Small Area Estimation for Semicontinuous Data

Hukum Chandra[1,3] and Ray Chambers[2]

[1]Indian Agricultural Statistics Research Institute, New Delhi, India

[2]University of Wollongong, Wollongong, NSW, 2522, Australia

[3]Corresponding author: Hukum Chandra, email: hchandra@iasri.res.in

Abstract

Survey data often contain variables which are semicontinuous in nature, i.e. they either take a single fixed value (typically 0) or they have a continuous, often skewed, distribution on the positive real line. This type of variables is very common in agricultural, environmental, ecological, epidemiological and economic surveys. Standard methods for small area estimation based on the use of linear mixed models can be inefficient for such variables. We discuss small area estimation techniques for semicontinuous variable under a two part random effects model which takes care of presence of excess zeros as well as skewed nature of the non-zero values of the responses variable. Empirical results suggest that the proposed method works well and produces an efficient set of small area estimates. An application to real survey data from the Australian Agricultural Grazing Industry Survey demonstrates the satisfactory performance of the method. We also propose a parametric bootstrap method to estimate the mean squared error (MSE) of the proposed estimator of small areas. The bootstrap estimates of the MSE are compared to the true MSE in simulation study.

Keywords: skewed data, zero-inflated, small area estimation, mixture model, mean squared error

1. Introduction

Many variables of interest in business and agricultural surveys are semicontinuous in nature. This article focuses on a particular type of semicontinuous variable frequently encountered in practice, a mixture of zeros and continuous skewed distributed positive values. Linear models are not appropriate for the semicontinuous variables. As a consequence, commonly used methods for small area estimation (SAE) based on the use of linear mixed models (LMMs), for example, the empirical best linear unbiased predictor (EBLUP) can be inefficient for such variables. Chandra and Chambers (2011a) and Berg and Chandra (2012) investigated SAE methods for skewed variables, focussing on those that follow a LMM following a logarithmic (log) transformation. Chandra and Chambers (2011a) described two predictors for SAE of skewed variables. The first predictor, a model-based direct estimator (MBDE) defined as a weighted sum of the sampled units, where the weights are defined to give the minimum mean squared error (MMSE) linear predictor of the population mean if the parameters of the LMM on log scale were known. The second predictor is based on prediction based approach of Karlberg (2000), that is, empirical predictor under a LMM on log scale. This empirical predictor is analogous to the synthetic estimator under a LMM. The MBDE is a direct estimator and unbiased in the presence of between area heterogeneity, but can yield unstable estimates if sample sizes are too small. On the other hand, the synthetic type empirical predictor only accounts for between area variability through the covariates and therefore can lead to biased estimates when heterogeneity exists between the areas. Berg and Chandra (2012) described an empirical best predictor in the sense that it has minimum MSE in the class of unbiased predictors. These approaches to SAE are suitable for skewed variables but their application is restricted to strictly positive variables only. Hence these approaches of SAE cannot be applied to a semicontinuous variable. Two part random effects model, also referred as a mixture model, is widely for SAE with zero-inflated variables, see for example, Pfeffermann *et al*. (2008) and Chandra and Sud (2012). We describe a SAE method for semicontinuous variables under a two part random effects model. We used a parametric bootstrap method to estimate the MSE of the proposed estimator of small areas.

2. Small Area Estimation Under Mixture Model

Let us consider a finite population U of size N which consists of D non-overlapping domains $U_i (i = 1, ..., D)$ and from this finite population a sample of size n is drawn. We assume that there is a known

number N_i of population units in small area *i*, with n_i $(i=1,....,D)$ of these sampled. The total number of units in the population is $N=\sum_{i=1}^{D}N_i$, with corresponding total sample size $n=\sum_{i=1}^{D}n_i$. We use *s* to denote the collection of units in sample, with s_i the subset drawn from small area *i*, and use expressions like $j\in i$ and $j\in s$ to refer to the units making up small area *i* and sample *s* respectively. Similarly, r_i denote the set of *j* units in small area *i* that are not in sample, with $|r_i|=N_i-n_i$ and $U_i=s_i\cup r_i$. Let y_{ij} denotes the values of variable of interest *Y* for the unit *j* in area *i* and $\mathbf{x}_{ij}$ denotes the $(m-1)$ vector of values of auxiliary variables in area *i* associated with y_{ij}. With this, the quantity of interest is the small area mean of *Y*, $m_i=N_i^{-1}\sum_{j=1}^{N_i}y_{ij}$. We consider a situation where the variable of interest follows a LMM on log transformation. Then we write a log scale LMM for the variable of interest, y_{ij}, as

$$\log(y_{ij})=l_{ij}=\mathbf{z}_{ij}^T\boldsymbol{\beta}+u_i+e_{ij} \tag{1}$$

where $\mathbf{z}_{ij}=\left(1_i,\log(\mathbf{x}_{ij})\right)$ is the $m\times 1$ vector of appropriately transformed covariates, $\boldsymbol{\beta}$ is a $m\times 1$ vector of fixed effects, u_i is the area-specific random effect associated with area *i* and e_{ij} is an individual level random errors for unit *j* in small area *i* with $(u_i,e_{ij})\sim N(\mathbf{0},diag(\sigma_u^2,\sigma_e^2))$. Let $\boldsymbol{\varphi}=(\boldsymbol{\beta},\sigma_u^2,\sigma_e^2)^T$ be the vector of model parameters, and let $\hat{\boldsymbol{\varphi}}=(\hat{\boldsymbol{\beta}},\hat{\sigma}_u^2,\hat{\sigma}_e^2)^T$ be the ML or REML estimator of $\boldsymbol{\varphi}$. In particular, $\boldsymbol{\sigma}^2=(\sigma_u^2,\sigma_e^2)^T$ is referred to as the variance components of the model and $\hat{\boldsymbol{\sigma}}^2=(\hat{\sigma}_u^2,\hat{\sigma}_e^2)^T$ denote the estimator of $\boldsymbol{\sigma}^2$. Given the sample data, we can estimate the unknown parameters (including the area effect) of model (1) and hence define the log-scale predictions as $\hat{l}_{ij}=\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}}+\hat{u}_i$, where $\hat{\boldsymbol{\beta}}$ is the estimates of $\boldsymbol{\beta}$, and $\hat{u}_i=\hat{\gamma}_i(\bar{l}_{is}-\bar{\mathbf{z}}_{is}^T\hat{\boldsymbol{\beta}})$ is the predictor of random area effect, where $\hat{\gamma}_i=\hat{\sigma}_u^2(\hat{\sigma}_u^2+n_i^{-1}\hat{\sigma}_e^2)^{-1}$ is the estimated shrinkage effect γ_i. Here, $\bar{l}_{is}=n_i^{-1}\sum_{j\in s_i}\log(y_{ij})$ and $\bar{\mathbf{z}}_{is}=n_i^{-1}\sum_{j\in s_i}\mathbf{z}_{ij}$ are the sample means of l_{ij} and $\mathbf{z}_{ij}$ respectively in area *i*. With this using a prediction-based approach similar to the Karlberg (2000), under model (1) a synthetic type predictor for the area mean m_i is given by (Chandra and Chambers, 2011a)

$$\hat{m}_i^{SYN-EP}=N_i^{-1}\left\{\sum_{s_i}y_{ij}+\sum_{r_i}\hat{y}_{ij}^{SYN-EP}\right\} \tag{2}$$

where $\hat{y}_{ij}^{SYN-EP}=\exp\left\{\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}}+0.5\left(\hat{\sigma}_u^2+\hat{\sigma}_e^2\right)-0.5\mathbf{z}_{ij}^T\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{z}_{ij}-0.25\hat{V}(\hat{\sigma}_u^2+\hat{\sigma}_e^2)\right\}$.

Chandra and Chambers (2011a) described a model-based direct estimator of the form $\sum_{j\in s_i}w_{ij}y_{ij}$, where w_{ij} is an estimator of the weight that gives the BLUP of the population mean if the parameters of the model (1) are known. To derive the predictor, Chandra and Chambers (2011a), use the approximation,

$$E(y_{ij})\approx\alpha_0+\alpha_1\hat{y}_{ij}^{SYN-EP}\text{, and} \tag{3}$$

$$Cov(y_{ij},y_{ik})\approx\hat{y}_{ij}^{SYN-EP}\hat{y}_{ik}^{SYN-EP}\left\{\exp(\hat{\sigma}_u^2)-1+\exp(\hat{\sigma}_u^2)\left(\exp(\hat{\sigma}_e^2)-1\right)I[j=k]\right\}, \tag{4}$$

where $\hat{y}_{ij}^{SYN-EP}$ is given in (2). Let us denote by $\mathbf{y}_U=(\mathbf{y}_s^T,\mathbf{y}_r^T)^T$, where $\mathbf{y}_s$ and $\mathbf{y}_r$ are the vectors of sampled and non-sampled units of *Y* respectively. Similarly, we define $\hat{\mathbf{y}}_s^{SYN-EP}$ and $\hat{\mathbf{y}}_r^{SYN-EP}$ as vectors

containing $\hat{y}_{ij}^{SYN-EP}$ for the sampled and non-sampled units and then $\mathbf{J}_U = (\mathbf{J}_s^T, \mathbf{J}_r^T)^T$ $= \left((\mathbf{1}_s^T, \mathbf{1}_r^T)^T, ((\hat{\mathbf{y}}_s^{SYN-EP})^T, (\hat{\mathbf{y}}_r^{SYN-EP})^T)^T\right)$. For known parameters, model given in (3) and (4) is a linear model for the mean of y_{ij} and then the BLUP of population mean of *Y* is given by $N^{-1}\mathbf{w}_s^T\mathbf{y}_s$, with weights

$$\mathbf{w} = (w_j; j \in s) = \mathbf{1}_s + \mathbf{H}_s^T(\mathbf{J}_U^T\mathbf{1}_U - \mathbf{J}_s^T\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}_s^T\mathbf{J}_s^T)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r, \quad (5)$$

where $\mathbf{H}_s = (\mathbf{J}_{s\,s}^T\mathbf{V}_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}_s^T\mathbf{V}_{ss}^{-1}$. Clearly, these weights are calibrated since $\sum_{j\in s} w_j = N$ and $\sum_{j\in s} w_j\hat{y}_j = \sum_{j\in U}\hat{y}_j$. The model-based direct estimator of the small area mean, m_i is

$$\hat{m}_i^{CC} = N_i^{-1}\sum_{j\in s_i} w_{ij}y_{ij}, \quad (6)$$

where w_{ij} is the element of weight $\mathbf{w}$ associated with unit (*i*,*j*) and given by (5). Under (1), following Berg and Chandra (2012), the minimum mean squared error (MMSE) predictor for m_i is given by

$$\hat{m}_i^{EBP} = N_i^{-1}\left\{\sum_{s_i} y_{ij} + \sum_{r_i}\hat{y}_{ij}^{EBP}\right\} \quad (7)$$

where $\hat{y}_{ij}^{EBP} = \exp\left\{\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\bar{l}_{is} - \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}}\right) + 0.5\hat{\sigma}_e^2(1 + n_i^{-1}\hat{\gamma}_i)\right\}$. We used Taylor linear approximation to obtain this bias correction due to back transformation. Then a bias corrected version of predictor (7) is

$$\hat{m}_i^{EBP-BC} = N_i^{-1}\left\{\sum_{s_i} y_{ij} + \sum_{r_i}\hat{y}_{ij}^{EBP-BC}\right\}, \quad (8)$$

where $\hat{y}_{ij}^{EBP-BC} = \exp\left\{\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\bar{l}_{is} - \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}}\right) + 0.5\hat{\sigma}_e^2(1 + n_i^{-1}\hat{\gamma}_i) - \frac{1}{2}\left(\mathbf{a}_{ij} + \hat{c}_{i1} + \hat{c}_{i2} + 2\hat{c}_{i3}\right)\right\}$ with,

$\mathbf{a}_{ij} = \left(\mathbf{z}_{ij}^T - \hat{\gamma}_i\bar{\mathbf{z}}_{is}^T\right)^T\hat{V}(\hat{\boldsymbol{\beta}})\left(\mathbf{z}_{ij}^T - \hat{\gamma}_i\bar{\mathbf{z}}_{is}^T\right)$. See Berg and Chandra (2012) for the details of $\hat{c}_{i1}$, $\hat{c}_{i2}$ and $\hat{c}_{i3}$.

Let us suppose that the response variable y_{ij} is a semicontinuous variable. We then describe an approach based on two part random effects model (also referred as a mixture model) to model this type of variables. Following the ideas of Olsen and Schafer (2001), Pfeffermann *et al.* (2008), Chandra and Chambers (2011b) and Chandra and Sud (2012), we express a semicontinuous variable $y_{ij} = \delta_{ij}\tilde{y}_{ij}$ as a product of two components. Here, first component $\tilde{y}_{ij}$ is referred to as log-linear component of y_{ij} and assume to follow a linear mixed model on log transformed scale, like (1). Second component $\delta_{ij} = I\left(y_{ij} > 0\right)$, is a binary (0/1) variable, assume to follow a generalized linear mixed model (GLMM) with logit link function, i.e. a logistic linear mixed model, referred as the logistic component of y_{ij}. The logarithmic component $\tilde{y}_{ij}$ is positively skewed and follows a linear mixed model on log scale similar to (1). The proposed mixture model based approach of SAE is implemented in three steps. First a linear mixed model is fitted for positive (non-zero) skewed values of the response variable on logarithmic transform scale and then in the second step a logistic linear mixed model is fitted for probability of the positive values. Finally, the two models are combined at

estimation stage. Chandra and Chambers (2011b) used a similar mixture model for SAE of the zero-inflated skewed data. However, they adopted this approach to derive the sample weights via 'fitted value' model and to define the MBDE estimator for small area means (Chandra and Chambers, 2009). They also described the MSE estimation of the proposed MBDE estimator. The proposed approach of SAE is an indirect method and hence it is expected to be efficient even for the areas for which only small samples data are available.

For the logistic component, the model linking the probability p_{ij} of positive values with the covariates is a logistic linear mixed model in area *i* of the form

$$\log it(p_{ij}) = \ln\left\{ p_{ij} / (1-p_{ij}) \right\} = \eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\theta} + v_i \ (i = 1,...,D) \tag{9}$$

with $p_{ij} = \exp(\eta_{ij})\left\{1+\exp(\eta_{ij})\right\}^{-1} = \exp(\mathbf{x}_{ij}^T\boldsymbol{\theta} + v_i)\left\{1+\exp(\mathbf{x}_{ij}^T\boldsymbol{\theta} + v_i)\right\}^{-1}$. Here $\boldsymbol{\theta}$ is a vector of unknown fixed effects parameters and v_i is the random area effect associated with area *i*, assumed to have a normal distribution with zero mean and constant variance. The estimation of unknown parameters of the logistic component was followed from the procedure described in Manteiga *et al.* (2007). In particular, we used an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\theta}$ and v_i with REML estimation of the variance component parameters. The estimation procedure was implemented in statistical software R. Using the estimated values, the predicted probabilities of the logistic component are obtained as:

$$\hat{p}_{ij} = \exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\left\{1+\exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\right\}^{-1} \tag{10}$$

In order to estimate the parameters of the second component, we denote by $s_+ = \left\{ j \in s, y_j > 0 \right\}$ the subset of the sample for which the response variable is non-zeros and have a skewed distribution, and $n_+ = \sum_{j\in s} \delta_j$ denotes the number of non-zeros sample units. Accordingly, we use a sub script of '+' to denote the quantity associated with the non-zeros sample units s_+ of size n_+. Using the sample data s_+, we the fit model (1) to obtain the estimate of fixed effect parameters and the prediction of random effects. Further we use a '+' to denote the parameter estimates based on sample s_+ of size n_+.

The predicted values of y_{ij}, that is, the linear component of y_{ij} can be obtained using expression (2) or (8). Further, we see that $E(y_{ij}) = p_{ij}\mu_{ij}$, where $\mu_{ij} = E(y_{ij} \mid \delta_{ij} = 1)$. Here we have the two parts, first for the probabilities of positive values of the response variable and second for the individual with positive value. Recall that we assume negligible correlations between the random effects in the two part of the model. In context of small area prediction problem, empirical evidence reported in Pfeffermann *et al*. (2008), Chandra and Chambers (2011b) and references therein clearly shows that this is a reasonable assumption. As a consequence, μ_{ij} and p_{ij} are assumed to be uncorrelated (or have negligible correlation). Using the estimated values of the parameters lead to plug-in predicted values of y_{ij} as $\hat{E}(y_{ij}) = \hat{p}_{ij}\hat{\mu}_{ij}$, where $\hat{p}_{ij}$ is given by (10), whereas $\hat{\mu}_{ij} = \hat{E}(\hat{y}_{ij} \mid \delta_{ij})$ is defined by (2) and (8) as $\hat{\mu}_{ij}^{MixEP}$ and $\hat{\mu}_{ij}^{MixMMSE}$ respectively. These two mixture model based indirect estimators are denoted by *Mix*EP and *Mix*EBP respectively. We used a bootstrap procedure of MSE estimation of these two estimators, i.e. *Mix*EP and *Mix*EBP. We further note that $E\left\{\hat{E}(y_{ij})\right\} = E(y_{ij})$. Linear mixed model based EBLUP is denoted by *Lin*EBLUP while the mixture model based MBDE estimator is denoted by *Mix*MBDE.

3. Results from Simulation Studies

We used model-based simulation to generate artificial population and sample data. We used two measures of the relative performance of the different SAE methods that were used in our simulations. These are the average percent relative bias (RB) and the average percent relative root mean squared error (RRMSE). In the simulation studies we choose a population size $N = 15{,}000$ and a sample size $n = 600$ and then generated randomly N_d, $d = 1,...,D; \sum_d N_d = N$ and $n_d = N_d(n/N); \sum_d n_d = n$. The average sizes of small area population and sample are 500 and 20 respectively with total of $D = 30$ small areas. These are fixed for all simulations. We first generated population values of $y_{di}\ (i = 1,...,N_d; d = 1,...,D)$ from a multiplicative model $y_{di} = \beta_o x_{di}^{\beta_1} u_d e_{di}$ with $\beta_o = 5.0$, $\beta_1 = 0.5$ and then created zero values for y_{di} randomly. The unit level random errors e_{di} are independently generated from a lognormal distribution, LN $(0, \sigma_e = 0.5)$. The random area effects u_d are generated from LN $(0, \sigma_u = 0.3)$. The covariate values x_{di} are generated from LN $(2, \sigma_x = 3)$. From this model, values of the y_{di} (that contains zeros values as well) are generated for $D = 30$ small areas of sizes N_d and then random samples of sizes n_d are drawn from each area. In our simulations, we created data with p= 0.50, 0.70 and 0.90 for all small areas at population level. Here p is proportion of positive values defined as total number of positive values in the population divided by total number of values in the population. The simulation runs were replicated K=1000 times and for each sets and in every simulation the values of small area estimates were calculated using different SAE methods described in previous Sections. Results from these simulations are reported in Table 1. For estimating the MSE using bootstrap method in each simulation run B=500 bootstrap samples were generated and the small area estimates were calculated and then the MSE estimates were computed. The related results for the MSE estimation, that is, the averages over the small areas of the true RMSEs (*Av*TRMSE) and the estimated RMSEs (*Av*ERMSE), the average percentage relative bias (*Av*RB) and the average percentage coverage rates of nominal 95 percent confidence intervals (*Av*CR) of MSE estimators of different estimators are calculated.

Two things standout from the values of percentage average relative biases (*AvRB*) reprted in Table 1, first the *Lin*EBLUP is highly biased and the biases are singnificantly greater than all the mixture model based SAE methods (i.e., *Mix*EBP, *Mix*EP and *Mix*MBDE). This clearly reveals that the *Lin*EBLUP is not suitable for semicontinuous data. Second, among the mixture model based SAE methods, the biases of *Mix*EBP are smaller than both *Mix*MBDE and *Mix*EP. With the increase in proportion of zeros, the average relative biases increases for all the methods. Turning to the values of percentage *AvRRMSE*, again with the larger area specific proportion of zeros the percentage *AvRRMSE* of all the methods increases. Again we see that the *Lin*EBLUP has very large values of *AvRRMSE*s as compared to the mixture model based methods. Among the mixture model based methods, the *Mix*EBP dominates the other methods. Overall, the proposed *Mix*EBP has both smaller bias and high efficiency. The proposed predictor under a two part random effect model (*Mix*EBP) offers substantial bias and efficiency gains over the other predictors when the study variable is semicontinuous. In this case, the *Lin*EBLUP is not recommneded to be used in practcie since it is very

ustable with huge biases. In costrast, the propsoed mixture model based SAE method particualrly the *Mix*EBP has proven to be a good method for such data. The actual coverage rates of nominal 95 percent confidence intervals achieved by the bootstrap MSE estimators of *Mix*EBP and *Mix*EP are 95 percent. This good performance is also confirmed by the results that the area averages of the true RMSEs and the estimated RMSEs obtained using bootstrap MSE estimator are very close. Overall, the bootstrap MSE estimator for the *Mix*EBP approximates the true MSE very well and with a good coverage property.

Reference

Berg, E., and Chandra, H. (2012) "Small Area Prediction for a Unit Level Lognormal Model". *Proceedings of the 2012 Federal Committee on Statistical Methodology Research Conference*, Washington, DC, USA, January 10-12, 2012.

Chandra, H. and Sud, U.C. (2012) "Small Area Estimation for Zero-Inflated Data," *Communications in Statistics - Simulation and Computation*, 41 (5), 632–643.

Chandra, H. and Chambers, R. (2011a) "Small Area Estimation Under Transformation To Linearity," *Survey Methodology*, 37 (1), pp. 39-51.

Chandra, H. and Chambers, R. (2011b) "Small Area Estimation for Skewed Data in Presence of Zeros," *The Bulletin of Calcutta Statistical Association*, 63, pp. 249-252.

Chandra, H. and Chambers, R. (2009) "Multipurpose Weighting for Small Area Estimation," *Journal of Official Statistics*, 25 (3), 379-395.

Karlberg, F. (2000) "Population Total Prediction Under a Lognormal Superpopulation Model," *Metron*, LVIII, 53-80.

Manteiga, G.W., Lombardìa, M.J., Molina, I., Morales, D., and Santamarìa, L. (2008) "Bootstrap Mean Squared Error of a Small-Area EBLUP," *Journal of Statistical Computation and Simulation*, 78(5), 443-462.

Olsen, M.K. and Schafer, J. L. (2001) "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association,* 96 (454), 730-745.

Pfeffermann, D., Terryn, B. and Moura, F.A.S. (2008) "Small Area Estimation under a Two-part Random Effects Model with Application to Estimation of Literacy in Developing Countries," *Survey Methodology,* 34 (2), 235-249.

Table 1. Percentage average relative bias (*AvRB*) and percentage average relative RMSE (*AvRRMSE*) of different estimators in model based simulations.

p	*Mix*EBP	*Mix*EP	*Mix*MBDE	*Lin*EBLUP	*Mix*EBP	*Mix*EP	*Mix*MBDE	*Lin*EBLUP
	AvRB				*AvRRMSE*			
0.9	0.50	1.11	0.68	13.06	15.07	31.03	18.98	77.88
0.7	0.58	1.07	0.99	12.42	19.39	32.86	26.96	77.83
0.5	0.75	1.22	1.12	13.95	24.65	35.61	36.83	96.90