2008

Estimation of small domain means for zero contaminated skewed data

Hukum Chandra
University of Wollongong, hchandra@uow.edu.au

R. Chambers
University of Wollongong, ray@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

07-08

Estimation of small domain means for zero contaminated skewed data

Hukum Chandra and Ray Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Estimation of small domain means for zero contaminated skewed data

Hukum Chandra[a*] and Ray Chambers [b]

[a] *Division Social Statistics, University of Southampton, U.K.,*
Email: hchandra@soton.ac.uk
[b] *Centre for Statistical and Survey Methodology, University of Wollongong, Australia*
Email: ray@uow.edu.au

Abstract

For skewed data linear model assumptions are questionable. Consequently, the standard techniques for small domain estimation based on linear mixed model can be inefficient. The estimation methods for small domains for skewed data that are linear following a log-log transformation are investigated by Chandra and Chambers (2006). However, application of their methods is limited to strictly positive survey variables. In many surveys (e.g. business and enterprises, income and expenditure, agricultural and ecological surveys etc) variables that are skewed often take zero values. In this paper we introduce small domain estimation techniques for skewed data in presence of zeros. In this context, following Fletcher *et at*. (2005) and Karlberg (2000) we extend Chandra and Chambers (2006) approach of small domain estimation under a mixture model. Our empirical results show the method works well and produces an efficient set of small area estimates.

Key Words: Skewed data; Small domain estimation; Mixture model; Expected-value model

1. Introduction

Reliable estimates for small domains are often required for regional planning, fund allocation and formulating policies. A domain is regarded as small if the domain specific sample information is not large enough (i.e. domain specific sample size is too small) to produce usual design unbiased direct estimates of adequate precision. This sensitivity of

* Corresponding author. Division Social Statistics, School of Social Sciences, University of Southampton, Highfield, Southampton –SO171BJ, U.K , Telephone: 0044-23 8059 4083 Fax: 0044-23 8059 3846
Emails:hchandra@soton.ac.uk

sample sizes has led the theory of small area estimation (SAE). The term small domain and small area are interchangeably used in the literature. The linear mixed models provide a better and efficient approach to SAE by incorporating random area effects that account for between area or domain dissimilarities beyond that is explained by covariates included in the model. See Rao (2003). In many surveys (e.g. business and enterprises, income and expenditure, agricultural and ecological surveys etc) data are skewed, and linear model provides a poor fit. Commonly used methods for SAE based on linear mixed models lead to inefficient estimates. Chandra and Chambers (2006) proposed SAE for skewed data that are linear following a log-log transformation. See Chandra (2006). In this case, they extended the model-based direct (MBD) approach of SAE discussed in Chandra and Chambers (2005). In particular, they derived sample weights via model calibration (Wu and Sitter, 2001) based on log-log transform model with random area-specific effects and then defined the MBD estimators for small area quantities. However, application of this method of SAE is limited to the strictly positive survey variables. In practice, skewed data often contains many zeros. In this situation, Chandra and Chambers (2006) method cannot be implemented. A naïve approach would be to add a constant (usually 1) to the survey variables and then apply their framework with adjusted variables. An obvious disadvantage, choice of constant is arbitrary and may influence the results. Among several methods proposed in the literature to model such data, mixture model is commonly employed. See Fletcher *et at*.(2005), Welsh *et al*.(1996) and Lambert (1992).

In this paper we discuss the small area estimation methods for skewed data in presence of zeros under the mixture model. Following Fletcher *et at*.(2005) and Karlberg (2000), our approach works in three stages. First a log-log linear mixed model is fitted

for positive values and then in the second stage a logistics model is fitted for probability of positive values. Finally, two models are combined in estimation. We then adopt Chandra and Chambers (2006) approach to derive the sample weights via 'expected value' model (also called the 'fitted value' model) and to define the MBD estimators for small area means. The next section, illustrates the mixture model, defines 'expected value' model and the related estimators for small domain means and their mean squared error estimators. In section 3 we present some empirical results. Finally, section 4 is devoted to concluding remarks and further research topics.

2. Estimation under mixture model

In this section we first define mixture model underpinning the skewed data with zeros and we then derive an 'expected value' model. To start, let Y_d be the $N_d \times 1$ vector of values of the variable of interest *y* and X_d be the $N_d \times (m-1)$ matrix of values of the auxiliary variables in small area *d*, N_d is the number of population units in the small area *d* (*d* =1,2…,*D*) and *D* is the total number of small areas. We assume that the survey variable *y* is positively skewed with both zero and non-zero values. Our aim is then to estimate the population mean for *y* in small area *d*, i.e. $\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_i$. We used subscript *d* for restriction to area. For estimation of population level quantities for the skewed survey variable with zeros, Karlberg (2000) advocated the used of combination of log-log normal and logistics model, assuming that population units are independent. This independence assumption is not true when our interest is estimation of small areas. Following Karlberg (2000), we express the survey variable $y_i = a_i \tilde{y}_i$ as a product of two components, where $\tilde{y}_i$ is referred as a log-log normal or logarithmic component and a_i

as a logistic component. We define a_i as a Bernoulli random variable for occurrence of a positive value of *y*. That is $a_i = 1$ if $y_i > 0$ and $a_i = 0$ otherwise. The variable $\tilde{y}_i$ is assumed to be linear on log-log scale.

For the logarithmic component, we follow the log-log linear mixed model defined in Chandra and Chambers (2006). That is $\tilde{L}_d = \log(\tilde{Y}_d)$ and Z_d follows a linear mixed model in the small area *d* of the form

$$\log(\tilde{Y}_d) = Z_d\beta + G_d u_d + e_d \tag{1}$$

where $Z_d = \left(1, \log(X_d)\right)$ is the $N_d \times m$ matrix of covariates, β is a $m \times 1$ vector of fixed effects, G_d is a $N_d \times q$ matrix of known covariates, u_d is the area-specific random effect associated with area *d* and e_d is a $N_d \times 1$ vector of individual level random errors. The two random effects u_d and e_d are assumed to be independently distributed, with zero means and variances $Var(u_d) = \Sigma(\theta)$ and $Var(e_d) = \sigma_e^2 I_{N_d}$ respectively, and they are assumed to be normal. Here $Var(\tilde{L}_d) = \sigma_e^2 I_{N_d} + G_d \Sigma(\theta) G_d' = V_d$. Note that the covariance matrix V_d depends on a vector θ of fixed parameters, usually referred as the variance components of the model. Throughout this article we assumed that sampling is uninformative given the values of the auxiliary variables, so the sample data also follow the population model and expectation and variance are under the model.

Under the assumption of spatial independence between small areas, aggregating *D*-area level models (1) over the population, we are led to the population level model

$$\tilde{L} = Z\beta + Gu + e \tag{2}$$

where $\tilde{L} = (\tilde{L}_1', .., \tilde{L}_D')'$, $Z = (Z_1', .., Z_D')'$, $G = diag(G_d; 1 \le d \le D)$, $u = (u_1', ..., u_D')'$ and $e = (e_1', .., e_D')'$. The covariance matrix of $\tilde{L}$ is $V = diag(V_d; 1 \le d \le D)$. Similar to

Chandra and Chambers (2006) we consider the decomposition of $\tilde{L}$, Z, G and V into sample and non-sample components so that Z_s is the $n \times m$ matrix of sample values of the auxiliary variables, G_s is the corresponding $n \times q$ matrix of sample components of G and V_{ss} is the $n \times n$ covariance matrix associated with the n sample units that make up the $n \times 1$ sample vector $\tilde{L}_s$. A subscript of r is used to denote corresponding quantities defined by the $N-n$ non-sample units, with V_{rs} denoting the $(N-n) \times n$ matrix defined by $Cov(\tilde{L}_r, \tilde{L}_s)$. In what follows we denote 1_N, 1_n and 1_r as vectors of 1's and I_N, I_n and I_r as identity matrices of order N, n and $N-n$ respectively. We use similar notation at the small area level by introducing an extra subscript d to denote small area. For example, we denote by s_d the set of n_d sample units in area d, r_d the corresponding $N_d - n_d$ non-sampled units in the area and put $V_{dss} = \sigma_e^2 I_{n_d} + G_{ds}\Sigma(\theta)G'_{ds}$ and $V_{dsr} = G_{ds}\Sigma(\theta)G'_{dr}$. In practice the variance components that define the covariance matrix V are unknown and we estimate them from the sample data under the model (2) with suitable estimation methods such as maximum likelihood (ML), restricted maximum likelihood (REML) or methods of moment. See for example Harville (1977). Then the estimated covariance matrix of $\tilde{L}$ is $\hat{V} = diag(\hat{V}_d; 1 \le d \le D)$ with $\hat{V}_d = \hat{\sigma}_e^2 I_{N_d} + G_d \Sigma(\hat{\theta}) G'_d$.

For the logistic component, we assume that a_i given Z_i are independent Bernoulli random variable with $\pi_i = P(y_i > 0 \mid Z_i) = P(a_i = 1 \mid Z_i)$. That is a_i takes the values 1 and 0 with distinct population values of a_i's are independently distributed. There can be three possible options for estimating probability π_i for the positive values. These are (i) fitting a generalised linear mixed model (GLMM) to the logit of the probability π_i (ii)

fitting a generalised linear model (GLM) to the logit of the probability π_i and (iii) simple area specific proportions of number of positive values to the total sample size. Fletcher *et at*. (2005) and Karlberg (2000) used the generalised linear model and fit the logit of the probability π_i to estimate probabilities. In the context of SAE, GLMM is widely used to model such probabilities. Although we are not presenting the results here in this paper, empirical investigations show the performance of proposed estimators for SAE do not have much differences due to these three methods/choices of estimating the probabilities. In other words, the estimates of probabilities for the positive values by the area specific proportions produce the equivalent result to that based on GLMM or GLM based methods. Consequently, we motivate to use the area specific proportions to estimate these probabilities, which are straightforward and easy to work. However, authors do have empirical evidence supporting this statement and reader can get it on request. The theoretical descriptions of the GLMM and GLM based estimation methods for SAE are not given in this paper since these are not pursued furthermore. See for example Rao (2003) for further details.

2.1 Estimation of parameters

For the estimation of logarithmic component of the survey variable under a log-log linear mixed model (2), we denote by $s_p = \{i \in s, y_i > 0\}$ the subset of the sample with respect to the positive values of the survey variable, and $n_p = |s_p| = \sum_{i \in s} a_i$ denotes the number of positive sample units. Let us denote by L_{ps}, Z_{ps}, G_{ps} and V_{pss} the corresponding vector and matrices related to strictly positive survey variable values. At small area level we use similar notation by introducing an extra subscript *d*. With these

notation, and assuming model (2) holds, the empirical best linear unbiased estimator of β is

$$\hat{\beta} = \left(\sum_{d=1}^{D} Z'_{pds} \hat{V}_{pdss}^{-1} Z_{pds}\right)^{-1} \left(\sum_{d=1}^{D} Z'_{pds} \hat{V}_{pdss}^{-1} L_{pds}\right) \text{ with } E(\hat{\beta} \mid y > 0) = \beta \text{, and}$$

$$Var(\hat{\beta} \mid y > 0) = \left(\sum_{d=1}^{D} Z'_{pds} \hat{V}_{pdss}^{-1} Z_{pds}\right)^{-1}.$$

For unit $i \in d$, we can see that

$$E(\hat{\tilde{y}}_i \mid a_i = 1) = E\left\{\exp(\hat{\tilde{l}}_i) \mid a_i = 1\right\} = \exp(Z_i\beta + \frac{b_{ii}}{2})$$

$$\neq \exp(Z_i\beta + \frac{v_{ii}}{2}) = E\left\{\exp(\tilde{l}_i) \mid a_i = 1\right\} = E(\tilde{y}_i \mid a_i = 1)$$

where $v_{ii} = \sigma_e^2 + G_i \Sigma(\theta) G_i'$ and $b_{ii} = Z_i \left(\sum_{d=1}^{D} Z'_{pds} \hat{V}_{pdss}^{-1} Z_{pds}\right)^{-1} Z_i'$. This indicates that back-transformation leads to biased predictor. The second order bias corrected predictors are

$$\hat{\tilde{y}}_i = k_i^{-1} \exp(Z_i\hat{\beta} + \frac{\hat{v}_{ii}}{2}); i \in d \tag{3}$$

where $k_i = \exp\left[\left(b_{ii} + Var(\hat{v}_{ii})/4\right)/2\right]$ is the bias correction with

$b_{ij} = Z_i \left(\sum_{d=1}^{D} Z'_{pds} \hat{V}_{pdss}^{-1} Z_{pds}\right)^{-1} Z_j' \to 0$ as $n \to \infty$ and $Var(\hat{v}_{ii})$ is the asymptotic covariance matrix of $\hat{v}_{ii} = \hat{\sigma}_e^2 + G_i \Sigma(\hat{\theta}) G_i'$, see Chandra and Chambers (2006).

For logistic component, as we mentioned earlier, estimated probabilities are given by

$$\hat{\pi}_i = n_{dp} / n_d; i \in d \tag{4}$$

where n_{dp} is number of positive values and n_d is sample size in area *d*.

We now grouped (3) and (4) at the area level in vector from as follows

$\hat{\tilde{Y}}_d = (\hat{\tilde{Y}}_{d1},.....,\hat{\tilde{Y}}_{dN_i})' = k_d^{-1}\, \mathrm{e}^{Z_d\hat{\beta}+\frac{\hat{v}_d}{2}}$ so that $E(\hat{\tilde{Y}}_d - \tilde{Y}_d \mid a_d = 1) \approx 0$ with $k_d = (k_{d1},....,k_{dN_i})'$, $\hat{v}_d = (\hat{v}_{d11},....,\hat{v}_{dN_iN_i})'$, and $\hat{\pi}_d = (\hat{\pi}_{d1},..,\hat{\pi}_{dN_i})'$ so that $E(\hat{\pi}_d) - E(a_d) = E(\hat{\pi}_d) - \pi_d \approx O(n^{-1})$. That is $\hat{\pi}_d$ is a biased predictor of π_d, however, we assume that $E(\hat{\pi}_d) - \pi_d \approx 0$ as $n \to \infty$. Further,

$$E(\hat{\tilde{Y}}_d - \tilde{Y}_d) = E\{E(\hat{\tilde{Y}}_d - \tilde{Y}_d \mid a_d = 1) \mid a_d = 1\} \approx 0$$

$$\Rightarrow E(\hat{\tilde{Y}}_d \mid a_d = 1) \approx E(\tilde{Y}_d \mid a_d = 1) = \mathrm{e}^{Z_d\beta+\frac{v_d}{2}}.$$

In order to obtain an unbiased predictor for the survey variable Y_d, we assume that $\hat{\tilde{Y}}_d$ and $\hat{\pi}_d$ are uncorrelated. Although they are not exactly, but in reality it is approximately true (Karlberg, 2000). This leads to

$$\begin{aligned} E(\hat{\pi}_d\hat{\tilde{Y}}_d) &= E\left\{E(\hat{\pi}_d\hat{\tilde{Y}}_d \mid a_d = 1)\right\} = E\left\{\hat{\pi}_d E(\hat{\tilde{Y}}_d \mid a_d = 1)\right\} \\ &= E(\hat{\pi}_d)E\left\{E(\hat{\tilde{Y}}_d \mid a_d = 1)\right\} \approx \pi_d\, \mathrm{e}^{Z_d\beta+\frac{v_d}{2}} = E(Y_d) \end{aligned} \quad (5)$$

This indicates $\hat{E}(y_i) = \hat{\pi}_i\hat{E}(\tilde{y}_i \mid a_i = 1) = \hat{\pi}_i\left\{k_i^{-1}\,\mathrm{e}^{Z_i\hat{\beta}+\frac{\hat{v}_{ii}}{2}}\right\} = \hat{\pi}_i\hat{\tilde{y}}_i; i \in d$. An approximately model-unbiased predictor of survey variable Y_d is

$$\hat{Y}_d = \hat{\pi}_d\left\{k_d^{-1}e^{Z_d\hat{\beta}+\frac{\hat{v}_d}{2}}\right\} = h(Z_d,\hat{\eta}) \quad (6)$$

Note that expression (6) is a second order bias corrected first moment. To get second moment, using the properties of lognormal and Bernoulli distribution (Casella and Berger, 1990), covariance between y_i and y_j (which is product of lognormal and Bernoulli variable) is evaluated as below. For units in small area *d*, under normality of the random errors, we have

$$\omega_{ij} = Cov(y_i, y_j) = \begin{cases} e^{(Z_i+Z_j)\beta} e^{\frac{1}{2}(v_{ii}+v_{jj})} (\pi_{ij} e^{v_{ij}} - \pi_i \pi_j) & for\ i \neq j \\ e^{2Z_i\beta} e^{v_{ii}} (\pi_i e^{v_{ii}} - \pi_i^2) & for\ i = j \end{cases}. \quad (7)$$

Here $E(y_i y_j) = P(a_i = 1, a_j = 1) E\left(E(y_i y_j) \mid a_i = 1, a_j = 1\right) = \pi_{ij} e^{(Z_i+Z_j)\beta} \mathrm{e}^{(v_{ii}+v_{jj}+2v_{ij})/2}$ and

$E(y_i) = \pi_i \mathrm{e}^{Z_i\beta + \frac{v_{ii}}{2}}$.

From (7) covariance matrix of Y_d is written as

$$V(Y_d) = \Omega_d = A_d \, \Delta_d \, A_d' \quad (8)$$

where $A_d = \left\{ diag(e^{Z_{di}\beta}); 1 \leq i \leq N_d \right\}$ and $\Delta_d = [\delta_{dij}]$ is $N_d \times N_d$ positive definite matrix with $\delta_{dij} = \left\{ \pi_{ij} \mathrm{e}^{v_{ij}} - \pi_i \pi_j \right\} \mathrm{e}^{(v_{ii}+v_{jj})/2}$. The area-specific approximately bias corrected predictor (6) and covariance matrix (8), grouped at population level define the population level version of 'expected value' or 'fitted value' model as a general model with first and second moment as

$$E(Y \mid h) = \alpha_0 1_N + \alpha_1 h(Z;\eta) = \alpha J \text{ and } Var(Y \mid h) = \Omega \quad (9)$$

where $Y = (Y_1', ..., Y_D')'$, $\Omega = diag(\Omega_d; 1 \leq d \leq D)$, $\alpha = (\alpha_0, \alpha_1)'$ is a vector of unknown parameters and J denotes the 'design matrix' for the linear model (9) linking Y and $h(Z;\eta)$. Under model (9), we use Wu and Sitter (2001) model calibration approach to derive the sample weights. The key idea of this approach is provided model (9) is a reasonable one, Y is then (at least approximately) a linear function of its 'fitted value' $h(Z;\eta)$. Under this model we carry out linear estimation using these 'fitted values' as auxiliary variables and we then derive the sample weights to define small area estimators, see Chandra and Chambers (2006).

2.2 Small area means estimators

With an appropriate sample and non-sample partition of Y, J and Ω, as below equation (2), the EBLUP type sample weights under the model (9) are

$$w_{EBLUP}^{h} = 1_n + \hat{H}_h'(J'1_N - J_s'1_n) + (I_n - \hat{H}_h'J_s')\hat{\Omega}_{ss}^{-1}\hat{\Omega}_{sr}1_r \tag{10}$$

where $\hat{H}_h = (J_s'\hat{\Omega}_{ss}^{-1}J_s)^{-1}J_s'\hat{\Omega}_{ss}^{-1}$. See Royall (1976). Following Chandra and Chambers (2005) we now define estimator for small area means. There can be two types of model-based direct (MBD) estimators for small area means, Hájek type and Horvitz-Thompson (HT) type. Chandra and Chambers (2006) considered both the Hájek type and Horvitz-Thompson type of the MBD estimators for small area means defined by using the model calibration sample weights derive under fitted-value model for the skewed data. They suggested that for the model calibration sample weights an appropriate (and efficient) MBD estimator is one defined as the Horvitz-Thompson (HT) type. The sample weights (10) associated with the sample units in the small area d can be used to define the HT type of MBD estimators for the small area d mean, $\bar{Y}_d$ as

$$\hat{\bar{Y}}_d^{HT,MBD} = \sum\nolimits_{s_d} w_i y_i \Big/ N_d \tag{11}$$

The estimator (11) also depends on how the model sample weights (10) are specified. That is whether the 'fitted value' model (9) has the ratio or the regression specification. For $\alpha_0 = 0$ in model (9) we refer as ratio specification of this model, otherwise regression specification. However, Chandra and Chambers (2006) concluded that the estimator (11) have equivalent performance for both ratio and regression specification of model (9). Consequently, we used only the ratio specification for the model (9). Then we considered the HT type of MBD (11) under ratio specification of the fitted values model (9) and denote this estimator as TrMBD.

Besides the MBD estimator (TrMBD) defined by (11), we also define an empirical best predictor for the small area d mean of Y (denoted by TrEBP) under 'fitted value' model (9) as

$$\hat{\bar{Y}}_d^{EBP} = N_d^{-1}\left\{\sum_{s_d} y_i + \sum_{r_d} \hat{\pi}_i \hat{\tilde{y}}_i\right\} = N_d^{-1}\left\{\sum_{s_d} y_i + \sum_{r_d} \left(\hat{\pi}_i k_i^{-1} e^{Z_i\hat{\beta}+\hat{v}_{ii}/2}\right)\right\} \tag{12}$$

where $\hat{k}_i; i \in s_d$ is define below (3).

2.3 Mean squared error estimation

For the estimation of mean squared error (MSE) of (11) we follow Chandra and Chambers (2005, 2006) approach and adapt standard methods for estimating the mean squared error of a weighted linear estimator. This approach treats (11) as simple weighted domain mean estimate. Under this approach the sample weights derived from (9) are treated as fixed and the prediction variance of (11) is estimated using a standard robust variance estimator. See Royall and Cumberland (1978). A "plug-in" estimate of the squared bias of (11) under this model is added to this estimated prediction variance to finally define a simple estimate of the mean squared errors. This MSE estimator is consistent for MSE of MBD under linear mixed model (Chambers, Chandra and Tzavidis, 2007). In contrast, MSE estimation of EBP (13) is not straightforward. We can use resampling methods for MSE estimation of (12). See Jiang, Lahiri and Wang (2002) and Maiti (2004). In this paper we do not pursue the MSE estimation of (12).

3. Monte Carlo simulation experiment

In this section, we design series of simulation studies to contrast the performance of different SAE estimators. In particular, we considered four different SAE estimators in our simulation studies. These are described as:

(i) the HT type MBD estimator for SAE of skewed data with zeros (11) based on model-calibrated sample weights (10) derived via 'fitted value' model (9), denoted by TrMBD

(ii) the empirical best predictor (12) under 'fitted value' model (9), denoted by TrEBP

(iii) the Hájek types MBD estimators based on sample weights derived under a linear mixed model (Chandra and Chambers, 2005), denoted by MBD0, and

(iv) the empirical best linear unbiased predictor under a linear mixed model (Rao, 2003), denoted by EBLUP.

Note that the model-calibrated EBLUP weights (10) derived under fitted value model (9) within small areas produces extremely variable estimates of the small area population sizes, implying that these weights cannot be considered as 'multipurpose'-they function well when used with variables that are reasonably correlated with the variable that defines the fitted value model, but can fail with other, less well correlated, variables (e.g. the indicator variable for small area inclusion). Obviously, as mentioned earlier, the HT type of MBD estimator is better choice in this situation. See Chandra and Chambers (2006). We further note that this problem does not arise with the 'standard' EBLUP weights, as the Hájek type and HT type MBD estimators derived under a linear mixed model are very close in their performances. Consequently, for sample weights derived under raw scale linear mixed model we considered the Hájek type of MBD estimator (MBD0).

We computed three measures of estimation performance using estimates generated in the simulation study. These are the relative bias (RB) and the relative root mean squared error (RRMSE), both expressed as percentages, of regional mean estimates and the

coverage rate (CR) of nominal 95 per cent confidence intervals for regional means. In the evaluation of coverage performances intervals are defined by the small area mean estimates plus or minus twice their standard error. These are defined as below.

The percentage relative bias, defined as

$$RB(\hat{T}_d) = \left(R^{-1}\sum_{r=1}^{R} T_{d(r)}\right)^{-1}\left\{\left(R^{-1}\sum_{r=1}^{R}\hat{T}_{d(r)}\right)-\left(R^{-1}\sum_{r=1}^{R} T_{d(r)}\right)\right\}\times 100\,. \quad (13)$$

The percentage relative root mean squared error, defined as

$$RRMSE(\hat{T}_d) = \left(R^{-1}\sum_{r=1}^{R} T_{d(r)}\right)^{-1}\left\{\sqrt{R^{-1}\sum_{r=1}^{R}\left(\hat{T}_{d(r)}-T_{d(r)}\right)^2}\right\}\times 100\,. \quad (14)$$

The coverage rate, defined as

$$CR(\hat{T}_d) = R^{-1}\sum_{r=1}^{R} 1\left\{T_d \in \left(\hat{T}_{d(r)} \pm 2\sqrt{mse(\hat{T}_{d(r)})}\right)\right\}. \quad (15)$$

Here $\hat{T}_d$ is the estimator (e.g. for the mean) for the small area *d* for parameter T_d and $\hat{T}_{d(r)}$ is the specific outcome of $\hat{T}_d$ obtained in the simulation $r\,(r = 1,\ldots..,R = 1000)$, $mse(\hat{T}_{d(r)})$ is the estimate of the MSE of $\hat{T}_{d(r)}$ given by the data for the r^{th} simulation.

In simulation studies, population and sample data are generated under the model. We choose a population size $N = 15{,}000$ and a sample size $n = 600$ and then generated randomly small area population sizes N_d, $\mathrm{d} = 1,...,D; \sum_d N_d = N$ and sample sizes as $n_d = N_d(n/N); \sum_d n_d = n$. The average sizes of small area population and sample are 500 and 20 respectively with total of $D = 30$ areas. These are fixed for all simulations. We carry out following model-based simulations:

1. We generated population values of $y_{di}\left(i = 1,...,N_d; d = 1,...,D\right)$ from a multiplicative model $y_{di} = 5.0x_{di}^2 u_d e_{di}$, which is linear on log-log scale. Then we created zero values for y_{di} randomly. The random errors e_{di} are independently

generated from a lognormal distribution, LN $(0, \sigma_e = 1.0)$. The random effects u_d are generated from LN $(0, \sigma_u = 0.5)$. The covariate values x_{di} are generated from LN $(5, \sigma_x = 1.2)$. From this model, values of the y_{di} (that contains zeros values as well) are generated for 30 small areas of sizes N_d and then random samples of sizes n_d are drawn from each area. We consider following two combinations for the simulation experiments:

Set 1: We created data with p= 0.50 and 0.75 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. Results from this simulation are presented in Table 1.

Set 2: We created data with proportion of positive values p=0.90 for 25 areas and different p values for 5 selected areas (these are area numbers 5, 10, 15, 20 and 25) at population level. The proportion of positive values for area numbers 5, 10, 15, 20 and 25 are 0.25, 0.35, 0.50, 0.65 and 0.75 respectively. Results generated from this simulation are set out in Table 2 and Figures 1-2.

2. This simulation set examines the performances of different methods of SAE under model misspecifications. Here population values for $y_{di}'s \left(i = 1,...,N_d; d = 1,....,D\right)$ are generated from $y_{di} = 5.0 x_{di}^2 \left\{\exp\left[\log(x_{di})\right]^2\right\} u_d e_{di}$. Note that this model is not linear on log-log scale. We then generated zero values for some of the y_{di} randomly. The values of covariate x_{di} are generated from LN $(3, \sigma_x = 0.2)$. The random errors e_{di} and random area effects u_d are independently generated from a lognormal distribution, with LN $(0, \sigma_e = 1.0)$ and LN $(0, \sigma_u = 0.5)$. Then values of the y_{di} (that contains both zeros and positive values) are generated for 30 small

areas of sizes N_d and then random samples of sizes n_d are drawn from each area. The combination of this simulation is denoted by set-3 as below.

Set 3: Like set-2, at population level we created data with proportion of positive values p=0.90 for 25 areas and varying p values for rest 5 areas (these are area 5, 10, 15, 20 and 25 with p=0.25, 0.35, 0.50, 0.66 and 0.75 respectively). Results from this simulation are shown in Table 3.

Table 1 presents the average values of relative bias, ratio of relative root mean squared error to EBLUP and the average coverage rate for the different methods from simulation set-1. In this simulation set the proportion of zeros (p=0.50 and p=0.75) are same for all areas. We observed an improvement (in terms of biases, RMSE and coverage rate) in the performance of all methods as proportion of zeros decreases (from 50% to 25%) in the data. These results show the average relative bias and the average root mean squared error (RRMSE) of MBD0 and EBLUP are larger than both TrMBD and TrEBP. The average relative bias of TrMBD is marginally higher than TrEBP, however, the average relative RMSE of TrMBD is smaller than the TrEBP. With same magnitude of average relative biases, EBLUP method dominates to MBD0 in terms of RRMSEs. In terms of coverage performances there is not much to choose.

Note that simulation set-1 and results reported in Table 1 corresponds to data that contains equal proportion of zeros in all areas. However, in simulation set-2 proportions of zeros varies for different areas from 25 to 90%. Table 2 set out the average (and median) values of relative bias, ratio of relative root mean squared error to EBLUP and the average coverage rate generated by different methods from this simulation set. Figure 1 and 2 present region-specific results. These results further show that TrMBD and TrEBP dominate EBLUP and MBD0. Area-specific results reflect that areas with

relatively large proportion of zeros (For example, areas 5, 10, 15 and 20 with p= 0.25, 0.35, 0.50 and 0.65 respectively) EBLUP method is very unstable. In these areas synthetic part of EBLUP contributed extremely high, which results in over estimation. In such cases MBD estimator still works well except in presence of outliers data since the MBD methods are sensitive to outliers (Chambers and Chandra, 2006). These results clearly show TrMBD is efficient overall. In contrast, TrEBP methods with marginally large biases and higher values of RRSME than TrMBD are sensitive to presence of zeros especially when a GLM method was used for estimating probabilities (although we have not presented the results based on GLM methods of estimating the probability). Use of area-specific proportion for the probability of positive values seems to working reasonably well for both TrMBD and TrEBP. Overall mixture model based methods of SAE for skewed data with zeros lead to efficient estimates for small areas with smaller relative biases and relative root mean squared errors and with relatively good coverage properties.

The values of relative biases and relative root mean squared errors generated by simulation set-3, both expressed in terms of percentage are presented in Table 3. These results are generated under wrong model choices. In Tables 3 we see that region-specific results contain lot of outlying estimates. These results show median relative bias of MBD0 is smaller overall. Between 'expected-value' model based methods, TrMBD has smaller bias than TrEBP. Although results are not reported here, under TrMBD, it hardly makes any difference due to three methods used for estimation the probabilities of positive values. Under TrEBP, the use of proportions seems to be more appropriate. We note for the areas with large proportion of zeros use of GLM based estimation leads to very unstable results with larges biases for TrEBP. In contrast, use of GLMM in this

situation is as good as area specific proportions. That is for zeros contaminated areas it is important to have area effects in estimating the probabilities when using TrEBP. Further, as noticed earlier region-specific results show in zeros inflated regions (5, 10, 15 and 20) EBLUP is very unstable and MBD0 is relatively better. In Table 3 we further noticed that even under wrong model choices median values of RRMSE of TrMBD are smaller overall. The MBD0 is performing better in terms of relative bias but very unstable. These results conclude that under wrong model choices proposed method have large biases than MBD0 and EBLUP but smaller RRMSE. If model holds, the method produce estimates with both smaller bias and RMSE. Note that for zero inflated regions, TrMBD works well for all three choices for estimating the probabilities. However, in this situation, GLM is not a good choice for estimating the probabilities for the TrEBP (area specific proportion or GLMM is preferable) since it produces biased results. Overall mixture model based methods of SAE for the skewed data with zeros shows a significant gain when model hold. In the event slight model misspecifications, the methods still work well with marginal gain.

4. Concluding remarks

In this paper we introduced small area estimation for skewed data that contain a substantial proportion of zeros. We use a mixture type of model for this purpose. The idea of using a mixture model for skewed data that contain a large proportion of zeros is not new. See Welsh *et al*.(1996), Lambert (1992), Karlberg (2000) and Fletcher *et at*. (2005). However, we apply this approach in context of small area estimation. Our results from simulation experiments show the method works well and produce efficient set of small area estimates. We described two different estimators based on 'expected-value'

model (9) derived from mixture model. We conclude that TrMBD (H-T type MBD estimator based on ratio specification of expected value model) is more efficient. Further, identification of appropriate model relationship on transform scale is very crucial in application this method otherwise results can be misleading. In this paper we used mixture type of model, however it is interesting to model such data under generalized linear mixed model with Gamma and Poisson (for count data) or other class of distributions for skewed data with zeros. We are currently working on these issues.

References

Chambers, R.L. and Chandra H., 2006. Improved direct estimators for small areas. Methodology Working Paper M06/07. Southampton Statistical Science Research Institute, University of Southampton, U.K.

Chambers, R.L., Chandra, H. and Tzavidis, N., 2007. On mean squared error estimation for linear predictors of small area means. Paper presented in the SAE 2007 conference PISA, Italy.

Chandra, H. and Chambers, R., 2006. Small area estimation with skewed data. Submitted.

Chandra, H., 2006. Small area estimation for business surveys. Proceedings of the Amer. Statist. Assoc. Section on Survey Research Methods. 2803-2809.

Chandra, H., and Chambers, R.L., 2005. Comparing EBLUP and C-EBLUP for small area estimation. Statistics in Transition, 7, 637-648.

Casella, G. and Berger, R.L., 1990. Statistical Inference. Duxbury Press, Belmont, California.

Jiang, J., Lahiri, P. and Wang, S-M., 2002. A unified jackknife theory for empirical best prediction with M- estimation. Ann. Statist. 30, 1782-1810.

Fletcher, D., MacKenzie, D. and Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. J. Envir. and Eco. Statist. 12 (1), 45-54.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc. 72, 320–338.

Karlberg, F., 2000. Survey estimation for highly skewed populations in the presence of zeros. J. Off. Statist. 16 (3), 229-241.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 34, 1–14.

Maiti, T., 2004. Applying jackknife method of mean squared prediction error estimation in SAIPE. Statistics in Transition. 6(5), 685-695.

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

Royall, R.M., 1976. The linear least-squares prediction approach to two-stage sampling. J. Amer. Statist. Assoc. 71, 657-664.

Royall, R.M., and Cumberland, W.G., 1978. Variance estimation in finite population sampling. J. Amer. Statist. Assoc.73, 351-358.

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.

Wu, C., and Sitter, R.R., 2001. A model calibration approach to using complete auxiliary information from survey data. J. Amer. Statist. Assoc. 96, 185-193.

Table 1 Average (ARB) values of relative bias, average (ARRMSE) values of relative root mean squared error and average (ACR) coverage rate for simulation set-1.

	ARB,%		Ratio of ARRMSE to EBLUP		ACR	
Methods	p =0.50	p =0.75	p =0.50	p=0.75	p =0.50	p =0.75
TrMBD	1.15	0.76	0.83	0.73	0.90	0.89
TrEBP	0.72	0.39	0.84	0.80	*	*
MBD0	-9.69	-8.02	2.04	2.03	0.83	0.86
EBLUP	-9.57	-7.54	1.00	1.00	0.87	0.87

* we do not pursue MSE estimation for the TrEBP.

Table 2 Average (ARB) and median (MRB) values of relative bias, average (ARRMSE) values of relative root mean squared error and average (ACR) coverage rate for simulation set-2.

Methods	ARB,%	MRB,%	Ratio of ARRMSE to EBLUP	ACR
TrMBD	0.03	0.59	0.62	0.89
TrEBP	0.78	0.84	0.69	*
MBD0	-3.73	-3.02	2.28	0.88
EBLUP	4.16	-9.10	1.00	0.88

* we do not pursue MSE estimation for the TrEBP.

Table 3 Relative biases and relative RMSE for simulation set-3.

	Relative Bias				Relative RMSE			
Region	TrMBD	TrEBP	MBD0	EBLUP	TrMBD	TrEBP	MBD0	EBLUP
1	-8.09	-9.62	-1.55	-5.85	51.4	66.4	119.1	68.0
2	-7.04	-10.58	3.42	-7.12	54.8	70.8	164.8	75.6
3	-9.58	-11.45	12.90	-8.36	46.2	64.4	233.7	67.1
4	-8.61	-11.32	-10.34	-9.85	78.6	83.2	146.4	86.2
5	-19.81	-9.26	-0.95	110.70	69.7	65.0	211.7	157.2
6	-9.94	-9.97	-6.72	-8.71	45.3	61.4	111.0	63.9
7	-11.32	-8.84	-15.09	-7.84	46.7	61.0	118.4	64.2
8	-5.89	-6.68	-9.04	-4.21	42.1	60.6	107.7	65.4
9	-8.70	-10.96	-7.87	-8.67	54.7	67.2	124.6	69.3
10	-22.96	-8.07	-11.11	94.10	81.0	80.4	207.7	146.1
11	-8.63	-9.26	-7.97	-8.09	47.6	63.3	122.0	66.3
12	-7.99	-8.35	0.30	-2.60	42.9	60.9	143.4	72.2
13	0.06	-8.63	17.02	-1.42	54.2	63.9	196.7	68.4
14	-8.78	-12.23	5.11	-7.01	44.1	59.5	211.8	60.7
15	-22.27	-9.13	-7.36	46.30	69.3	73.6	142.6	98.9
16	-9.18	-7.11	-4.37	-4.05	49.0	57.6	121.4	66.1
17	-6.66	-10.39	15.04	-6.35	41.2	57.7	245.9	60.9
18	-9.51	-9.79	-3.46	-7.03	38.7	63.0	144.8	64.0
19	-7.51	-9.96	0.46	-6.80	43.1	59.7	164.2	63.8
20	-14.24	-8.20	2.01	18.14	82.1	90.1	154.8	98.7
21	-6.58	-4.49	3.30	-0.83	40.3	56.6	126.8	57.1
22	-10.08	-13.87	2.46	-9.60	46.4	66.4	191.6	64.4
23	-11.91	-12.19	-12.55	-10.36	51.8	72.2	119.8	79.1
24	-7.69	-5.36	-8.56	-3.66	44.3	58.7	110.6	57.9
25	-11.26	-11.56	4.24	3.79	54.5	62.2	288.8	77.2
26	-7.73	-11.04	0.65	-5.69	39.4	54.5	131.7	60.9
27	-15.00	-13.12	-9.82	-11.98	37.3	47.3	95.6	50.8
28	-12.36	-11.11	-3.29	-8.73	53.8	69.5	144.5	74.4
29	-8.96	-10.20	-2.86	-7.91	45.3	62.6	113.0	67.0
30	-11.87	-15.01	-11.91	-11.87	68.5	81.5	112.1	85.8
Average	-10.34	-9.93	-2.26	3.28	52.1	65.4	154.4	75.2
Median	-9.07	-9.97	-3.08	-6.91	47.2	63.2	143.0	67.0

Figure 1 Region-specific relative biases for TrMBD (solid line), TrEBP (thin line), EBLUP (dash line) and MBD0 (dotted line) methods from simulation set-2.

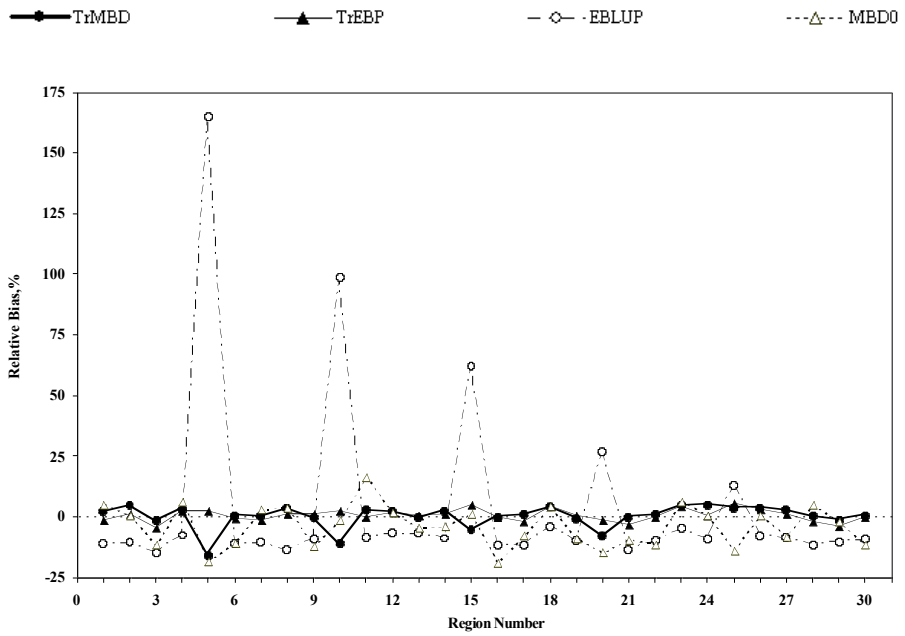

Figure 2 Region-specific relative root mean squared errors for TrMBD (solid line), TrEBP (thin line) and EBLUP (dash line) methods from simulation set-2.