An Outlier Robust Block Bootstrap for Small Area Estimation

Payam Mokhtarian and Ray Chambers
National Institute for Applied Statistics Research Australia
University of Wollongong, Wollongong, NSW 2522, AUSTRALIA
Corresponding author: Payam Mokhtarian, e-mail: payam@uow.edu.au

Abstract

Small area inference based on mixed models, i.e. models that contain both fixed and random effects, are the industry standard for this field, allowing between area heterogeneity to be represented by random area effects. Use of the linear mixed model is ubiquitous in this context, with maximum likelihood, or its close relative, REML, the standard method for estimating the parameters of this model. These parameter estimates, and in particular the resulting predicted values of the random area effects, are then used to construct empirical best linear unbiased predictors (EBLUPs) of the unknown small area means. It is now well known that the EBLUP can be unstable when there are outliers in sample data, and an outlier-robust EBLUP, or REBLUP, has been proposed by Sinha and Rao (2009), based on modifying the parameter estimating functions to make them less sensitive to sample outliers. Unfortunately, these modified estimating functions can be numerically unstable, and mean squared error estimation for the REBLUP is not straightforward. Taking a somewhat different approach, Chambers and Mokhtarian (2013) proposed an outlier robust block bootstrap approach to fitting a linear mixed model in the presence of both area level and unit level outliers. A natural extension of this bounded block bootstrap can then be used to define an outlier robust version of the EBLUP and a simple way of estimating its mean squared error. This approach is described in this paper, together with simulation results that provides some evidence for our claim that the new method is robust to the influence of outliers. In particular, it leads to an easily computed version of the REBLUP and an easily computed and stable estimate of its mean squared error.

Keywords: Mixed models, robust estimation, unit-level models, variance components, random effect block bootstrap.

1. Introduction

A standard approach to small area estimation (SAE) for means or totals uses mixed models to express these targets of inference as linear combinations of both fixed and random effects. The standard method of calculating the corresponding small area estimates is to then use maximum likelihood (ML) or restricted maximum likelihood (REML) to estimate the parameters of the mixed model and to substitute these estimated parameter values in the expression for the Best Linear Unbiased Predictor (BLUP) of the small area characteristic of interest. The resulting estimator is referred to as the Empirical Best Linear Unbiased Predictor (EBLUP), see Henderson (1975) and Harville (1977). Although the EBLUP is simple to implement and is efficient under normality assumptions, it can be unreliable in the presence of sample outliers.

Chambers and Tzavidis (2006) and Tzavidis *et al.* (2010) address this issue by fitting outlier robust M-quantile models to the small area data, using a generalization of quantile regression for SAE. Sinha and Rao (2009) also tackle this issue, using the M-estimation methods set out in Richardson and Welsh (1995) to construct a plug-in outlier robust version of the EBLUP, which they call the Robust EBLUP, or REBLUP. The REBLUP has a low prediction variance but its prediction bias can be high when the sample outliers in a small area of interest are drawn from a distribution with a different mean from the rest of the values from that area. Chambers *et al*. (2013) discuss this issue and suggest that both the M-quantile estimator and the REBLUP can be improved by addition of an outlier robust bias correction. They also propose two

different analytical mean-squared error estimators for these bias-corrected outlier robust small area estimators.

Small area estimators based on a standard two level linear mixed model are dependent on EBLUPs of random area (level two) effects, which in turn depend on estimates of the variance components of the model. Chambers and Chandra (2012) describe a semiparametric two level block bootstrap method for making robust inferences about the parameters of a linear mixed model. This idea is adapted in Chambers and Mokhtarian (2013) to obtain alternative robust estimates of linear mixed model parameters and random effect predicted values using standard REML fitting methods within the replication structure of a bounded block bootstrap. In this paper we extend this idea to SAE under a linear mixed model, using the bounded block bootstrap to generate a robust alternative to the REBLUP. A key advantage of this approach is its ease of computation, both of the estimate and of an estimate of its MSE, since the bootstrap only computes REML estimates. Since analytic estimation of the mean squared error of outlier robust estimators like REBLUP is either very complex (see Chambers *et al*., 2013) or has well known computational difficulties, this alternative method of outlier robust small area inference seems promising.

The reminder of this paper is as follows. In section 2, after introducing basic concepts and notation for SAE under a linear mixed model, we briefly review recent work on outlier robust small area estimation under this model. The proposed outlier robust block bootstrap method of small area estimation is set out in section 3. In section 4 we use model-based simulation of outlier contaminated mixture data to evaluate the performance of the proposed approach, both in terms of estimation of small area means in this situation as well as estimation of mean squared error. Section 5 concludes the paper with some final remarks, and discussion of future research on outlier robust small area inference.

2. Robust Small Area Prediction

We suppose that data for a sample of n individuals from a population of size N are available. These data consist of unit record values from G areas, with n_i individuals in sample in area i. For individual j in area i, let y_j denote the value of the variable of interest, with $\mathbf{x}_j$ the value of a $p \times 1$ vector of individual level covariates and $\mathbf{z}_i$ the value of a known $q \times 1$ vector of area level covariates. The number N_i of individuals within each area i is assumed to be known, as are the corresponding small area average $\bar{\mathbf{x}}_i$ of $\mathbf{x}_j$. The target of inference is the unknown small area mean $\bar{y}_i$, and estimation based on a two level linear mixed model is proposed (Battese *et al*., 1988; Rao, 2003; Chambers and Clark, 2012). Throughout, sampling is assumed to be non-informative given the population values of $\mathbf{x}_j$ and $\mathbf{z}_i$.

Let $\mathbf{y}_U$, $\mathbf{X}_U$ and $\mathbf{Z}_U$ denote the population level vector of variable of interest and associated matrices of covariates. Then

$$\mathbf{y}_U = \mathbf{X}_U\beta + \mathbf{Z}_U\mathbf{u} + \mathbf{e}_U \tag{1}$$

where $\mathbf{u} \sim N(\mathbf{0}, \Sigma_u)$ is a vector of Gq area random effects, $\mathbf{e}_U \sim N(\mathbf{0}, \Sigma_u)$ is a vector of N individual specific random effects and $\mathbf{u}$ and $\mathbf{e}_U$ are assumed to be mutually independent. The variance-covariance matrix of $\mathbf{y}_U$ is $\mathbf{V}_U = \mathbf{\Sigma}_e + \mathbf{Z}_U\mathbf{\Sigma}_u\mathbf{Z}_U^T$. The parameters of the covariance matrices Σ_u and Σ_e, typically referred to as the variance components of (1), are denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ while the vector β is referred to as the fixed effects parameter of this model. Let $\hat{\beta}$ and $\hat{\mathbf{u}}$ denote the vectors of estimated and predicted values respectively for the fixed and area level random effects in (1). The EBLUP of the area *i* mean $\bar{y}_i$ under (1) is then

$$\hat{\bar{y}}_i^{\text{EBLUP}} = N_i^{-1}\left\{n_i\bar{y}_{si} + (N_i - n_i)\hat{\bar{y}}_{ri}\right\} \quad (2)$$

where $\hat{\bar{y}}_{ri} = \bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_i^T\hat{\mathbf{u}}$ is the predicted value of non-sample mean of the variable of interest in area i and the indices *s* and *r* denote sample and non-sample quantities, respectively. That is, $\bar{y}_{si}$ is the average of the n_i sample values of y_j in area i and $\bar{\mathbf{x}}_{ri}$ is the vector of the area i average values of the $N_i - n_i$ non-sample values of $\mathbf{x}_j$. The predictor (2) can be sensitive to sample outliers. Consequently, Sinha and Rao (2009) proposed replacing $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ by outlier robust versions in order to make (2) insensitive to these outliers. That is, they replaced (2) by the robust EBLUP (REBLUP)

$$\hat{\bar{y}}_i^{\text{REBLUP}} = N_i^{-1}\left\{n_i\bar{y}_{si} + (N_i - n_i)\hat{\bar{y}}_{ri}^{\text{REBLUP}}\right\} \quad (3)$$

where $\hat{\bar{y}}_{ri}^{\text{REBLUP}} = \bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}}^{rob} + \mathbf{z}_i^T\hat{\mathbf{u}}^{rob}$. Here $\hat{\boldsymbol{\beta}}^{rob}$ and $\hat{\mathbf{u}}^{rob}$ are outlier robust estimates of the fixed and random effects vectors in (1), obtained by solving an outlier robust version of either the maximum likelihood or the restricted maximum likelihood (REML) estimating equations for these parameters. In particular, Sinha and Rao (2009) use a modified version of the robust estimating equations for linear mixed model parameters proposed by Richardson and Welsh (1995). The robust predictors of the random area effects in (1) are then obtained by substituting these outlier robust estimates of the fixed effects and the variance components of (1) into the Fellner (1986) estimating equations for outlier robust predicted values of random effects in linear mixed models.

3. SAE via Outlier Robust Block Bootstrapping

We now describe our proposed alternative outlier robust small area estimation method, based on application of a bounded block bootstrap (hereafter RREB) under (1). The RREB was proposed by Chambers and Mokhtarian (2013), and is an outlier robust extension of the random effect block (REB) bootstrap method of Chambers and Chandra (2012). Also, for notational convenience, we restrict ourselves from now on to the special (but widely used) case of (1) where $\mathbf{z}_i \equiv 1$, so Σ_u is scalar below.

The bootstrap technique (Efron and Tibshirani, 1993) was originally developed for parametric inference given independent and identically distributed data. The block bootstrap extends this to accommodate the hierarchical dependence structure of the clustered and multilevel data that are characteristic of SAE. Although the bootstrap technique is typically used for estimating parametric estimation uncertainty given an assumed model, Chambers and Mokhtarian (2013) adapt this technique to modelling data with group structure, e.g. a linear mixed model. The REB bootstrap replications rely on level-specific empirical residuals to construct bootstrap samples. The RREB restricts the influence of the sample outliers by bounding these residuals. Like the REB, the RREB is semiparametric in the sense that although the bootstrap model is based on estimated parameters defined by fitting the model to the sample data, the dependence structure in the bootstrap model residuals is generated nonparametrically, by replicating groups and then individuals within groups rather than re-sampling the sample data at random. This ensures that the RREB, like the REB, is robust to failure of the distribution assumptions of the model (1).

The RREB procedure for estimating a set of small area means is as follows:

1. Fit a two level linear mixed model (1) to the sample data and use the marginal residuals $\mathbf{r}_s = \mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}$ to calculate the area (level 2) average residuals $\mathbf{r}^{(2)} = \mathbf{D}_G\mathbf{r}_s$, where $\mathbf{D}_G$ is *G*-block diagonal matrix whose diagonal elements are $n_i^{-1}\mathbf{1}_{n_i}^T$. Here $\mathbf{1}_m$ denotes a m-vector of ones.
2. Bound the effect of the mean zero level 2 residuals, $\mathbf{r}^{(2)c} = \mathbf{r}^{(2)} - av(\mathbf{r}^{(2)})\mathbf{1}_G$, using a suitable bounded influence function. Here $av(\mathbf{w})$ denotes the

averaging operator for the vector $\mathbf{w}$. Also, before we bound these residuals, we scale them to recover the REML estimate $\hat{\Sigma}_u$ of the variance of the area random effect. The vector of outlier robust area specific (level 2) residuals is

$$\mathbf{r}^{(2)R} = \psi\left[\left\{G^{-1}\left(\mathbf{r}^{(2)c}\right)^T \mathbf{r}^{(2)c}\right\}^{-1/2} \mathbf{r}^{(2)c}\hat{\Sigma}_u^{1/2}\right].$$

Here ψ is the Huber influence function with tuning constant is $2\hat{\Sigma}_u^{1/2}$.

3. Calculate mean zero individual effect (Level 1) residuals as $\mathbf{r}^{(1)c} = \left(\mathbf{r}_s - \mathbf{r}^{(2)R} \otimes \mathbf{1}_{n_i}\right) - av\left(\mathbf{r}_s - \mathbf{r}^{(2)R} \otimes \mathbf{1}_{n_i}\right)\mathbf{1}_n$. As with the Level 2 residuals, these are now scaled to recover the REML estimate $\hat{\Sigma}_e$ of the Level 1 variance and then bounded using the Huber ψ function with tuning constant $2\hat{\Sigma}_e^{1/2}$. The set of outlier robust Level 1 residuals is therefore

$$\mathbf{r}^{(1)R} = \left(\mathbf{r}_i^{(1)R}\right) = \psi\left[\left\{n^{-1}\left(\mathbf{r}^{(1)c}\right)^T \mathbf{r}^{(1)c}\right\}^{-1/2} \left(\mathbf{r}^{(1)c}\right)\hat{\Sigma}_e^{1/2}\right].$$

4. Let $srswr(A,m)$ denote a sample of size m taken randomly and with replacement from the set A. Level 1 and level 2 bootstrap errors are then defined by sampling independently and with replacement from each set of outlier robust residuals separately:

$$\mathbf{r}^{*(2)R} = srswr(\mathbf{r}^{(2)R}, G);$$
$$\mathbf{r}_i^{*(1)R} = srswr(\mathbf{r}^{(1)R}_{srswr(\{1,\ldots,G\},1)}, n_i);$$
$$\mathbf{r}^{*(1)R} = \left(\mathbf{r}_i^{*(1)R}\right).$$

5. Generate outlier robust bootstrap sample data $\mathbf{y}_s^{*R}$ via

$$\mathbf{y}_s^{*R} = \mathbf{X}_s\hat{\beta} + \mathbf{Z}_s\mathbf{r}^{*(2)R} + \mathbf{r}^{*(1)R} \quad (4)$$

6. Fit a two-level linear mixed model to the bootstrap values in (4) to obtain REML estimates of fixed effects parameters and variance components as well as predicted random effects.
7. Repeat steps 4 - 6 B times to obtain the RREB values of the parameter estimates. Denote these by $\left\{\hat{\beta}^{(b)\text{RREB}}, \hat{\Sigma}_u^{(b)\text{RREB}}, \hat{\Sigma}_e^{(b)\text{RREB}}; b = 1,\ldots,B\right\}$.

Let $\hat{\beta}^{\text{RREB}}$, $\hat{\Sigma}_u^{\text{RREB}}$ and $\hat{\Sigma}_e^{\text{RREB}}$ denote the bootstrap averages of $\hat{\beta}^{(b)\text{RREB}}$, $\hat{\Sigma}_u^{(b)\text{RREB}}$ and $\hat{\Sigma}_e^{(b)\text{RREB}}$ respectively. The RREB estimate of the mean of small area i is then

$$\hat{\bar{y}}_i^{\text{RREB}} = N_i^{-1}\left\{n_i\bar{y}_{si} + (N_i - n_i)\hat{\bar{y}}_{ri}^{\text{RREB}}\right\} \quad (5)$$

where $\hat{\bar{y}}_{ri}^{\text{RREB}} = \bar{\mathbf{x}}_{ri}^T\hat{\beta}^{\text{RREB}} + \hat{u}^{\text{RREB}}$. There are three versions of $\hat{\mathbf{u}}^{\text{RREB}}$ (and hence three versions of $\hat{\bar{y}}_i^{\text{RREB}}$) depending on the type of bootstrap averaging used to obtain this predicted value.

$$\hat{u}^{\text{RREB-1}} = B^{-1}\sum_{b=1}^{B}\left\{\left(n_i^{-1}\hat{\Sigma}_e^{(b)\text{RREB}} + \hat{\Sigma}_u^{(b)\text{RREB}}\right)^{-1}\hat{\Sigma}_u^{(b)\text{RREB}}\left(\bar{y}_{si} - \bar{\mathbf{x}}_{si}^T\hat{\beta}^{(b)\text{RREB}}\right)\right\}$$

$$\hat{u}^{\text{RREB-2}} = \left\{B^{-1}\sum_{b=1}^{B}\left(n_i^{-1}\hat{\Sigma}_e^{(b)\text{RREB}} + \hat{\Sigma}_u^{(b)\text{RREB}}\right)^{-1}\hat{\Sigma}_u^{(b)\text{RREB}}\right\}\left\{\bar{y}_{si} - \bar{\mathbf{x}}_{si}^T\hat{\beta}^{\text{RREB}}\right\}$$

$$\hat{u}^{\text{RREB-3}} = \left\{\left(n_i^{-1}\hat{\Sigma}_e^{\text{RREB}} + \hat{\Sigma}_u^{\text{RREB}}\right)^{-1}\hat{\Sigma}_u^{\text{RREB}}\right\}\left\{\bar{y}_{si} - \bar{\mathbf{x}}_{si}^T\hat{\beta}^{\text{RREB}}\right\}$$

Finally, we note that the MSE of the RREB estimator (9) can be easily estimated using the observed variability in the RREB bootstrap replications, via

$$\widehat{\text{MSE}}^{\text{RREB}}(\hat{\bar{y}}_i^{\text{RREB}}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\bar{y}}_i^{(b)\text{RREB}} - \hat{\bar{y}}_i^{\text{RREB}}\right)^2. \quad (6)$$

4. Model-Based Simulations

A series of model-based simulation experiments were used to evaluate the performance of the different small area estimation methods discussed in the previous section. In these simulations, we generated data using a two-level model of the form $y_{ij} = 100 + 5x_{ij} + u_i + e_{ij}$ where we fixed the total number of areas at $G = 40$. The population and sample size were the same for all areas and were fixed at $N_i = 100$ and $n_i = 5$, and sampling was at random and without replacement in each area. Values of a covariate x were generated independently and identically from a log-normal distribution with a mean of 1.0 and a standard deviation of 0.5 on the log-scale. The area specific random effects and the individual specific random effects were generated according to the four contamination-type scenarios set out in Table 1.

Table 1. Simulation scenarios.

Scenario	*Area effect distribution*	*Individual effect distribution*
[0,0]: no outlier	$u \sim N(0,3)$	$e \sim N(0,6)$
[0,e]: individual outliers only	$u \sim N(0,3)$	$e \sim \delta N(0,6) + (1-\delta)N(20,150)$ where δ is an independently generated Bernoulli random variable with $\Pr(\delta = 1) = 0.97$
[u,0]: area outliers only	areas 1-36: $u \sim N(0,3)$ areas 37-40: $u \sim N(9,20)$	$e \sim N(0,6)$
[u,e]: outliers in both area and individual effects	areas 1-36: $u \sim N(0,3)$ areas 37-40: $u \sim N(9,20)$	$e \sim \delta N(0,6) + (1-\delta)N(20,150)$ where δ is an independently generated Bernoulli random variable with $\Pr(\delta = 1) = 0.97$

Table 2. Model-based simulation results for predictors of small area means.

Predictor	*Results for different outlier scenarios and areas*					
	[0,0] 1-40	[0,e] 1-40	[u,0] 1-36	[u,0] 37-40	[u,e] 1-36	[u,e] 37-40
	Median values of RB (%)					
EBLUP	0.02	-0.20	0.10	-0.54	0.17	-1.59
REBLUP	0.03	-0.39	0.11	-0.47	-0.30	-1.00
RREB-1	0.04	-0.34	0.91	-6.71	0.63	-6.81
RREB-2	0.02	-0.17	0.08	-0.42	0.10	-0.78
RREB-3	0.04	0.32	0.85	-6.70	0.58	-6.78
	Median values of RRMSE (%)					
EBLUP	0.81	1.22	0.85	0.97	1.37	2.36
REBLUP	0.82	1.01	0.84	1.02	0.99	1.44
RREB-1	1.71	1.89	1.92	7.55	1.84	7.61
RREB-2	0.81	1.03	0.85	0.97	1.02	1.42
RREB-3	0.83	1.23	0.82	2.18	1.39	2.21

A total of 500 independent Monte Carlo simulations (population generation, then sample selection) were carried out for each simulation scenario, and within each simulation we calculated the EBLUP (2), the REBLUP (3) and all three versions of the REBB (5) using 1000 bootstrap replications, this number of simulations and bootstrap samples being suitable for evaluating 95 per cent percentile confidence intervals; see Caers *et al*. (1998). These estimators were then assessed using the median values of their area specific relative bias (RB) and relative root mean-squared error (RRMSE). These performance measures are set out in Table 2. In Table 3 we show the same performance measures, but this time for the bootstrap MSE estimator (6) for the three versions of the RREB estimator (5). Note that the MSE estimator for the EBLUP in this Table is the estimator of Prasad and Rao (1990).

Table 3. Model-based simulation results for RREB bootstrap estimators of MSE

Predictor	*Results for different outlier scenarios and areas*					
	[0,0] 1-40	[0,e] 1-40	[u,0] 1-36	[u,0] 37-40	[u,e] 1-36	[u,e] 37-40
			Median values of RB			
EBLUP	-0.34	1.74	3.82	-17.31	11.32	-40.86
RREB-1	4.08	4.05	4.94	4.87	5.91	5.37
RREB-2	-0.91	-0.89	-0.94	-0.82	-0.95	-0.88
RREB-3	-0.90	-0.90	-0.94	-0.77	-0.95	-0.84
			Median values of RRMSE			
EBLUP	6.24	18.57	7.20	17.90	22.28	43.19
RREB-1	48.57	43.38	52.80	51.33	63.64	57.62
RREB-2	16.75	19.08	16.89	12.92	25.65	22.88
RREB-3	16.27	19.12	16.94	13.03	26.06	23.01

5. Conclusions

Our simulation results show that the bootstrap averaging used in RREB-1 and in RREB-3 is inferior to that used in RREB-2. The reason for RREB-1's overall poor performance is because bootstrap averaging of predicted random effects is effectively the same as zeroing them, resulting in what is essentially a synthetic estimate. In contrast, the reason for RREB-3's poor performance is overshrinkage due to bias in robust estimates of the variance components. However, it is clear RREB-2 performs well in all the scenarios investigated in the study, with bias and mean squared error that is very similar to that of the REBLUP. Furthermore, the easily calculated bootstrap MSE estimator (6) performs well when used with RREB-2.

References

Battese, G., Harter, R. and Fuller, W. (1988) An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, **83**, 28-36.

Caers, J., Beirlant, J. and Vynckier, P. (1998) Bootstrap confidence intervals for tail indices, *Computational Statistics and Data Analysis*, **16**, 259-277.

Chambers, R. and Chandra, H. (2012) A random effect block bootstrap for clustered data, *The Journal of Computational and Graphical Statistics*, *DOI: 10.1080/10618600.2012.681216*

Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2013) Outlier robust small area estimation, *Journal of the Royal Statistical Society, Series B*, **75** (5), To appear.

Chambers, R. and Clark, R. (2012) *An Introduction to Model-based Survey Sampling with Applications*. Oxford: Oxford University Press.

Chambers, R. and Mokhtarian, P. (2013) Outlier robust block bootstrap fitting of linear mixed models, *CSSM Working Paper*, University of Wollongong.

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

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall.

Fellner, W.H. (1986) Robust estimation of variance components, *Technometrics*, **28**, 51-60.

Harville, D. (1977) The use of linear model methodology to rate high school or college football teams, *Journal of the American Statistical Association*, **72**, 278-289.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model, *Biometrics*, **31** (2), 423-447.

Prasad, N.G.N. and Rao, J.N.K. (1990) The estimation of the mean squared error of small area estimators, *Journal of the American Statistical Association*, **85**, 163-171.

Rao, J.N.K. (2003) *Small Area Estimation*. New York: Wiley.

Richardson, A.M. and Welsh, A.H. (1995) Robust restricted maximum likelihood in mixed linear models, *Biometrics*, **52**, 1429-1439.

Sinha, S.K. and Rao, J.N.K. (2009) Robust small areas estimation, *Canadian Journal of Statistics*, **37**, 381-399.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010) Robust prediction of small area means and distributions, *Australian & New Zealand Journal of Statistics*, **52**, 167-186.