STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 503—516

SMALL AREA ESTIMATION UNDER A MIXTURE MODEL

Hukum Chandra, HVL Bathla and U C Sud[1]

ABSTRACT

Small area estimation (SAE) under a linear mixed model may not be efficient if data contain substantial proportion of zeros than would be expected under standard model assumptions (hereafter zero-inflated data). We discuss the SAE for zero-inflated data under a mixture model (Fletcher *et al*., 2005 and Karlberg, 2000) that account for excess zeros in the data. Our results from simulation studies show that mixture model based approach for SAE works well and produces an efficient set of small area estimates. An application to real survey data from the National Sample Survey Organisation of India demonstrates the satisfactory performance of the approach.

Key Words*:* Linear mixed model, Small area estimation, EBLUP, Zero-inflated data, mixture model.

1. Introduction

In recent years, demand for reliable small areas statistics has greatly increased worldwide due to their growing use in formulating policies and programs, allocation of government funds, regional planning, and marketing decisions at local level. Sample surveys are usually planned to produce estimates for larger domains or areas and are therefore not appropriate to produce small area statistics due to small sample sizes. Due to cost and operational considerations, it is seldom possible to procure a large enough overall sample size to support direct estimates of adequate precision for all areas of interest. It is often necessary to employ indirect estimates for small areas that can increase the effective area sample size by borrowing strength from related areas through linking models, using census and administrative data and other auxiliary data associated with the small areas. The linear mixed models have been widely used in SAE. The empirical best linear unbiased predictor (EBLUP) is the most popular approach for estimation under these models, see Rao (2003). However, the EBLUP is model dependent and

[1] Division of Sample Survey, Indian Agricultural Statistics Research Institute, PUSA Campus, New Delhi-110012, India. Email: hchandra@iasri.res.in.

sensitive to model failure. To guard against model failure, the model-assisted approaches are proposed in the literature. For example, the Pseudo-EBLUP described in Prasad and Rao (1999) and the model-assisted empirical best predictor of Jiang and Lahiri (2006), hereafter JL-EBP. If linear mixed model is true, neither will be as efficient as the EBLUP. For the robustness, both estimators rely on the design consistency. Relying on a large sample property of a small sample statistic seems rather optimistic.

In practice, survey data often contain large proportion of zero values (for example, agricultural and environmental surveys etc.) than would be expected under standard model assumptions. Presence of excess zeros in the data makes the model assumptions invalid, see McCullagh and Nelder (1989). Consequently, problems with inference are liable to occur by ignoring this feature of the data. In classical regression literatures, mixture models which separately model the non-zero values and the occurrence of zero values are widely used to account for excess zeros in data, see, for example, Lambert (1992), Welsh *et al.* (1996) and Fletcher *et al.* (2005). In survey estimation Karlberg (2000) applied mixture model to estimate the population total for highly skewed data with many zeros. In the context of SAE, Chandra and Chambers (2006) and Chandra *et al*. (2007) observed that the EBLUP is ill-suited for small areas with large proportions of zeros. This indicates that standard methods of SAE under a linear mixed model may not be efficient for such data. In this article we explore the small area estimation under the mixture model for zero-inflated data. Following Fletcher *et al.* (2005) and Karlberg (2000), our approach for SAE works in three steps. First, a linear mixed model is fitted for positive values and then, in the second stage, a generalized linear mixed model (GLMM) is fitted for probability of positive values. Finally, two models are combined in estimation.

The structure of the paper is as follows. In the next section we first illustrate a linear mixed model and related estimators for small areas and then we introduce the mixture model and small area estimator under this model. In section 3 we present results from model-based as well as design-based simulation to evaluate the proposed approach of SAE, with the latter based on real survey data from Debt-Investment Survey 2002–03 of the National Sample Survey Organisation (NSSO) for the rural areas of the state of Uttar Pradesh in India. Finally, section 4 is devoted to concluding remarks and further research topics.

2. Estimation of Small Area Means

2.1. Small area estimation under a linear mixed model

Let U denote a population of size N and assume that population is partitioned into D small areas (or areas) $U_i (i = 1, ..., D)$. Let N_i and n_i is population and sample size respectively for area i. The total number of units in the population is

$N = \sum_{i=1}^{D} N_i$, with corresponding total sample size $n = \sum_{i=1}^{D} n_i$. Let $\mathbf{y}_i$ denote the N_i-vector of population values of a characteristic Y of interest and $\mathbf{x}_i$ denote the corresponding $N_i \times p$ matrix of population values in area i. Throughout, we use i to index the D small areas of interest, and j to index the distinct population units in these areas. We use s to denote the collection of units in a sample, with s_i the subset drawn from area i. Our aim is estimation of population mean of Y in small area i, i.e. $m_i = N_i^{-1} \sum_{j \in U_i} y_j$.

A commonly used class of models in small area inference is the class of linear mixed models. We consider the following linear mixed model for the distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$ in area i:

$$\mathbf{y}_i = \mathbf{x}_i \boldsymbol{\beta} + \mathbf{g}_i \mathbf{u}_i + \mathbf{e}_i \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{g}_i$ is a $N_i \times q$ matrix of known covariates characterising differences between small areas, $\mathbf{u}_i$ is a q-vector of random area effects associated with area i and $\mathbf{e}_i$ is a $N_i \times 1$ vector of individual level random errors. The area specific effects $\{\mathbf{u}_i; i = 1, ..., D\}$ are assumed to be independent and identically distributed realisations of a random vector of dimension q with zero mean and covariance matrix Σ_u. Similarly, the scalar individual effects making up $\mathbf{e}_i$ are assumed to be independent and identically distributed realisations of a random variable with zero mean and variance σ_e^2, with area and individual effects mutually independent. The covariance matrix of $\mathbf{y}_i$ is $\mathbf{v}_i = \sigma_e^2 \mathbf{I}_{N_i} + \mathbf{g}_i \Sigma_u \mathbf{g}_i'$, depends on a vector of parameters $\theta = \left(\Sigma_u, \sigma_e^2\right)$ typically referred to as the variance components of (1). We assume that the sampling method used is uninformative given the values of the auxiliary variables, so the sample data also follow the population model (1).

By aggregating the area-specific models (1) over the D-small area, we are led to the population level linear mixed model

$$\mathbf{y}_U = \mathbf{x}_U \beta + \mathbf{g}_U \mathbf{u} + \mathbf{e}_U \tag{2}$$

where $\mathbf{y}_U = (\mathbf{y}_1', ..., \mathbf{y}_D')'$, $\mathbf{x}_U = (\mathbf{x}_1', ..., \mathbf{x}_D')'$, $\mathbf{g}_U = diag\{\mathbf{g}_i; 1 \le i \le D\}$, $\mathbf{u}' = \left(\mathbf{u}_1', ..., \mathbf{u}_D'\right)$ and $\mathbf{e}_U = (\mathbf{e}_1', ..., \mathbf{e}_D')'$. Under (2), the covariance matrix of $\mathbf{y}_U$ is $\mathbf{v}_U = diag(\mathbf{v}_i; 1 \le i \le D)$. Given a sample s of size n from this population, we

can partition $\mathbf{v}_U = \begin{bmatrix} \mathbf{v}_{ss} & \mathbf{v}_{sr} \\ \mathbf{v}_{rs} & \mathbf{v}_{rr} \end{bmatrix}$ into their sample and non-sample components. Here, $r = U - s$ denotes the population units that are not in the sample. In particular, under (2) we have $\mathbf{v}_{ss} = diag\left\{\mathbf{v}_{iss}; i = 1,\ldots,D\right\} = diag\left\{\mathbf{g}_{is}\Sigma_u\,\mathbf{g}'_{is} + \sigma_e^2\mathbf{I}_{is}; i = 1,\ldots,D\right\}$ and $\mathbf{v}_{sr} = diag\left\{\mathbf{v}_{isr}; i = 1,\ldots,D\right\} = diag\left\{\mathbf{g}_{is}\Sigma_u\,\mathbf{g}'_{ir}; i = 1,\ldots,D\right\}$. Here $\mathbf{g}_{is}$ and $\mathbf{g}_{ir}$ denote the restriction of $\mathbf{g}_i$ to sampled and non-sampled units in area i respectively. Given estimated values $\hat{\theta} = \left(\hat{\Sigma}_u, \hat{\sigma}_e^2\right)$ of the variance components we can substitute these to obtain estimates $\hat{\mathbf{v}}_{ss}$ and $\hat{\mathbf{v}}_{sr}$ of $\mathbf{v}_{ss}$ and $\mathbf{v}_{sr}$ respectively.

Under (2), the EBLUP for small area i mean of Y is (Rao, 2003, section 6.2.3)

$$\begin{aligned} \hat{m}_i^{EBLUP} &= \hat{E}\left\{m_i \,|\, \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\right\} \\ &= N_i^{-1}\left[\sum_{j\in s_i} y_j + \mathbf{1}'_{ir}\left\{\mathbf{x}_{ir}\hat{\beta} + \hat{\mathbf{v}}_{irs}\hat{\mathbf{v}}_{iss}^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right] \\ &= N_i^{-1}\left[n_i\bar{y}_{is} + (N_i - n_i)\left\{\bar{\mathbf{x}}'_{ir}\hat{\beta} + \bar{\mathbf{g}}'_{ir}\hat{\Sigma}_u\mathbf{g}'_{is}\left(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'_{is} + \hat{\sigma}_e^2\mathbf{I}_{is}\right)^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right]. \end{aligned} \tag{3}$$

Here $\hat{E}$ denotes the expectation operator under (2) with unknown parameters replaced by estimates, $\mathbf{x}_{is}$ and $\mathbf{x}_{ir}$ are the matrices of sample and non-sample values of $\mathbf{X}$ in area i, $\mathbf{y}_{is}$ is the vector of sample values of Y in the same area, $\hat{\beta}$ is the 'empirical' BLUE of β, $\hat{\mathbf{v}}_{irs}$ is the transpose of the estimated value of $\mathbf{v}_{isr}$ with $\hat{\mathbf{v}}_{iss}$ the corresponding estimate of $\mathbf{v}_{iss}$, and $\mathbf{1}_{ir}$ is a vector of ones of length $N_i - n_i$. Note that the estimator (3) is model dependent and works well under (1).

Two alternative approaches in the literature under linear mixed model (1) are the pseudo-EBLUP (Rao, 2003, section 7.2.7) and the estimator of Jiang and Lahiri (2006). Recollect from (3) that the EBLUP is defined by replacing the unknown area i mean m_i by an estimate of its expected value given the observed sample values of Y in area i and the area i values of $\mathbf{X}$. Let π_{ij} denote the sample inclusion probability of population unit j in small area i. The pseudo-EBLUP is then defined by replacing m_i by an estimate of its expected value given the value of its design-consistent estimate

$$\hat{m}_i^{\pi} = \left(\sum_{j\in s_i}\pi_{ij}^{-1}\right)^{-1}\sum_{j\in s_i}\pi_{ij}^{-1}y_{ij} = \sum_{j\in s_i}\tilde{w}_{ij}y_{ij} \tag{4}$$

and the area *i* values of **X**. That is, under (1) the pseudo-EBLUP of m_i is (Rao, 2003, section 7.2.7)

$$\begin{aligned}\hat{m}_i^{psuedoEBLUP} &= \hat{E}\left\{m_i \middle| \hat{m}_i^{\pi}, \mathbf{x}_{is}, \mathbf{x}_{ir}\right\} \\ &= \overline{\mathbf{x}}_i'\hat{\beta}_{\tilde{w}} + \left(\overline{\mathbf{g}}_i'\hat{\Sigma}_{u\tilde{w}}\overline{\mathbf{g}}_{i\tilde{w}}\right)\left(\overline{\mathbf{g}}_{i\tilde{w}}'\hat{\Sigma}_{u\tilde{w}}\overline{\mathbf{g}}_{i\tilde{w}} + \hat{\sigma}_{e\tilde{w}}^2\sum\nolimits_{j\in s_i}\tilde{w}_{ij}^2\right)^{-1}(\hat{m}_i^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}_{\tilde{w}})\end{aligned} \tag{5}$$

where $\hat{\beta}_{\tilde{w}}$, $\hat{\Sigma}_{u\tilde{w}}$ and $\hat{\sigma}_{e\tilde{w}}^2$ are pseudo-maximum likelihood estimates based on the weights $\tilde{w}_{ij}$ and $\overline{\mathbf{g}}_{i\tilde{w}}$ and $\overline{\mathbf{x}}_{i\tilde{w}}'$ are design-consistent estimates of $\overline{\mathbf{g}}_i$ and $\overline{\mathbf{x}}_i$ that are defined in exactly the same way as $\hat{m}_i^{\pi}$ above. Under the same model the Jiang and Lahiri (2006) approach leads to an estimator that is also defined by conditioning on the value of $\hat{m}_i^{\pi}$,

$$\begin{aligned}\hat{m}_i^{JL} &= \sum\nolimits_{j\in s_i}\tilde{w}_{ij}\hat{E}\left\{\hat{E}\left(y_{ij}\middle|\mathbf{x}_{ij},\mathbf{u}_i\right)\middle|\hat{m}_i^{\pi},\mathbf{x}_i\right\} \\ &= \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta} + \left\{\tilde{\mathbf{w}}_{is}'\left(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}_{is}' + \hat{\sigma}_e^2\mathbf{I}_{is}\right)\tilde{\mathbf{w}}_{is}\right\}^{-1}\left\{\tilde{\mathbf{w}}_{is}'\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'\tilde{\mathbf{w}}_{is}\right\}\left(\hat{m}_i^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}\right)\end{aligned} \tag{6}$$

where $\tilde{\mathbf{w}}_{is}$ is the vector of sample weights $\tilde{w}_{ij}$ in area *i*. Note that in (6) we use optimal (i.e. maximum likelihood (ML) or restricted maximum likelihood (REML)) estimates for model parameters. Both (5) and (6) are essentially motivated by the idea of estimating the area *i* mean by its conditional expectation under (1) given the value of the usual design-consistent estimator (4) for this quantity. Under (1), neither will be as efficient as the EBLUP.

2.2.Small area estimation under a mixture model

Let us consider that survey variable *Y* is zero-inflated. We then introduce a mixture model to accommodate excess zeros in the data. In the dual regime model described in section 1, following the ideas of Fletcher *et al*. (2005) and Karlberg (2000), the variable of interest *Y* is expressed as product:

$$\mathbf{y}_i = \mathbf{y}_i^*\boldsymbol{\delta}_i \tag{7}$$

where $\mathbf{y}_i^*$ is the linear component and assume to follow a linear mixed model, like (1) and $\boldsymbol{\delta}_i = I\left(\mathbf{y}_i > 0\right)$, is a binary (0/1) variable, assume to follow a generalized linear mixed model (GLMM) with logit link function (Breslow and Clayton, 1993), i.e. logistic linear mixed model, referred as the logistic component of (7). We assume that δ_j given $\mathbf{x}_j$ are independent Bernoulli random variables with probability $p_j = \mathrm{P}(y_j > 0) = \mathrm{P}(\delta_j = 1)$. With these

notations, the model linking the probability of positive values with the covariates is the logistic linear mixed model in small area *i* of the form (Manteiga *et al.*, 2007):

$$\log it(\mathbf{p}_i) = \ln\left\{\mathbf{p}_i /(1-\mathbf{p}_i)\right\} = \boldsymbol{\eta}_i = \mathbf{x}_i\boldsymbol{\alpha} + \mathbf{g}_i\mathbf{b}_i \ (i = 1,...,D) \qquad (8)$$

with $\mathbf{p}_i = \exp(\boldsymbol{\eta}_i)\left\{1+\exp(\boldsymbol{\eta}_i)\right\}^{-1} = \exp(\mathbf{x}_i\boldsymbol{\alpha} + \mathbf{g}_i\mathbf{b}_i)\left\{1+\exp(\mathbf{x}_i\boldsymbol{\alpha} + \mathbf{g}_i\mathbf{b}_i)\right\}^{-1}$. Here $\boldsymbol{\alpha}$ is a vector of unknown fixed effects parameters and $\mathbf{b}_i$ $(i = 1,...,D)$ is the random area effect associated with area *i* which is assumed to be normal with zero mean and constant variance.

For estimation of parameters for linear-component, we denote by $s_+ = \left\{j \in s, y_j > 0\right\}$ the subset of the sample for which the survey variable is non-zeros, and $n_+ = \sum_{j\in s}\delta_j$ denotes the number of non-zeros sample units. As below (2), we denote by $\mathbf{y}^*_{s_+}, \mathbf{x}_{s_+}$, $\mathbf{g}_{s_+}$ and $\mathbf{V}_{s_+s_+}$ the corresponding vector and matrices related to non-zeros survey variable values of the sample. Accordingly, at area level we use similar notation by introducing an extra subscript *i*. Assuming model (1), the empirical-BLUE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{x}_{is_+}\right)^{-1}\left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{Y}^*_{is_+}\right)$ with $E(\hat{\boldsymbol{\beta}} \mid \boldsymbol{\delta}_j) = \boldsymbol{\beta}$ and $V(\hat{\boldsymbol{\beta}} \mid \boldsymbol{\delta}_j) = \left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{x}_{is_+}\right)^{-1}$. The expectation and variance are under the model. Under (1), the predicted values for the linear component of (7) are (Henderson, 1953):

$$\hat{E}(y^*_j) = \hat{\mu}_j = \mathbf{x}_j\hat{\boldsymbol{\beta}} + \mathbf{g}_j\hat{u}_i; j \in i \text{, with } \hat{\mathbf{u}}_i = \hat{\Sigma}_u\mathbf{g}'_{is_+}\hat{\mathbf{v}}^{-1}_{is_+s_+}(\mathbf{y}_{is_+} - \mathbf{x}'_{is_+}\hat{\boldsymbol{\beta}}). \qquad (9)$$

For the estimation of unknown parameters of logistic components in (8), we used an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\alpha}$ and $\mathbf{b}_i$ with REML estimation of variance component parameters as described in Saei and Chambers (2003) and Manteiga *et al.* (2007). The predicted probabilities of the logistic component of (7) are:

$$\hat{p}_j = \exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}} + \mathbf{g}_j\hat{\mathbf{b}}_i)\left\{1+\exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}} + \mathbf{g}_j\hat{\mathbf{b}}_i)\right\}^{-1}; j \in i. \qquad (10)$$

Using results from appendix and collecting (10) and (11), an approximately model-unbiased estimate of $E(y_j)$ is

$$\hat{E}(y_j \mid \mathbf{x}_j, \mathbf{g}_j, i) = \hat{\theta}_j = \left\{\mathbf{x}_j\hat{\boldsymbol{\beta}} + \mathbf{g}_j\hat{\Sigma}_u\mathbf{g}'_{is_+}\hat{\mathbf{v}}^{-1}_{is_+s_+}(\mathbf{y}_{is_+} - \mathbf{x}'_{is_+}\hat{\boldsymbol{\beta}})\right\}.$$

$$\left\{ e^{\mathbf{x}_j\hat{\boldsymbol{\alpha}}+\mathbf{g}_j\hat{\mathbf{b}}_i} \left(1+e^{\mathbf{x}_j\hat{\boldsymbol{\alpha}}+\mathbf{g}_j\hat{\mathbf{b}}_i}\right)^{-1} \right\}. \quad (11)$$

Consequently using (11) the estimator for population mean of Y in area i is

$$\hat{m}_i^{mix} = \hat{E}\left\{m_i \left| \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir} \right.\right\} = N_i^{-1}\left[\sum_{j\in s_i} y_j + \sum_{j\in r_i} \hat{y}_j\right] = N_i^{-1}\left[\mathbf{1}_{is}'\mathbf{y}_{is} + \mathbf{1}_{ir}'\hat{\boldsymbol{\theta}}_{ir}\right] \quad (12)$$

Besides (10), we also consider the estimated probabilities under a logistic linear model, with no area effect in (8):

$$\hat{p}_j = \exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}})\left\{1+\exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}})\right\}^{-1}; j \in i \quad (13)$$

3. Empirical Evaluations

In this section we present results from two simulation studies that were used to contrast the performance of different small area estimators set out in Table 1 and described in section 2. The first is a model based simulation in which small area population and sample data were simulation under the model. The second is a design based simulation in which a fixed population containing number of small areas was repeatedly sampled, holding the sample size in each small area fixed.

The performance of different small area estimators were evaluated with respect to two basic criteria–the relative bias and the relative root mean squared error, both expressed as percentages, of area mean estimates. The bias was measured as

$$\%AvRB = \underset{i}{mean}\left|\left\{M_i^{-1}\left(K^{-1}\sum_{k=1}^{K}\hat{m}_{ik}\right)-1\right\}\right|\times 100$$

Note that the subscript of k here indexes the K simulations, with m_{ik} denoting the value of the small area i mean in simulation k and $\hat{m}_{ik}$ denoting the area i estimated value in simulation k. The actual area i mean value (the average over the simulations) is denoted by $M_i = K^{-1}\sum_{k=1}^{K} m_{ik}$. The root mean squared error was measured as

$$\%AvRRMSE = \underset{i}{mean}\left[M_i^{-1}\left\{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\hat{m}_{ik}-m_{ik}\right)^2}\right\}\right]\times 100\,.$$

3.1. Model-Based Simulations

In these simulations we fixed population size $N = 15,000$ and number of small areas $D = 30$. Population sizes in the small areas were uniformly distributed over the interval [443, 542] and were kept fixed over simulations. In each simulation we generated N population values of auxiliary variable x from the chi-square distribution with 10 degrees of freedom. For each area, population values for y were generated under the two-level model $y_{ij} = 10 + 2x_{ij} + u_i + e_{ij}$ $\left(j = 1,...,N_i; i = 1,..,30\right)$. The area-specific random effects u_i and individual random effects e_{ij} were independently drawn from $N(0, \sigma_u^2 = 4)$ and $N(0, \sigma_e^2 = 16)$ distributions respectively. Zero-inflated population values for y were generated by first generating Bernoulli (0/1) random variable with fixed probability p (i.e. proportion of non-zero values in the data) and multiplying them to y for each area. A sample of size $n = 600$ was then selected from the simulated population, with area sample sizes proportional to the fixed area populations. Sampling was via stratified random sampling, with the strata defined by the small areas with average sample size of 25. All population values independently regenerated at each simulation and an independent sample drawn from the population each time. A total of $K = 1000$ simulations were carried out. We used simulations scenario with different proportion of non-zero values p = 0.90, 0.75, 0.65 and 0.50 for all small areas.

The average relative bias and the average relative root mean squared error over these simulations for different estimators are presented in Table 2. These results show an increase in biases and RMSEs for all estimators with increase in proportion of zeros in the data. We noted relatively unstable performance of the direct estimator (DIR), so it is not an option for SAE. Hereafter we do not discuss DIR. Among the linear mixed model based estimators, in terms of biases and RMSEs, the EBLUP is better overall. Both the EBLUP and Pseudo have smaller biases and RMSE compared to JL. However, EBLUP is either dominating marginally or at par with Pseudo. In contrast, among mixture model based estimators, MIX1 and MIX2 are almost identical. The area-specific relative bias and relative RMSE for p=0.65 are shown in Figure 1. The results in Figure 1 reflect the consistently better performance of the proposed approach. Similar conclusions are also true for other values of p. Overall, mixture model based approach is superior than the standard linear mixed model based estimators for zero-inflated data (see Table 1 and Figure 1).

3.2. Design-Based Simulations

In these simulations we used the data of Debt-Investment Survey 2002–03 for rural areas of the state of Uttar Pradesh in India conducted by National Sample Survey Organisation in the year 2002–03. In the original survey there were 11,814 sample households (with population of 22,145,951) spread across 69 districts of Uttar Pradesh that participated in the Survey. However, in our simulation studies (to avoid computation difficulties arising out of computer memory space) we used the sample of 1693 households from $D = 10$ districts only and further divided the survey weights by 10 to reduce the overall population size. This sample of 1693 households was bootstrapped to create a realistic population of $N = 327,481$ households by re-sampling with replacement with probability proportional to a household's sample weight. A total of $K = 1000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample. Districts were the small area of interest. The Y variable of interest was amount of loan outstanding per household (with 43 percent zero values in the original sample) and the auxiliary variable X was land owned by household. The aim was to predict the district level average value of amount of loan outstanding per household.

Table 3 set out the districts-wide as well as average relative biases and relative RMSEs of different estimators based on the 1000 repeated independent stratified samples. We note that the districts have different proportions of non-zeros in the data which gives a realistic scenario to apply our approach. These results show that the mixture model based MIX2 estimator for small area mean is severely biased and relatively unstable. The GLM (13) used to estimate probability for occurrence of zero in MIX2 is not fitting well since it does not capture the area effects. In contrast, average bias and RMSE of the MIX1 is consistently smaller than linear mixed model based EBLUP, Pseudo and JL estimators for small mean. Further, among linear mixed model based estimator, in terms of average bias and RMSE, the EBLUP performs better than JL which dominates the Pseudo. Furthermore, district-wide results reveal that in most of the districts MIX1 have smaller biases and RMSEs than alternative estimators. Overall the results in Table 3 clearly indicate that the mixture model based estimation is working well and better than the linear mixed model estimation if data contain substantial proportion of zeros.

4. Conclusions

The results set out in section 3 conclude that commonly used methods of SAE under a linear mixed model lead to biased and unstable estimates for small area with many zeros. In this case, proposed mixture model based SAE takes care of

excess zeros and provides more efficient sets of small area estimates. An application to real survey data from the NSSO too shows satisfactory performance of the proposed approach. In this article we have not addressed the mean squared error (MSE) estimation for different estimator for small area means described in sections 2. The MSE of various estimators under the linear mixed model (EBLUP, Pseudo-EBLUP and JL) can be estimated via their analytical MSE expression already available in the literatures, see, for example, Rao (2003) and Jiang and Lahiri (2006). However, the MSE of the proposed estimators under mixture model still need to be developed. Alternatively, re-sampling based methods like Jacknife or bootstrap methods can be explored. Authors are currently working on the MSE estimation.

Appendix

From (7) we see that

$$\begin{aligned} E(y_j) &= \Pr(\delta_j = 1)E(y_j \mid \delta_j = 1) + \Pr(\delta_j = 0)E(y_j \mid \delta_j = 0) \\ &= \Pr(\delta_j = 1)E(y_j \mid \delta_j = 1) = p_j \mu_j \end{aligned}$$

where $\mu_j = E(y_j \mid \delta_j = 1)$. This leads to $\hat{E}(y_j) = \hat{p}_j \hat{\mu}_j = \hat{\theta}_j$. Assuming that $\hat{\mu}_j$ and $\hat{p}_j$ are uncorrelated (see Karlberg [7]) leads to

$$\begin{aligned} E(\hat{\theta}_j) = E(\hat{p}_j \hat{\mu}_j) &= E\left\{E(\hat{p}_j \hat{\mu}_j \mid \delta_j)\right\} \\ &= E\left\{\hat{p}_j E(\hat{\mu}_j \mid \delta_j)\right\} = E(\hat{p}_j) E\left\{E(\hat{\mu}_j \mid \delta_j)\right\} \\ &= p_j \mu_j = E(y_j). \end{aligned}$$

Table 1. Estimators evaluated in simulation studies

Estimators	Description
DIR	Direct estimation
Under linear mixed model (1)	
EBLUP	EBLUP (3)
Pseudo	Pseudo-EBLUP (5)
JL	JL-EBP (6)
Under mixture model (7)	
MIX1	Estimator (12) with estimated probabilities (10)
MIX2	Estimator (12) with estimated probabilities (13)

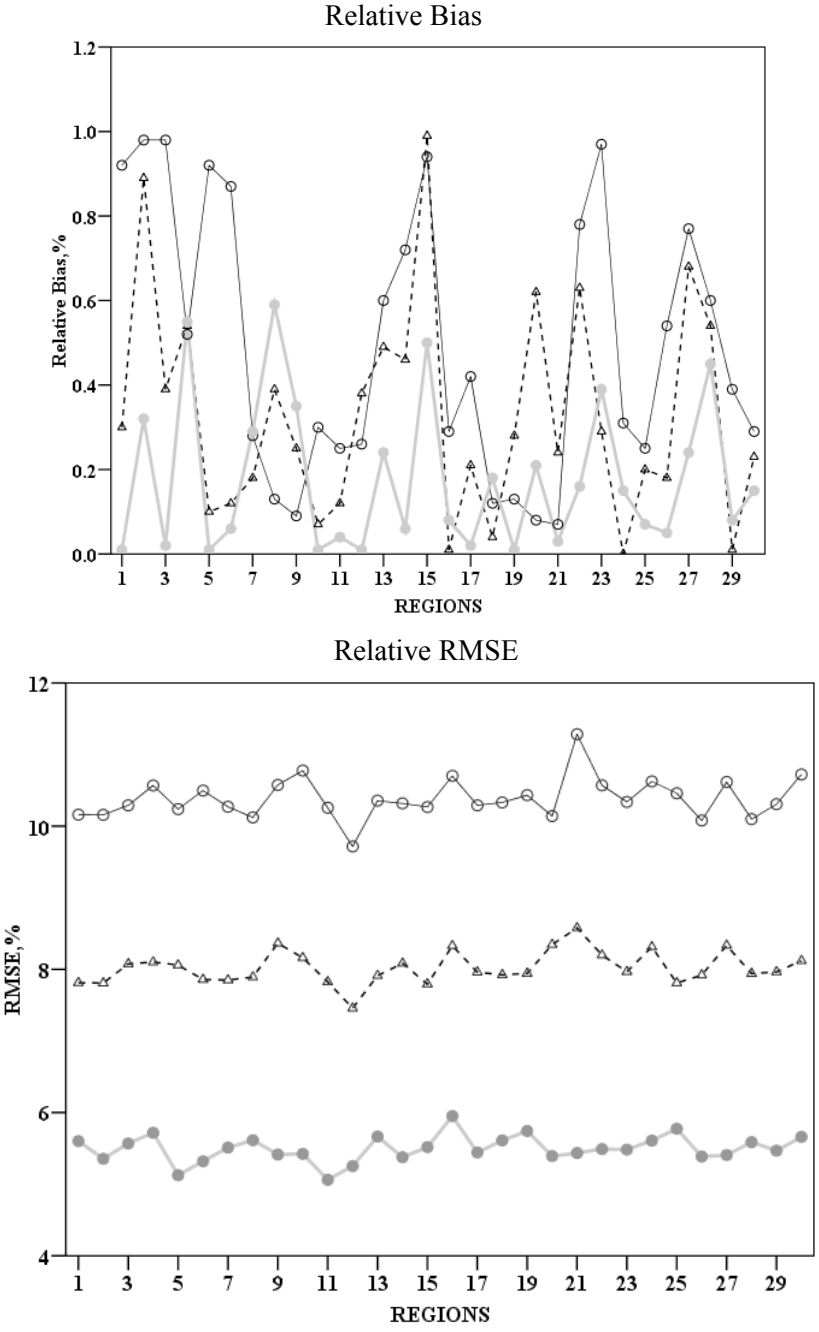
Table 2. Percentage average relative bias (*%AvRB*) and percentage average relative RMSE (*%AvRRMSE*) of different estimators under model based simulations

	p			
Estimators	0.90	0.75	0.65	0.50
	%AvRB			
DIR	0.48	0.54	0.61	0.90
EBLUP	0.24	0.28	0.33	0.37
Pseudo	0.25	0.30	0.33	0.39
JL	0.41	0.44	0.49	0.51
MIX1	0.18	0.16	0.21	0.30
MIX2	0.17	0.17	0.18	0.29
	%AvRRMSE			
DIR	10.74	15.14	18.00	24.40
EBLUP	5.78	7.26	8.02	9.72
Pseudo	5.87	7.32	8.06	9.68
JL	8.67	9.77	10.39	11.77
MIX1	3.91	5.19	6.03	8.04
MIX2	3.62	4.73	5.50	7.25

Table 3. Districts-wide (and average) percentage relative bias and percentage relative RMSE of different estimators under design based simulations

Districts	*p*	EBLUP	Pseudo	JL	MIX1	MIX2
		% Relative bias				
1	0.59	4.7	6.6	4.9	4.7	21.6
2	0.62	4.0	5.0	4.0	4.0	24.3
3	0.42	5.8	6.4	5.5	0.3	10.9
4	0.46	9.1	7.4	9.0	2.7	5.6
5	0.50	3.3	3.2	3.2	5.0	10.5
6	0.31	30.0	33.4	30.0	19.4	67.1
7	0.44	5.7	8.0	6.0	15.1	21.9
8	0.52	19.1	17.0	19.2	18.0	24.5
9	0.27	2.4	5.8	2.5	7.0	47.8
10	0.49	0.7	1.4	0.4	5.4	2.7
Average	0.46	8.5	9.4	8.5	8.2	23.7
		% Relative RMSE				
1	0.59	16.5	17.9	17.2	16.0	24.5
2	0.62	16.6	17.0	17.3	15.6	26.4
3	0.42	15.6	17.3	15.9	14.5	18.7
4	0.46	15.7	17.1	15.8	14.3	14.6
5	0.50	21.0	22.2	22.4	19.8	20.1
6	0.31	42.1	46.2	43.2	31.8	75.6
7	0.44	15.0	17.4	15.1	20.8	25.4
8	0.52	29.3	28.9	29.1	28.7	30.6
9	0.27	24.3	26.9	24.3	26.9	60.8
10	0.49	12.0	13.7	11.7	12.0	11.2
Average	0.46	20.8	22.4	21.2	20.0	30.8

Figure 1. Region-specific performance measures of the EBLUP (dashed line, Δ), JL (thin line, O) and MIX2 (solid line,●) for p=0.65 under model based simulations.



REFERENCES

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistics Association*, **88**, 9–25.

CHANDRA, H. and CHAMBERS, R. (2006). Multipurpose Weighting for Small Area Estimation. *Journal of Official Statistics*, accepted for publication.

CHANDRA, H., SALVATI, N. and CHAMBERS, R. (2007) Small Area Estimation for Spatially Correlated Populations. A Comparison of Direct and Indirect Model-Based Methods. *Statistics in Transition,* **8**, pp. 887–906.

FLETCHER, D., MACKENZIE, D. and VILLOUTA, E. (2005). Modelling Skewed Data With Many Zeros: A Simple Approach Combining Ordinary and Logistic Regression. *Journal of Environmental and Ecological Statistics*, **12** (1), 45–54.

HENDERSON, C.R (1953). Estimation of Variance and Covariance Components. *Biometrics*, **9**, 226–252.

JIANG, J. and LAHIRI, P. (2006). Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach. *Journal of the American Statistical Association*, **101**, 301–311.

KARLBERG, F. (2000). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes. *Journal of Official Statistics*, **16**, 229–241.

LAMBERT, D. (1992). Zero-Inflated Poisson Regression, With An Application To Defects in Manufacturing. *Technometrics*, 34, 1–14.

MCCULLAGH, P. and NELDER, J.A. (1989) *Generalized Linear Models* (New York: Chapman and Hall)

MANTEIGA, G.W., LOMBARDÌA, M.J., MOLINA, I., MORALES, D., and SANTAMARÌA, L. (2007). Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model. *Computational Statistics & Data Analysis*, **51**(5): 2720–2733.

PRASAD, N.G.N. and RAO, J.N.K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, **25**, 67–72.

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

SAEI, A. and CHAMBERS, R. (2003). Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper No. M03/15. University of Southampton, UK.

WELSH, A.H., CUNNINGHAM, R.B., DONNELLY, C.F., and LINDENMAYER, D.B. (1996). Modelling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros. *Ecological Modelling*, 88, 297–308.