub/12-001-x/2011002/article/11604-eng.pdf. Accessed 28 June 2021.
12. SAS Institute Inc (2009) SAS/STAT 9.2 Users Guide. SAS Institute Inc, Cary, NC, 2nd edn. On-line documentation.
13. Mukhopadhyay, P. K., McDowell, A. Small Area Estimation for survey data analysis using SAS software. *SAS Global Forum 2011 paper* 336-2011. **2011** Available at http://support.sas.com/resources/papers/proceedings11/336-2011.pdf. Accessed 28 June 2021
14. Australian Bureau of Statistics (2007) *2006 Census community profile series: Basic community profile. Cat. No. 2001.0.* Available at. https://www.abs.gov.au/websitedbs/censushome.nsf/home/communityprofiles?opendocument&navpos=230. Accessed 28 June 2021.
15. Glover, J., Tennant, S. *Social Health Atlas of New South Wales (incl. ACT) Local Government Areas, 2010.* Electronic report. Available at https://phidu.torrens.edu.au/social-health-atlases/data-archive/data-archive-social-health-atlases-of-australia#social-health-atlas-of-australia-data-released-august-november-december-2011-by-statistical-local-area-local-government-area-based-on-the-asgc-2006-and-medicare-local. Accessed 28 June 2021.
16. Prasad, N. G. N., Rao, J. N. K. On robust small area estimation using a simple random effects model. *Survey Methodology*, **1999**, 25, 67–72.
17. Hindmarsh D.M. Small area estimation for health surveys. Doctoral thesis. University of Wollongong: School of Mathematics and Applied Statistics, Wollongong, NSW, Australia, March 2013.