UNIVERSITY
OF WOLLONGONG
AUSTRALIA

University of Wollongong
Research Online

Faculty of Engineering and Information Sciences - Papers: Part A

Faculty of Engineering and Information Sciences

2016

Small area estimation for semicontinuous data

Hukum Chandra
Indian Agricultural Statistics Research Institute, hchandra@uow.edu.au

Raymond L. Chambers
University of Wollongong, ray@uow.edu.au

Publication Details

Chandra, H. & Chambers, R. L. (2016). Small area estimation for semicontinuous data. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 58 (2), 303-319.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Small area estimation for semicontinuous data

Abstract
Survey data often contain measurements for variables that are semicontinuous in nature, i.e. they either take a single fixed value (we assume this is zero) or they have a continuous, often skewed, distribution on the positive real line. Standard methods for small area estimation (SAE) based on the use of linearmixed models can be inefficient for such variables. We discuss SAE techniques for semicontinuous variables under a two part random effects model that allows for the presence of excess zeros as well as the skewed nature of the nonzero values of the response variable. In particular, we first model the excess zeros via a generalized linear mixed model fitted to the probability of a nonzero, i.e. strictly positive, value being observed, and then model the response, given that it is strictly positive, using a linear mixed model fitted on the logarithmic scale. Empirical results suggest that the proposed method leads to efficient small area estimates for semicontinuous data of this type. We also propose a parametric bootstrap method to estimate the MSE of the proposed small area estimator. These bootstrap estimates of the MSE are compared to the true MSE in a simulation study.

Disciplines
Engineering | Science and Technology Studies

Publication Details
Chandra, H. & Chambers, R. L. (2016). Small area estimation for semicontinuous data. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 58 (2), 303-319.

This journal article is available at Research Online: http://ro.uow.edu.au/eispapers/4020

Small area estimation for semicontinuous data

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

[1] Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India.
[2] National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, NSW, 2522, Australia.

Abstract

Survey data often contain measurements for variables that are semicontinuous in nature, i.e. they either take a single fixed value (we assume this is zero) or they have a continuous, often skewed, distribution on the positive real line. Standard methods for small area estimation (SAE) based on the use of linear mixed models can be inefficient for such variables. We discuss SAE techniques for semicontinuous variables under a two part random effects model that allows for the presence of excess zeros as well as the skewed nature of the non-zero values of the response variable. In particular, we first model the excess zeros via a generalized linear mixed model fitted to the probability of a non-zero, i.e. strictly positive, value being observed, and then model the response, given that it is strictly positive, using a linear mixed model fitted on the logarithmic scale. Empirical results suggest that the proposed method leads to efficient small area estimates for semicontinuous data of this type. We also propose a parametric bootstrap method to estimate the MSE of the proposed small area estimator. These bootstrap estimates of the MSE are compared to the true MSE in a simulation study.

Key words: Mean squared error; Parametric bootstrap; Skewed data; Small area estimation; Zero-inflated.

1. Introduction

Many variables of interest in business, agricultural, environmental, ecological and epidemiological surveys are semicontinuous in nature, i.e. they either take a single fixed value (typically zero) or they have a continuous, often skewed, distribution on the positive real line. This article focuses on a particular type of semicontinuous variable frequently encountered in practice, a mixture of zeros and continuous strictly positive

[*] ***Corresponding author:*** e-mail: hchandra12@gmail.com, Phone: 0091-11-25841475, Fax: 0091-11-25841564

values that are generally skewed. Such a semicontinuous variable is quite different from one that has been left-censored or truncated, because the zeros are valid self-representing data values, not proxies for negative or missing responses. It is therefore natural to view a semicontinuous response of this type as the result of two processes, one determining whether the response is zero and the other determining the actual level if it is non-zero (Olsen and Schafer, 2001). Measurements of indebtedness, investment, production or amount of stock on hand all represent situations where semicontinuous data are typically collected in household and business surveys. For example, Amount of Loan Outstanding (collected in the 59th Round of the National Sample Survey, or NSS, in India), and Closing Beef Cattle, or BEEFCL (collected in the Australian Agricultural Grazing Industries Survey, or AAGIS) are just two cases of important survey output variables that are, by their definition, semicontinuous. In both, the target variable is either zero or some positive value, with these positive values then having a skewed distribution. Unlike the NSS data, an anonymised version of the AAGIS data is available, and so these data are used in the empirical evaluations presented in Section 5, which focus on regional estimation for BEEFCL. See Figure 1 and Table 4 for the distributions of regional sample sizes and proportions of zero values in the AAGIS sample data, while the sample distribution of BEEFCL in these data is shown in Figure 2. It is clear from Figures 1 and 2 that BEEFCL is zero-inflated with highly skewed non-zero values.

Since a linear model is not appropriate for a semicontinuous variable, commonly used methods for small area estimation based on the use of linear mixed models (e.g. the empirical best linear unbiased predictor or EBLUP) can be inefficient for such variables (see Rao, 2003). Chandra and Chambers (2011a) and Berg and Chandra (2012) investigate small area estimation methods for skewed variables, focussing on the case where a linear mixed model is appropriate after a logarithmic (log) transformation. Chandra and Chambers (2011a) describe two methods of small area estimation for such positively skewed variables. The first, a model-based direct estimator or MBDE, is defined as a weighted sum of the sampled units in the small area, with weights constructed so as to lead to the minimum mean squared error linear predictor of the overall population mean if the parameters of the log scale linear mixed model were known. The second, based on the approach of Karlberg (2000), uses an empirical

predictor based on a log scale linear mixed model that is analogous to the synthetic estimator under a linear mixed model. 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 between area variation in the model covariates, and can therefore lead to biased estimators when there is significant residual between area heterogeneity. Berg and Chandra (2012) also describe an empirical best predictor that has minimum mean squared error in the class of unbiased predictors when a log scale linear mixed model is appropriate. This predictor allows for between area variation and is indirect, i.e. it uses information from all the small areas. However, all these approaches are restricted to a strictly positive variable, and so cannot be directly applied to a semicontinuous variable.

The presence of excess zeros in survey data is a well known problem, and a variety of approaches have been suggested for addressing it. However, much less is known when the focus is on small area estimation using these data, even though presence of excess zeros within a small area are clearly much more influential than they are in the larger overall sample. A two part random effects model (Olsen and Schafer, 2001), also referred as a mixture model (Fletcher *et al*., 2005), is widely used for small area estimation with zero-inflated variables, see for example, Pfeffermann *et al*. (2008) and Chandra and Sud (2012). In what follows we therefore develop a small area estimation method for semicontinuous variables under a two part random effects model. Here we first model the excess zeros via a generalized linear mixed model fitted to the probability of a non-zero, i.e. strictly positive, value being observed, and then model the response, given that it is strictly positive, using a log scale linear mixed model. These two model components are combined in estimation. We also propose a parametric bootstrap method that can be used estimate the mean squared error (MSE) of our proposed two part estimator.

The structure of the paper is as follows. In Section 2 we develop a number of predictors for a small area mean based on a log scale linear mixed model. In Section 3 we then introduce the two part random effects model (or mixture model) and discuss different

approaches to small area estimation under this model. Section 4 then focuses on MSE estimation via a parametric bootstrap approach. In Section 5 we present results from both model-based as well as design-based simulations which are used illustrate the performances of the different methods of small area estimation discussed in Section 3, with the design-based simulations based on survey data from the AAGIS. Finally, in Section 6 we summarize our main findings and discuss avenues for future research.

2. Small area estimation under transformation to linearity

We assume that a non-informative sampling method is used to draw a sample of size n from a finite population U of size N which consists of D non-overlapping domains U_i $(i = 1,....,D)$. Following standard practice, we refer to these domains as small areas or just areas. We further assume that there is a known number N_i of population units in small area i, with n_i 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 (i.e. $|s_i| = n_i$), 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 denotes the set of 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} denote the value of the variable of interest Y for unit j in area i and $\mathbf{x}_{ij}$ denote the vector of length $m-1$ containing the known values of the auxiliary variables for unit j in area i. Throughout we assume that 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 log scale linear mixed model. That is, y_{ij} satisfies

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

where $\mathbf{z}_{ij} = \left(1_i, \mathrm{g}(\mathbf{x}_{ij})\right)$ is the $m \times 1$ vector of covariates defined by appropriate transformation of the auxiliary variables, $\boldsymbol{\beta}$ is a $m \times 1$ vector of fixed effects, u_i is a random effect associated with area i and e_{ij} is an individual level random effect for unit j

in small area *i*. Following standard practice, we assume that the area and individual effects are mutually independent, with the area effects independently and identically distributed as $u_i \square N(0,\sigma_u^2)$ and the individual effects independently and identically distributed as $e_{ij} \square N(0,\sigma_e^2)$. The sample observations $\{y_{ij}; i=1,....D; j \in s_i\}$ are assumed to be available. We further assume that the population values of $\mathbf{z}_{ij}$ are available, and that they can be linked to the sample. Consequently, the available data for area *i* are $\{(y_{ij}, \mathbf{z}_{ij}); i=1,....D; j \in s_i\} \cup \{\mathbf{z}_{ij}; i=1,....D; j \in r_i\}$. 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 Maximum Likelihood (ML) or the Restricted Maximum Likelihood (REML) estimator of $\boldsymbol{\varphi}$. In particular, $\boldsymbol{\sigma}^2 = (\sigma_u^2, \sigma_e^2)^T$ is usually referred to as the vector of variance components of the model with estimator $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$. Note that since we have assumed a non-informative sampling method, the sample and population distributions of the data are the same, and are given by (1).

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 estimator of $\boldsymbol{\beta}$, and $\hat{u}_i = \hat{\gamma}_i (\bar{l}_{is} - \bar{\mathbf{z}}_{is}^T \hat{\boldsymbol{\beta}})$ is the empirical best linear unbiased predictor (EBLUP) of the random area effect. Here $\hat{\gamma}_i = \hat{\sigma}_u^2 (\hat{\sigma}_u^2 + n_i^{-1} \hat{\sigma}_e^2)^{-1}$ is the plug-in estimator of the shrinkage effect $\gamma_i = \sigma_u^2 (\sigma_u^2 + n_i^{-1} \sigma_e^2)^{-1}$, and $\bar{l}_{is} = n_i^{-1} \sum_{j \in s_i} \log(y_{ij})$ 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*. Using a prediction-based approach similar to that described in Karlberg (2000), Chandra and Chambers (2011a) then propose a synthetic type predictor for the area mean m_i under model (1) of the form

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

where

$$\hat{y}_{ij}^{SYN-EP} = \left(\hat{c}_{ij}^{SYN-EP}\right)^{-1} \exp\left\{ \mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}} + 0.5\left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2\right) \right\}$$

and

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

is a Taylor series linearization-based correction for back transformation bias. Note that (2) is not an Empirical Best Predictor since it does not allow for between unit correlation within a small area when it predicts the value of a non-sample y_{ij} given the corresponding sample values for this variable in area *i*. It is therefore a synthetic predictor of the small area mean.

Chandra and Chambers (2011a) also propose a model-based direct estimator (MBDE) of m_i of the form $\sum_{j\in s_i} w_{ij}y_{ij}$, where w_{ij} is an estimator of the weight that leads to the best linear unbiased predictor (BLUP) of the population mean if the parameters of the model (1) are known. To derive this estimator, Chandra and Chambers (2011a) use the approximations,

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

and

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

where $\hat{y}_{ij}^{SYN-EP}$ is given in (2). The approximations (3) and (4) follow from the moment generating function of a normal distribution, and the fact that the covariance between two units from different areas is zero. Put $\mathbf{y}_U = (\mathbf{y}_s^T, \mathbf{y}_r^T)^T$, where $\mathbf{y}_s$ and $\mathbf{y}_r$ are the vectors of sampled and non-sampled units of *Y* respectively. Similarly, let $\hat{\mathbf{y}}_s^{SYN-EP}$ and $\hat{\mathbf{y}}_r^{SYN-EP}$ denote the vectors containing the values $\hat{y}_{ij}^{SYN-EP}$ for the sampled and non-sampled units and define $\mathbf{J}_U = (\mathbf{J}_s^T, \mathbf{J}_r^T)^T = \left((\mathbf{1}_s^T, \mathbf{1}_r^T)^T, ((\hat{\mathbf{y}}_s^{SYN-EP})^T, (\hat{\mathbf{y}}_r^{SYN-EP})^T)^T\right)$. We can then express (3) and (4) in matrix form as

$$\begin{aligned} E(\mathbf{y}_U) &\approx \mathbf{J}_U\boldsymbol{\alpha} \\ V(\mathbf{y}_U) \approx \mathbf{V}_U &= \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}, \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ and the elements of variance-covariance matrix $\mathbf{V}_U$ are given by (4). For known parameters, the model specified in (3) and (4) is referred to as a 'fitted

value' model and corresponds to a linear model for y_{ij}. The BLUP of the population mean $m_U = N^{-1}\sum_{i=1}^{D}\sum_{j=1}^{N_i} y_{ij}$ of *Y* under (5) is then $N^{-1}\mathbf{w}_s^T\mathbf{y}_s$, where

$$\mathbf{w}_s = (w_j; j \in s) = \mathbf{1}_s + \mathbf{H}_s^T(\mathbf{J}_U^T\mathbf{1}_U - \mathbf{J}_s^T\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}_s^T\mathbf{J}_s^T)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r, \quad (6)$$

where $\mathbf{H}_s = (\mathbf{J}_{s\,s}^T\mathbf{V}_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}_s^T\mathbf{V}_{ss}^{-1}$. Note that the weights (6) satisfy $\sum_{i=1}^{D}\sum_{j\in s_i} w_{ij} = N$ and $\sum_{i=1}^{D}\sum_{j\in s_i} w_{ij}\hat{y}_{ij}^{SYN-EP} = \sum_{i=1}^{D}\sum_{j=1}^{N_i}\hat{y}_{ij}^{SYN-EP}$. The MBDE of the small area mean m_i (Chandra and Chambers, 2011a) is then

$$\hat{m}_i^{CC} = N_i^{-1}\sum_{j\in s_i} w_{ij}y_{ij}, \quad (7)$$

where the w_{ij} are the weights (6) associated with the sample units in area *i*. We note that since (7) is a direct estimator, it can lead to unstable estimates when area sample sizes are too small. Balanced against this however is its inherent robustness to misspecification of the model for the y_{ij}.

Finally, Berg and Chandra (2012) use (1) to develop the empirical version of the minimum mean squared error (MMSE) predictor for m_i. This is

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

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 note that (8) allows for between unit correlations within a small area and is therefore an Empirical Best Predictor (EBP) under the normality assumptions of (1). To see this, observe that for non-sample unit $j \in r_i$ the conditional distribution of $l_{ij} = \log(y_{ij})$ given the area *i* sample data $x_{ij}, \{l_{ik}, x_{ik}; k \in s_i\}$ is normal, with

$$E\left(l_{ij}\left|x_{ij}, \{l_{ik}, x_{ik}; k \in s_i\}\right.\right) = E\left(l_{ij}\left|z_{ij}, \bar{l}_{is}, \bar{z}_{is}\right.\right) = \mathbf{z}_{ij}^T\boldsymbol{\beta} + \gamma_i\left(\bar{l}_{is} - \bar{\mathbf{z}}_{is}^T\boldsymbol{\beta}\right)$$

and

$$Var\left(l_{ij}\left|x_{ij}, \{l_{ik}, x_{ik}; k \in s_i\}\right.\right) = \sigma_u^2 + \sigma_e^2 - \sigma_u^4(\sigma_u^2 + n_i^{-1}\sigma_e^2)^{-1} = \sigma_e^2(1+n_i^{-1}\gamma_i)$$

so

$$E\left(y_{ij}\left|x_{ij}, \{y_{ik}, x_{ik}; k \in s_i\}\right.\right) = \exp\left\{\mathbf{z}_{ij}^T\boldsymbol{\beta} + \gamma_i\left(\bar{l}_{is} - \bar{\mathbf{z}}_{is}^T\boldsymbol{\beta}\right) + 0.5\sigma_e^2(1+n_i^{-1}\gamma_i)\right\},$$

which immediately leads to the empirical version (8) of the MMSE predictor (8). Consequently, when (1) holds, i.e. the y_{ij} are lognormally distributed, we expect (8) to dominate (2).

Note that

$$E\left[\hat{y}_{ij}^{EBP}\right] = E\left[\exp\left\{\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \gamma_i\left(\overline{l}_{is} - \overline{\mathbf{z}}_{is}^T\hat{\boldsymbol{\beta}}\right) + 0.5\hat{\sigma}_e^2(1+n_i^{-1}\hat{\gamma}_i)\right\}\right]$$
$$\neq \exp\left\{\mathbf{z}_{ij}^T\boldsymbol{\beta} + \gamma_i\left(\overline{l}_{is} - \overline{\mathbf{z}}_{is}^T\boldsymbol{\beta}\right) + 0.5\sigma_e^2(1+n_i^{-1}\gamma_i)\right\}.$$

That is, the MMSE predictor (8) is biased. Berg and Chandra (2012) use Taylor series approximation to bias correct this predictor. Following their development, a bias corrected version of (8) is

$$\hat{m}_i^{EBP-BC} = N_i^{-1}\left\{\sum\nolimits_{s_i} y_{ij} + \sum\nolimits_{r_i} \hat{y}_{ij}^{EBP-BC}\right\}, \tag{9}$$

where $\hat{y}_{ij}^{EBP-BC} = \left(\hat{c}_{ij}^{EBP}\right)^{-1}\hat{y}_{ij}^{EBP}$, with

$$c_{ij}^{EBP} = \exp\left\{0.5\left(\mathbf{a}_{ij} + \hat{c}_{i1}\hat{V}(\hat{\sigma}_e^2) + \hat{c}_{i2}\hat{V}(\hat{\sigma}_u^2) + 2\hat{c}_{i3}\hat{C}ov(\hat{\sigma}_e^2,\hat{\sigma}_u^2)\right)\right\}.$$

Put $\hat{d}_i = \left(\overline{l}_{is} - \overline{\mathbf{z}}_{is}^T\hat{\boldsymbol{\beta}}\right)$. Then

$$\mathbf{a}_{ij} = \left(\mathbf{z}_{ij}^T - \hat{\gamma}_i\overline{\mathbf{z}}_{is}^T\right)^T\hat{V}(\hat{\boldsymbol{\beta}})\left(\mathbf{z}_{ij}^T - \hat{\gamma}_i\overline{\mathbf{z}}_{is}^T\right),$$

$$\hat{c}_{i1} = \left\{0.5 + 0.5\frac{\hat{\gamma}_i^2}{n_i} - \frac{\hat{\gamma}_i^2\hat{d}_i}{n_i\hat{\sigma}_u^2}\right\}^2 - \left\{\frac{\hat{\gamma}_i^3}{n_i^2\hat{\sigma}_u^2} - \frac{2\hat{\gamma}_i^3\hat{d}_i}{n_i^2\hat{\sigma}_u^4}\right\},$$

$$\hat{c}_{i2} = \left\{\frac{(1-\hat{\gamma}_i)^2}{2} + \frac{\hat{\gamma}_i(1-\hat{\gamma}_i)\hat{d}_i}{\hat{\sigma}_u^2}\right\}^2 - \left\{\frac{\hat{\gamma}_i(1-\hat{\gamma}_i)^2}{\hat{\sigma}_u^2} + \frac{2\hat{\gamma}_i^2(1-\hat{\gamma}_i)}{\hat{\sigma}_u^4}\hat{d}_i\right\},$$

$$\hat{c}_{i3} = \left\{\frac{1}{2} + \frac{\hat{\gamma}_i^2}{2n_i} - \frac{\hat{\gamma}_i^2\hat{d}_i}{n_i\hat{\sigma}_u^2}\right\}\left\{\frac{(1-\hat{\gamma}_i)^2}{2} + \frac{\hat{\gamma}_i(1-\hat{\gamma}_i)\hat{d}_i}{\hat{\sigma}_u^2}\right\} + \left\{\frac{\hat{\gamma}_i^2(1-\hat{\gamma}_i)}{n_i\hat{\sigma}_u^2} - \frac{\hat{\gamma}_i^2(1-2\hat{\gamma}_i)}{n_i\hat{\sigma}_u^4}\hat{d}_i\right\}.$$

3. Small area estimation under a mixture model

We now consider the case where the response variable y_{ij} is semicontinuous. In particular, we shall assume that y_{ij} is either zero or has a skewed distribution over the strictly positive real line. We describe an approach based on modelling this variable via

a two part random effects model (also referred as a mixture model). That is, we shall assume that y_{ij} is drawn from a two-component mixture, where the first component corresponds to a fixed value (zero) and the second component corresponds to a strictly positive random variable with a skewed distribution. Following Olsen and Schafer (2001), Pfeffermann *et al.* (2008), Chandra and Chambers (2011b) and Chandra and Sud (2012), we define *I*(*A*) as the indicator function for the event *A* and write $y_{ij} = 0 \times I\left(y_{ij} = 0\right) + \tilde{y}_{ij} I\left(y_{ij} > 0\right) = \delta_{ij}\tilde{y}_{ij}$, where $\tilde{y}_{ij}$ is referred to as the *log-linear* component of y_{ij} and is assumed to follow the log scale linear mixed model (1). The second component $\delta_{ij} = I\left(y_{ij} > 0\right)$ is assumed to follow a generalized linear mixed model (GLMM) with logit link function (Breslow and Clayton, 1993), and is referred as the *logistic* component of y_{ij}. Note that values of $\tilde{y}_{ij}$ are only observed when $\delta_{ij} = 1$, whereas values of δ_{ij} are always observed.

Small area estimation under this mixture model is implemented in three steps. First, a logistic linear mixed model is fitted to the sample values of the indicator variable δ_{ij}. Second, a log scale linear mixed model is fitted to the positive sample values of the response variable. Finally, predicted values generated under these two models are combined at the estimation stage. Chandra and Chambers (2011b) used a similar mixture model for small area estimation of zero-inflated skewed data. However, their approach focuses on the MBDE estimator for this case, and uses sample weights obtained via the 'fitted value' linear model implied by the two part mixture model. They also develop a MSE estimator based on pseudo-linearization (Chambers *et al.*, 2011). However, as noted earlier, the MBDE is a direct estimator and can be unstable when area specific sample sizes are too small.

Fitting the logistic component of a two part random effects model poses computational challenges similar to those found when fitting generalized linear mixed models. Generally, an approximate Fisher scoring procedure based on higher order Laplace approximations is used to obtain maximum likelihood estimates for the fixed coefficients and variance components, see Olsen and Schafer (2001). Pfeffermann *et al.*

(2008) use a two part random effects model that allows for the random area effects in the two components of the model to be correlated. However, their simulation results show that this correlation does not significantly improve small area estimation. Furthermore, use of this correlation makes model fitting computationally intensive and sometimes numerically unstable. Consequently the area random effects in the two components of the two part random effects model are often assumed to be independent, see for example, Karlberg (2000) and references therein. We shall proceed similarly and assume that the two area random effects are uncorrelated. That is, following the Pfeffermann *et al.* (2008), Chandra and Chambers (2011b) and Chandra and Sud (2012) we assume that the correlation between the two random components δ_{ij} and $\tilde{y}_{ij}$ of the assumed mixture model is negligible. Note that this implies that the mixture model is *not* appropriate if there is reason to believe that the distributions of these components are dependent, e.g. if the observed zeros in the data are due to censoring of $\tilde{y}_{ij}$, as in a Tobit model.

We assume that, given $\mathbf{x}_{ij}$, the δ_{ij} are independent Bernoulli random variables with $\mathrm{P}(y_{ij} > 0) = \mathrm{P}(\delta_{ij} = 1) = p_{ij}$. The model linking the probability p_{ij} with the values of the covariates associated with unit *j* in area *i* is a logistic linear mixed model of the form

$$\mathrm{logit}(p_{ij}) = \ln\left\{p_{ij} / (1 - p_{ij})\right\} = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\theta} + v_i \qquad (10)$$

so $p_{ij} = \exp(\eta_{ij})\left\{1 + \exp(\eta_{ij})\right\}^{-1} = \exp(\mathbf{x}_{ij}^T\boldsymbol{\theta} + v_i)\left\{1 + \exp(\mathbf{x}_{ij}^T\boldsymbol{\theta} + v_i)\right\}^{-1}$. Here $\boldsymbol{\theta}$ is a vector of unknown fixed effects parameters and v_i is the random effect associated with area *i*, assumed to have a normal distribution with zero mean and constant variance σ_v^2. We estimate the parameters of (10) using the procedure described in Saei and Chambers (2003) and Manteiga *et al.* (2007). This is an iterative procedure, implemented in the statistical software package R, that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\theta}$ and v_i with REML estimation of the variance component parameters. Using a 'hat' to denote these estimated values, the predicted probabilities of the logistic component of the two part random effects model are:

$$\hat{p}_{ij} = \exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\left\{1 + \exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\right\}^{-1}. \qquad (11)$$

In order to estimate the parameters of the second log-linear component, of y_{ij}, we denote by $s_+ = \{ j \in s, y_j > 0 \}$ the subset of the sample for which the response variable is non-zero, with $n_+ = \sum_{j \in s} \delta_j$ denoting the number of non-zero sample units. In what follows, we will use a subscript of '+' to denote a quantity associated with these non-zero sample units. Using the data in s_+, we then fit the model (1) to obtain estimates of the fixed effect parameters and the predicted values of the random effects. In particular, the Empirical Best Linear Unbiased Estimator (EBLUE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^+ = \left(\sum_{i=1}^{D} \mathbf{x}_{is_+}^T \hat{\mathbf{v}}_{iss_+}^{-1} \mathbf{x}_{is_+} \right)^{-1} \left(\sum_{i=1}^{D} \mathbf{x}_{is_+}^T \hat{\mathbf{v}}_{iss_+}^{-1} \mathbf{y}_{is_+} \right).$$

Here $\hat{\mathbf{v}}_{iss_+} = diag\left\{ \hat{\sigma}_u^{2+} \mathbf{1}_{is_+} \mathbf{1}_{is_+}^T + \hat{\sigma}_e^{2+} \mathbf{I}_{is_+} \right\}$, with $\mathbf{1}_{is_+}$, $\mathbf{I}_{is_+}$ equal to the unit vector of length n_i^+ and the identity matrix of dimension n_i^+ respectively, where n_i^+ denotes the number of area *i* units in s^+. The corresponding Empirical Best Linear Unbiased Predictors (EBLUPs) for the random area effects are given by $\hat{u}_i^+ = \hat{\gamma}_i^+ \left(\bar{l}_{is_+} - \bar{\mathbf{z}}_{is_+}^T \hat{\boldsymbol{\beta}}^+ \right)$ with $\hat{\gamma}_i^+ = \hat{\sigma}_u^{2+} (\hat{\sigma}_u^{2+} + n_{i+}^{-1} \hat{\sigma}_e^{2+})^{-1}$. The estimated values of y_{ij} can then be obtained using (2) or (9). The first option leads to a synthetic type predictor while the second, after correction for back transformation bias, leads to an empirical version of the minimum mean squared error predictor, i.e. an EBP, for y_{ij}. The synthetic type predictor is

$$\hat{y}_{ij}^{SYN-EP+} = \exp\left\{ \left(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}^+ + \frac{1}{2} (\hat{\sigma}_u^{2+} + \hat{\sigma}_e^{2+}) \right) - \left(\frac{1}{2} \mathbf{z}_{ij}^T \hat{V}(\hat{\boldsymbol{\beta}}^+) \mathbf{z}_{ij} + \frac{1}{4} \hat{V}(\hat{\sigma}_u^{2+} + \hat{\sigma}_e^{2+}) \right) \right\}, \quad (12)$$

while the EBP is

$$\hat{y}_{ij}^{EPB-BC+} = \left(\hat{c}_{ij}^{EPB+} \right)^{-1} \exp\left\{ \mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}^+ + \hat{\gamma}_i^+ \left(\bar{l}_{is} - \mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}^+ \right) + \frac{1}{2} \hat{\sigma}_e^{2+} (1 + n_i^{-1} \hat{\gamma}_i^+) \right\}, \quad (13)$$

with

$$\hat{c}_{ij}^{EBP+} = \exp\left\{ 0.5 \left(\mathbf{a}_{ij}^+ + \hat{c}_{i1}^+ \hat{V}(\hat{\sigma}_e^{2+}) + \hat{c}_{i2}^+ \hat{V}(\hat{\sigma}_u^{2+}) + 2 \hat{c}_{i3}^+ \hat{Cov}(\hat{\sigma}_e^{2+}, \hat{\sigma}_u^{2+}) \right) \right\}$$

where

$$\mathbf{a}_{ij}^+ = \left(\mathbf{z}_{ij}^T - \hat{\gamma}_i^+ \bar{\mathbf{z}}_{is}^T \right)^T \hat{V}(\hat{\boldsymbol{\beta}}^+) \left(\mathbf{z}_{ij}^T - \hat{\gamma}_i^+ \bar{\mathbf{z}}_{is}^T \right)$$

and $\hat{c}_{i1}^{+}$, $\hat{c}_{i2}^{+}$ and $\hat{c}_{i3}^{+}$ are obtained from $\hat{c}_{i1}$, $\hat{c}_{i2}$ and $\hat{c}_{i3}$ by replacing the parameter estimates $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2\right)$ by $\left(\hat{\boldsymbol{\beta}}^{+}, \hat{\sigma}_e^{2+}, \hat{\sigma}_u^{2+}\right)$.

Let E_1 denote expectation with respect to unit level (level 1) variability in y_{ij}. That is, this expectation conditions on the random area effects in the logistic and log-linear components of the two part model. Then, setting $\mu_{ij} = E_1(\tilde{y}_{ij})$, we see that under independence of these area effects,

$$E_1(y_{ij}) = E_1\left(\delta_{ij}\tilde{y}_{ij}\right) = E_1\left(\delta_{ij}\right)E_1\left(\tilde{y}_{ij}\right) = p_{ij}\mu_{ij} \tag{14}$$

where p_{ij} was defined following (10). Substituting predicted values for p_{ij} and μ_{ij} in (14) leads to a plug-in predicted value for y_{ij},

$$\hat{E}(y_{ij}) = \hat{p}_{ij}\hat{\mu}_{ij}, \tag{15}$$

where $\hat{p}_{ij}$ is given by (11), and $\hat{\mu}_{ij} = \hat{E}_1\left(\tilde{y}_{ij}\right)$ can be calculated using either (12) or (13). That is, we have two different predicted values:

$$\hat{y}_{ij}^{MixEP} = \hat{p}_{ij}\hat{y}_{ij}^{SYN-EP+} \tag{16}$$

and

$$\hat{y}_{ij}^{MixEBP} = \hat{p}_{ij}\hat{y}_{ij}^{EPB-BC+}. \tag{17}$$

As usual. let a 'hat' denote an estimated value. Then, for non-sample unit *j* in area *i*, we see that we can write $\hat{y}_{ij}^{MixEP} = \hat{E}\left(y_{ij}\left|\mathbf{x}_{ij}\right.\right)$, while $\hat{y}_{ij}^{MixEBP} = \hat{E}\left(y_{ij}\left|\mathbf{x}_{ij}, \mathbf{y}_{is}, \mathbf{x}_{is}\right.\right)$.

The two predictors (16) and (17) allow us to define three different estimators for population mean of *Y* in small area *i* as follows:

(i) Using (16) we can calculate a synthetic type estimator of the form

$$\hat{m}_i^{MixEP} = \hat{E}\left\{m_i\left|\mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\right.\right\} = N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_{ij} + \sum\nolimits_{j\in r_i} \hat{y}_{ij}^{MixEP}\right], \tag{18}$$

which we denote by *Mix*EP in what follows;

(ii) The fitted values $\hat{y}_{ij}^{MixEP}$ that define the synthetic estimator (18) can also be used to define a 'fitted value' covariate in a linear model for y_{ij}. This model is then used to calculate sample weights w_{ij}^{MixEP} via (6), and an MBDE based on these weights computed as

$$\hat{m}_i^{MixMBDE} = N_i^{-1} \sum_{j \in s_i} w_{ij}^{MixEP} y_{ij} . \quad (19)$$

We denote this estimator by *Mix*MBDE in what follows;

(iii) Using (17) we can calculate an EBP type estimator of the form

$$\hat{m}_i^{MixEBP} = \hat{E}\left\{ m_i \left| \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir} \right. \right\} = N_i^{-1} \left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{MixEBP} \right], \quad (20)$$

which we denote by *Mix*EBP in what follows.

4. Mean squared error estimation

Analytic estimators of the MSE of nonlinear small area estimators are technically complex to derive and typically involve a considerable degree of approximation. As a consequence, a number of numerically intensive, but computationally tractable, methods for MSE estimation have been proposed, e.g. the jackknife method of Jiang, Lahiri and Wan (2002) and the bootstrap methods described in Hall and Maiti (2006) and Manteiga *et al.* (2007, 2008) and references therein. By construction, the small area predictors (18) and (20) are non-linear with complex structure and so obtaining a closed form expression for their corresponding MSEs is not straightforward. We therefore adopt a bootstrap approach when estimating the MSE of (18) and (20). In particular, we use the parametric bootstrap method defined by the steps in the following algorithm. Note that we use an estimator $\hat{m}_i$ of the area *i* mean m_i to motivate the algorithm, but it is generally applicable to estimators of any set of finite population parameters defined on the survey population.

Step 1. Fit the log scale linear mixed model (1) to the positive values y_{ij} in the sample data to obtain the estimates $\hat{\boldsymbol{\varphi}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$.

Step 2. Given the estimates $\hat{\boldsymbol{\varphi}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$, generate area-specific random errors from a lognormal distribution $u_i^* \sim LN\left(0, \hat{\sigma}_u^2\right), i = 1, \ldots, D$ and individual level random errors from an independent lognormal distribution $e_{ij}^* \sim LN\left(0, \hat{\sigma}_e^2\right)$, $j = 1, ..., N_i; i = 1, \ldots, D$.

Step 3. Similarly, fit the logistic linear mixed model (10) to the sample values of the binary variable δ_{ij} and compute $\hat{\boldsymbol{\theta}}$ and $\hat{v}_i$.

Step 4. Given $\hat{\boldsymbol{\theta}}$ and $\hat{v}_i$, calculate probabilities $\hat{p}_{ij}^* = \exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\left\{1 + \exp(\mathbf{x}_{ij}^T\hat{\boldsymbol{\theta}} + \hat{v}_i)\right\}^{-1}$, and hence generate independent binary values δ_{ij}^*, $j = 1, ..., N_i; i = 1, \ldots, D$ satisfying $\mathrm{P}\left(\delta_{ij}^* = 1\right) = \hat{p}_{ij}^*$.

Step 5. Calculate bootstrap population data $\left(y_{ij}^*, \mathbf{x}_{ij}\right)$ under the two part model using

$$y_{ij}^* = \left(\hat{\beta}_0 x_{ij}^{\hat{\beta}_1} u_i^* e_{ij}^*\right)\delta_{ij}^*, \; j = 1, ..., N_i; i = 1, \ldots, D, \tag{21}$$

and then calculate the corresponding value of the area *i* mean $m_i^* = N_i^{-1}\sum_{j \in i} y_{ij}^*$.

Step 6. Let $\mathbf{y}_s^* = \left(y_{ij}^*; \; j \in s_i; i = 1, \ldots, D\right)$ denote the vector of bootstrap sample values for this population. Using these values, calculate the estimate $\hat{m}_i^*$ of the area *i* population mean.

Step 7. Repeat steps 2 - 6 independently *B* times to generate the bootstrap distribution $\left(m_i^{*(b)}, \hat{m}_i^{*(b)}; b = 1, \ldots, B\right)$ of values for m_i^* and $\hat{m}_i^*$.

Step 8. Calculate the bootstrap estimate of the MSE of the actual sample-based estimate $\hat{m}_i$ of m_i as

$$\mathrm{mse}^{boot}(\hat{m}_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{m}_i^{*(b)} - m_i^{*(b)}\right)^2. \tag{22}$$

5. Empirical evaluations

In this Section we report the results from a limited set of empirical evaluations that illustrate the performance of the different estimators of small area means described in the preceding sections, and their corresponding MSE estimators. These estimators are set out in Table 1. Note that for the commonly used linear mixed model EBLUP, denoted by *Lin*EBLUP and which served as the baseline estimator in our simulations, we used the

MSE estimator of Prasad and Rao (1990). For the mixture model based MBDE (19) (*Mix*MBDE) we followed the Chambers *et al.* (2011) approach and used a pseudo-linearization-based MSE estimator. Finally, for the mixture model based indirect estimators *Mix*EP (18) and *Mix*EBP (20) we used the parametric bootstrap procedure detailed in Section 4.

We used two types of simulations in our empirical evaluations. The first used models to simulate population and sample data. In this case, at each simulation, population data were first generated under the model and a single sample was then taken from this simulated population by stratified simple random sampling without replacement, with the small areas defining the strata. The results from these simulations allow one to compare different estimators in terms of their sensitivity to model assumptions. The second type of simulation was design-based, using population data created by nonparametrically bootstrapping a real survey dataset. Here we evaluated estimators in the context of their performance under repeated sampling from this population under a pre-specified sample design. The results from these simulations allow one to assess the robustness of different estimators to the type of model misspecification seen in practice.

We use two measures of the relative performance for the different small area estimation methods that were considered in our simulations. These are the average percent relative bias

$$AvRB(m) = \underset{i}{mean}\left|\left\{\bar{m}_i^{-1}K^{-1}\sum\nolimits_{k=1}^{K}\left(\hat{m}_{ik} - m_{ik}\right)\right\}\right| \times 100$$

and the average percent relative root mean squared error

$$AvRRMSE(m) = \underset{i}{mean}\left\{\sqrt{K^{-1}\sum\nolimits_{k=1}^{K}\left(\frac{\hat{m}_{ik} - m_{ik}}{m_{ik}}\right)^2}\right\} \times 100$$

of the estimates $\hat{m}_{ik}$ generated by an estimation method. Here $\bar{m}_i = K^{-1}\sum_{k=1}^{K} m_{ik}$, with the subscript *i* indexing the small areas and the subscript *k* indexing the *K* Monte Carlo simulations, and with m_{ik} denoting the actual area *i* mean at simulation *k*, with predicted value $\hat{m}_{ik}$. Note that in the design-based simulations $m_{ik} = m_i$, so $\bar{m}_i = m_i$.

We also investigated the performance of the different MSE estimation methods considered in the simulations. Here we calculated the average relative bias of the MSE estimation method, defined by

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

Here $\hat{M}_{ik}$ denotes the simulation *k* value of the MSE estimator in area *i*, and M_i denotes the actual (i.e. Monte Carlo) MSE in area *i*. We also consider a secondary performance indicator. This is based on the fact that in many applications of small area estimation, MSE estimators are used to calculate Gaussian type confidence intervals for the small area quantities of interest. Consequently it is interesting to evaluate the coverage properties of such intervals. In particular, we focussed on 'two sigma' (i.e. nominal 95 percent) Gaussian intervals, and calculated the average percent coverage

$$AvCR(M) = \underset{i}{mean}\left\{ K^{-1} \sum_{k=1}^{K} I\left(\left| \hat{m}_{ik} - m_{ik} \right| \le 2\hat{M}_{ik}^{1/2} \right) \right\} \times 100 .$$

Table 1. Definitions of small area predictors used in the simulation studies.

Estimator	Description	Method of MSE estimation
	Mixture model based method	
*Mix*EBP	Empirical best predictor (20) defined by the predicted values (17)	Bootstrap MSE (22)
*Mix*EP	Empirical synthetic predictor (18) defined by the predicted values (16)	Bootstrap MSE (22)
*Mix*MBDE	MBDE estimator (19) defined by a 'fitted values' linear model, with the predicted values (16) used as the model covariate	Pseudo-linearization MSE estimator of Chambers *et al.* (2011)
	Raw scale linear mixed model based method	
*Lin*EBLUP	Standard linear mixed model EBLUP	Prasad and Rao (1990) MSE estimator

5.1 Model-based simulations

Model-based simulations are a standard way of illustrating the sensitivity of an estimation procedure to variation in assumptions about the structure of the population of interest. The model-based simulations reported in this paper are based on population data generated under model (1). We choose a population size $N = 15,000$ with $D = 30$ small areas and a sample size $n = 600$ and then randomly generated small area population sizes

N_i, $i = 1,...,D; \sum_i N_i = N$ and sample sizes as $n_i = N_i(n/N); \sum_i n_i = n$. The average small area population and sample sizes were 500 and 20 respectively. These were fixed in all simulations. Population values of y_{ij} $\left(j = 1,...,N_i; i = 1,...,D\right)$ were first generated via the model $\log(y_{ij}) = \log(5) + 0.5\log(x_{ij}) + u_i + e_{ij}$ with unit level random errors e_{ij} independently generated from the normal distribution N(0, $\sigma_e = 0.5$), and random area effects u_i independently generated from the normal distribution N(0, $\sigma_u = 0.3$). The covariate values $\log(x_{ij})$ were generated from the normal distribution N(log(2), $\sigma_x = 3$). We generated zero values for y_{ij} using Poisson sampling, i.e. we set y_{ij} to zero if the realized value of an independently generated uniform variate $U_{ij} \square Uniform(0, P^{-1})$ was such that $U_{ij} > p_{ij}$, where p_{ij} was computed using (10) with the same fixed effect coefficient values as (1) and with an independent area effect drawn from the normal distribution with zero mean and a standard deviation of 0.1. The value of *P* was chosen to generate differing numbers of zero values in the population. Thus with *P* = 0.9, approximately 10% of population values of *Y* are set to zero, while with *P* = 0.5, this increases to 50% and with *P* = 0.3 it becomes 70%. A random sample of (fixed) size $n_i = 20$ was drawn from each area *i*. We also repeated these simulations with a smaller sample of size $n = 300$ and with area sample sizes of $n_i = 10$. All simulations consisted of *K* = 1000 independent replications, with the results from these simulations set out in Table 2.

The percentage average relative bias (*AvRB*) values in Table 2 indicate that *Lin*EBLUP has a significantly larger bias than all three mixture model based small area estimation methods (*Mix*EBP, *Mix*EP and *Mix*MBDE). This implies that *Lin*EBLUP may not be suitable for semicontinuous data. Restricting ourselves to the mixture model based small area estimation methods, we see that the bias values reported for *Mix*EBP are smaller than those reported for *Mix*MBDE and *Mix*EP. Further, the bias advantage of *Mix*EBP appears larger for smaller sample sizes. For moderate sample sizes ($n_i = 20$) the *Mix*MBDE dominates the *Mix*EP in term of bias, but this is not the case for small sample sizes ($n_i = 10$). Average relative biases increase for all the methods as sample sizes

decrease or as the proportion of zero values in the population (i.e. the level of zero inflation in the data) increases. Turning now to the percentage average relative root mean square errors (*AvRRMSE*) values in Table 2, we see again that smaller area sample sizes or larger proportions of population zeros leads to an increase in the percentage average relative root mean square errors of all the methods. Also, *Lin*EBLUP continues to record very large values of relative root mean square error as compared to the mixture model based methods, reinforcing our previous comment that this method of small area estimation appears best avoided when faced with zero inflated skewed data. Among the mixture model based methods, the *Mix*EBP dominates the other methods. Overall, this predictor appears to offer substantial bias and efficiency gains over the other predictors that we considered in our simulations.

Table 2. 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	
	$n_i = 10$	$n_i = 20$	$n_i = 10$	$n_i = 20$	$n_i = 10$	$n_i = 20$	$n_i = 10$	$n_i = 20$
	AvRB							
0.90	0.61	0.50	1.04	1.11	0.94	0.68	27.52	13.06
0.50	1.02	0.75	1.07	1.22	2.41	1.12	30.18	13.95
0.30	2.06	1.84	2.29	2.37	2.59	3.09	94.44	21.97
	AvRRMSE							
0.90	20.25	15.07	33.42	31.03	27.11	18.98	243.74	77.88
0.50	30.53	24.65	38.49	35.61	52.36	36.83	303.73	96.90
0.30	39.53	34.23	44.68	41.67	62.32	53.92	386.60	112.46

We now turn to an examination of the performance of the MSE estimators associated with the different predictors. In particular, we present results from a limited model-based simulation study that was carried out to illustrate the empirical performance of the different MSE estimators defined in Table 1. Here we only considered a sample size $n = 300$ with area specific sample sizes of $n_i = 10$. We also only considered two zero inflation scenarios, corresponding to P = 0.50 and P = 0.90. These simulations were repeated K = 500 times. Note that bootstrap estimation of the MSE in each simulation was based on B = 500 bootstrap samples. The results for these simulations are set out in Table 3 and correspond to 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 per cent Gaussian confidence intervals (*Av*CR) based on the various MSE estimators.

Table 3. Average true RMSEs (*Av*TRMSE), average estimated RMSEs (*Av*ERMSE), average percentage relative bias (*Av*RB), and average percentage coverage rates of nominal 95 per cent Gaussian confidence intervals (*Av*CR) generated by MSE estimators of the different small area estimators defined in Table 1. Area sample sizes are $n_i = 10$. Averages are over the small areas.

P	*Mix*EBP	*Mix*EP	*Mix*MBDE	*Lin*EBLUP
	*Av*CR			
0.90	95	95	95	96
0.50	95	96	96	95
	*Av*ERMSE (*Av*TRMSE)			
0.90	8.39 (8.67)	14.24 (14.50)	11.92 (11.92)	66.47 (57.20)
0.50	7.10 (7.22)	9.22 (8.90)	12.50 (12.40)	29.72 (31.09)
	*Av*RB			
0.90	-2.84	-1.45	0.26	20.31
0.50	-0.61	4.65	2.14	10.61

From the results reported in Table 3, we see that all methods of MSE estimation lead to Gaussian confidence intervals with average actual coverage *Av*CR at or near nominal coverage. Furthermore, the MSE estimators (bootstrap and psuedo linearization) for the three mixture model based predictors (*Mix*EBP, *Mix*EP and *Mix*MBDE) all report average estimated RMSE values that are close to the true average RMSE values. In three out of the four cases of the bootstrap MSE estimator for *Mix*EBP and *Mix*EP we see that on average the estimated RMSE values are a little less than the true RMSE values, indicating a small downward bias. This is reflected in the average percentage relative bias (*Av*RB) values recorded for these cases. In contrast, the pseudo-linearization MSE estimator used with *Mix*MBDE has either virtually no bias or a very small upward bias (again reflected in its *Av*RB values), while the linear model based MSE estimator for *Lin*EBLUP seems somewhat unstable, being conservative when the proportion of zeros in the population is relatively small, but optimistic when this proportion is high. Overall, we can see that the average percentage relative bias (*Av*RB) values recorded by the MSE

estimators for the three mixture model based predictors are all small, in contrast to the bias values recorded by the linear model based MSE estimator for *Lin*EBLUP, which are much larger.

5.2 Design-based simulation

Our design-based simulations were based on actual survey data collected in the 1995-96 Australian Agricultural Grazing Industry Survey (AAGIS) conducted by the Australian Bureau of Agricultural and Resource Economics. The survey collects detailed financial (e.g. farm business receipts, assets, debt), physical (e.g. farm area and location) and socioeconomic information (e.g. age and education of farm operator) from farm businesses across Australia. The target population for the survey is broadacre farms operating in 3 broad agro-ecological zones, the pastoral zone, the wheat-sheep zone and the high rainfall zone. In this study we use the wheat-sheep zone, which consists of 12 regions (the small areas of interest). In the original sample there were 760 farms from 12 regions in the wheat-sheep zone. The variable of interest for this study is number of beef cattle on hand at the end of the financial year (BEEFCL) and the covariate is land area (LAND).

A linear model fit to the sample data was very poor ($R^2 = 0.18$ for the linear regression of BEEFCL on LAND). This fit improved slightly ($R^2 = 0.25$) when dummy variables corresponding to four out of the five broadacre industries: (i) specialist cropping farms, (ii) mixed livestock and cropping farms, (iii) sheep specialists, (iv) beef specialists and (v) mixed sheep and beef farms, were included as covariates of the linear model. It is noteworthy that the target variable BEEFCL is zero inflated with about 38 per cent of its values equal to zero. In particular, out of a total sample of 760 observations there are 286 zero values. The distribution of region sample sizes and proportion of zeros is given in Table 4 and displayed in Figure 1. We used the 474 farms with BEEFCL > 0 and fitted a model for BEEFCL in terms of corresponding values of LAND for these farms. However, we did not observe any improvement in the model fit ($R^2 = 0.18$) even after we included the dummy variables corresponding to industries (i), (iii), (iv) and (v) above ($R^2 = 0.23$).

Table 4. Region specific sample sizes and population sizes

Regions	Population size (N_i)	Sample size (n_i)	Sample size for $y > 0$	Sample size for $y = 0$	Proportion of zeros
1	3726	85	72	13	0.15
2	4770	73	57	16	0.22
3	5918	88	69	19	0.22
4	1776	44	21	23	0.52
5	2335	58	30	28	0.48
6	2929	54	38	16	0.30
7	1901	47	39	8	0.17
8	3731	87	79	8	0.09
9	1450	42	16	26	0.62
10	4090	63	26	37	0.59
11	4960	76	21	55	0.72
12	1983	43	6	37	0.86
Total	39569	760	474	286	0.38

A careful examination of the sample data indicates that the marginal distributions of both BEEFCL and LAND are highly skewed and there is clear evidence of non-linearity in their relationship (see the histograms displayed in Figure 2). When a linear model based on the logarithm of LAND and the four industry dummy variables referred to earlier was fitted to the logarithm of BEEFCL, the fit improved ($R^2 = 0.41$). The usual linear model assumptions of normality, homoscedasticity, etc., were also satisfied. As a consequence it was decided that a log scale linear model was appropriate for positive values of BEEFCL, with the covariates for the fixed part of the model defined by the logarithm of LAND and these four industry dummy variables. Given that the residuals from this model also displayed significant between region variability, a region random effect was included in the model, i.e. we fitted model (1). This improved the R^2 value to just under 50%, with all model coefficients highly significant. Furthermore, when we fitted the mixed logistic model (10) to the binary indicator for BEEFCL > 0 in these data, using the same covariates as in (1), the dummy variables corresponding to industries (i) and (v) and the logarithm of LAND were significant, with some evidence of overdispersion ($\hat{\sigma}_v^2 = 0.87269$, with a standard deviation of 0.93418). Finally, we carried out a crude check of whether the random effects in (1) and (10) might be

correlated by fitting a logistic model to the same binary indicator for BEEFCL > 0 but this time just using the EBLUPs from (1) as the model covariates. The fit of this diagnostic model was significant, with a Generalized R^2 of 14%, indicating potential correlation between the random effects in (1) and the random effect in (10). However, in our simulations we ignored this and proceeded on the basis of a working model defined by a zero correlation between these two sources of variability.

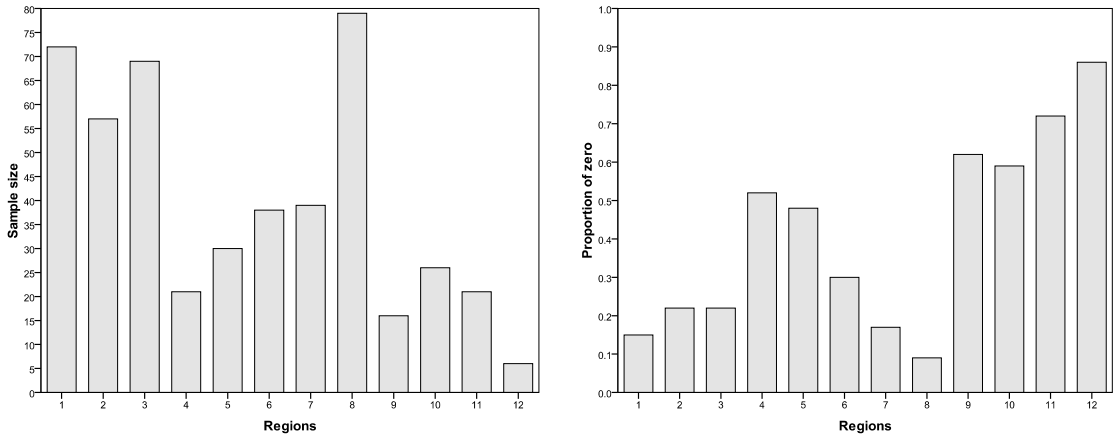

Figure 1. Distribution of regional sample sizes (left side) and regional proportions of zero observations (right side).

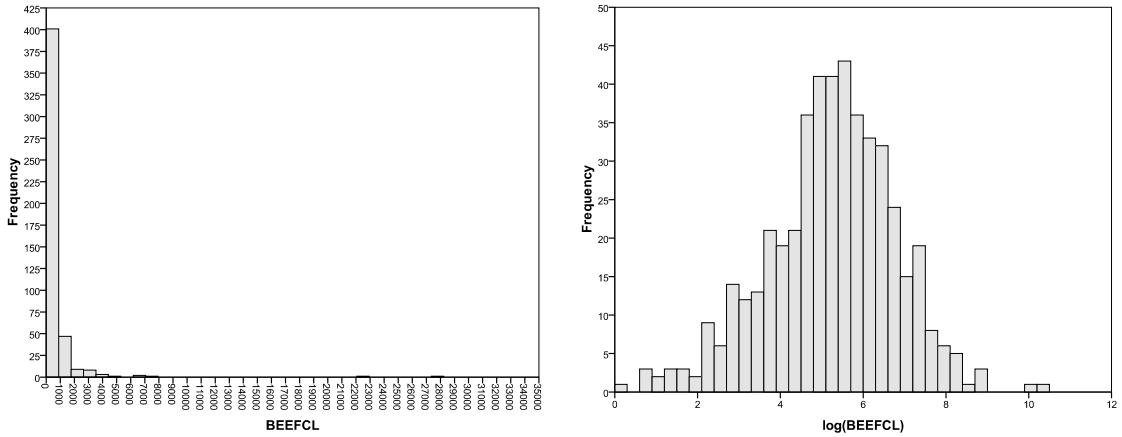

Figure 2. Histogram of BEEFCL (> 0) on raw scale (left plot) and on log scale (right).

We then used these AAGIS sample data to generate a synthetic population of $N = 39{,}569$ farms by re-sampling the original AAGIS sample of $n = 760$ farms with probability proportional to a farm's sample weight. Once created, this fixed population

was repeatedly sampled using stratified random sampling with regions corresponding to strata and with stratum sample sizes the same as in the original sample. Table 5 shows the average over the 12 regions of the percentage relative bias and percentage relative root mean squared error values of the different small area estimation methods based on $K = 1000$ independent stratified samples taken from this synthetic population.

Table 5. Region specific values of the percentage relative biases (RB) and percentage relative root mean squared errors (RRMSE) for different small area predictors.

Regions	*Mix*EBP	*Mix*EP	*Mix*MBDE	*Lin*EBLUP	*Mix*EBP	*Mix*EP	*Mix*MBDE	*Lin*EBLUP
	RB				RRMSE			
1	4.23	47.87	93.87	6.68	11.52	48.07	294.67	16.50
2	17.71	12.24	4.02	33.74	24.39	14.23	31.51	53.07
3	9.31	27.50	25.30	0.64	17.95	27.99	56.46	15.45
4	4.65	103.76	73.10	15.91	21.46	106.19	93.72	36.24
5	31.94	24.69	9.75	8.86	37.56	29.38	68.47	37.32
6	26.64	13.86	1.90	15.18	31.10	15.87	42.77	22.54
7	2.27	52.82	33.43	5.90	15.08	53.12	165.65	23.14
8	33.86	61.03	158.48	153.60	35.87	61.32	200.18	193.61
9	24.05	494.56	74.46	2.46	41.10	497.60	249.22	9.20
10	8.50	123.55	14.82	2.03	21.86	124.83	245.83	21.04
11	12.99	30.87	0.98	16.22	26.76	32.36	45.52	35.46
12	88.13	212.22	68.27	557.25	114.47	229.63	159.76	672.70
Average	22.02	100.41	46.53	68.21	33.26	103.38	137.81	94.69
Median	15.35	50.34	29.37	12.02	25.58	50.59	126.74	29.30

From the results set out in Table 5 we see that the *Mix*EBP predictor has generally smaller average bias and smaller average RRMSE than the other three predictors considered here, while the synthetic type predictor *Mix*EP performs poorly, recording the worst values for RB in 7 out of the 12 regions. This is not unexpected since the log scale linear mixed model underpinning *Mix*EP almost certainly does not hold exactly in the synthetic AAGIS population. Furthermore, since *Mix*EP does not explicity allow for heterogeneity between regions, it is sensitive to bias induced by region to region variability in the relationship between BEEFCL and LAND. On the other hand, even though *Lin*EBLUP is based on a clearly inappropriate model for BEEFCL, its performance as a predictor is reasonable in most cases, reflecting the fact that it includes

a between area adjustment (albeit on the raw scale rather than on the log scale). We also see that although the mixture model based direct estimator *Mix*MBDE has better RB values than *Mix*EP, its RRMSE tends to be large, reflecting the fact that it is a direct estimator. The large relative bias and relative RMSE of *Mix*MBDE and *Lin*EBLUP in region 8 is noteworthy. In this region the proportion of zero values is small, and the positive BEEFCL values highly skewed with many outliers. Here *Lin*EBLUP performs badly because its assumed linear model is a poor fit to these skewed data, while *Mix*MBDE fails because as a direct estimator it is sensitive to the presence of outliers. Overall, it is clear from the results in Table 5 that the mixture model based predictor *Mix*EBP performed better in our design based simulations than its competitors, both in terms of relative bias and relative root mean squared error.

We now consider the design-based performance of the parametric bootstrap procedure used to estimate the MSE of *Mix*EBP in these simulations. Here, for each sample from the fixed synthetic population, the bootstrap MSE estimate was based on $B = 100$ bootstrap samples. The average RMSE values generated by these region-specific bootstrap MSE estimates for *Mix*EBP are shown in Figure 3, as is the corresponding average of the true design-based RMSE for this predictor. We see that the value of the true design-based RMSE for region 8 is very high, while the corresponding bootstrap-based RMSE estimate tends to be low. As noted earlier, this region has highly skewed data, with extreme values persisting even after a logarithmic transformation. This generated large values for the true RMSE of *Mix*EBP. This behaviour was not replicated by the parametric bootstrap, as its bootstrap population data were generated under a distributional assumption that did not allow for such outliers. This raises questions about outlier robust MSE estimation that are beyond the scope of this paper however. Generally, we see that in the remaining regions, where the log scale linear model assumptions for BEEFCL are more appropriate, the bootstrap MSE estimator tracks the actual MSE of *Mix*EBP reasonably well and we are lead to the same conclusions about this MSE estimator as in the model based simulation study presented in Section 5.1.

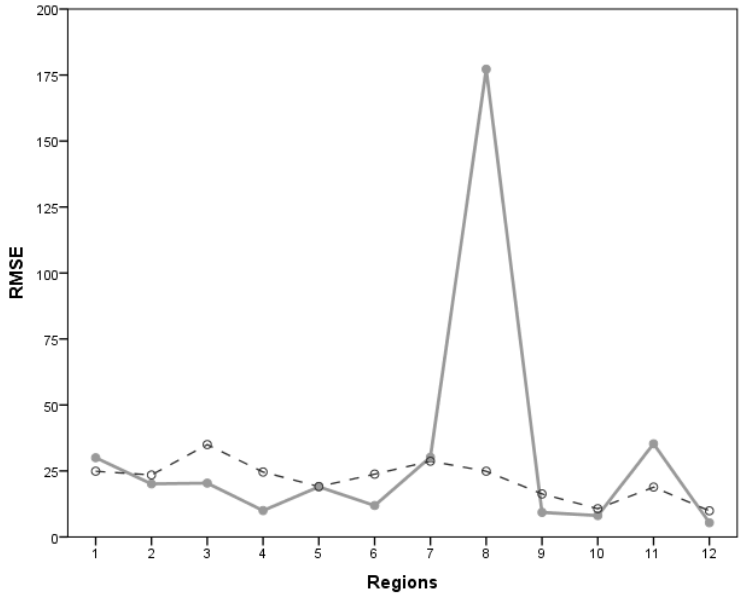

Figure 3. Region-specific values of true design-based RMSE (solid line) and average estimated RMSE (dashed line) for the *Mix*EBP obtained in the design-based simulations using the AAGIS data.

6. Conclusions

In this paper we explore small area estimation for semicontinuous variables, where the data are skewed and contain a substantial proportion of zeros Our approach assumes a mixture or two part random effects model, and we propose an empirical best predictor estimator for small area means for this case. We also propose a parametric bootstrap estimator for its MSE. Empirical results reported in the paper support the conclusion that the proposed mixture model based empirical best predictor (*Mix*EBP) is less biased and can be more efficient than both the corresponding synthetic type predictor (*Mix*EBP) as well as the model based direct type estimator (*Mix*MBDE) based on the 'fitted values' defined by the assumed mixture model. These results also suggest that ignoring the skewed and semicontinuous nature of the data and using a standard mixed linear model-based EBLUP estimator (*Lin*EBLUP) can lead to biased and unstable estimates. We note that, provided the mixture model assumptions are reasonable for the small area data, the proposed parametric bootstrap procedure seems to work well. An application to real agricultural survey data provides some empirical support for these observations.

It should be noted that we assume a log scale linear mixed model for non-zero skewed data. Although the log transformation is widely used in practice for such data, it is not

the only appropriate transformation to linearity, and other transformations (e.g. square root) can be explored in this context. We also assume that zero inflation in the data can be adequately modelled via a mixture of two independent components, a Bernoulli variable and a Lognormal variable. As noted earlier, this is not appropriate if in fact the zero values are essentially due to truncation, and indeed in the AAGIS data that we used in our design-based simulations, there is some evidence that the random area effect in the linear mixed model (1) and the random area effect in the logistic mixed model (10) are correlated. Furthermore, other models for zero inflated skewed data, e.g. those based a generalized linear mixed model with underlying Gamma or Poisson distributions are also possible. We are currently working on these issues.

Acknowledgment

The authors would like to acknowledge the valuable comments and suggestions of the Editor, Associate Editor and two anonymous referees. These led to a considerable improvement in the paper.

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.

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.

Chambers, R., Chandra, H., and Tzavidis, N. (2011).) On bias-robust mean squared error estimation for linear predictors for domains. *Survey Methodology,* **37 (2)**, pp. 1-17.

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.

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.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal Royal Statistical Society, Series B*, **68**, 221–238.

Jiang, J., Lahiri, P. and Wan, S. (2002). A unified Jackknife theory for empirical best prediction with *M*-estimation. *Annals of Statistics*, **30** (6), 1782-1810.

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.

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 and Data Analysis*, **51**(5), 2720-2733.

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.

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

Rao, J.N.K. (2003). *Small Area Estimation*. 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. (available from www.s3ri.soton.ac.uk)