2009

Small Area Estimation of Proportions in Business Surveys

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

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

N. Salvati
University of Pisa

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

15-09

Small Area Estimation of Proportions in Business Surveys

Hukum Chandra, Ray Chambers, Nicola Salvati

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

Small Area Estimation of Proportions in Business Surveys

Hukum Chandra
Indian Agricultural Statistics Research Institute,
Library Avenue, PUSA Campus, New Delhi-110012, India
Phone 0091-11-25841475, Fax 0091-11-25841564
Email: hchandra@iasri.res.in

Ray Chambers
Centre for Statistical and Survey Methodology,
University of Wollongong, Wollongong, NSW, 2522, Australia
Email: ray@uow.edu.au

Nicola Salvati
Dipartimento di Statistica e Matematica Applicata all'Economia
University of Pisa, Via Ridolfi, 10, 56124 - Pisa, Italy
E-mail: salvati@ec.unipi.it

ABSTRACT

Binary data are often of interest in business surveys, particularly when the aim is to characterise grouping in the businesses making up the survey population. When small area estimates are required for such binary data, use of standard estimation methods based on linear mixed models becomes problematic. We explore two model-based techniques of small area estimation for small area proportions, the empirical best predictor (EBP) under a generalized linear mixed model and the model-based direct estimator (MBDE) under a population level linear mixed model. Our empirical results show that both the MBDE and the EBP perform well. The EBP is a computationally intensive method, whereas the MBDE is easy to implement. In case of model misspecification, the MBDE also appears to be more robust. The mean squared error (MSE) estimation of MBDE is simple and straightforward, which is in contrast to complicated MSE estimation for the EBP.

KEY WORDS: Small area proportions, model-based direct estimation, generalised linear mixed model, empirical best predictor.

1. Introduction

The demand of reliable statistics for population characteristics at disaggregated geographical levels (small areas), when only reduced sample sizes are available, has promoted the development of statistical methods for small area estimation (SAE). Conventional estimates for small area quantities based on survey data alone are often unstable because of sample size limitations. From this perspective, model-based methodologies allow for the construction of efficient estimators and their confidence intervals by borrowing the strength through use of a suitable model. Small area models make use of explicit linking models based on random area-specific effects that take into account between areas variation beyond that explained by auxiliary variables included in the model. For continuous response variables, the empirical best linear unbiased predictor (EBLUP) approach under the linear mixed model (LMM) is very common and is known to be efficient for small area estimation, see Rao (2003). Chandra and Chambers (2005, 2009) described the model-based direct estimation (MBDE) method of SAE. The MBDE is a weighted linear estimator for small areas, defined by using sample weights derived under a population level LMM. By construction, the MBDE is a direct estimator and so enjoys the model robustness properties of this class of estimators. It is noteworthy that weights used to define the MBDE 'borrow strength' via a model that explicitly allows for small area effects. Besides ease of implementation, the MBDE is robust under model misspecifications. However, this robustness can be at the price of increased variability.

In this paper we consider the situation where the variable of interest is binary and small area estimates are required. Use of standard estimation methods based on linear mixed models (e.g. the EBLUP) becomes problematic in this case. The empirical best predictor (EBP) under a generalized linear mixed model (GLMM) with logistic link function is often used for SAE based on such data; see Rao (2003) and Saei and Chambers (2003). We observe

that the EBP is model dependent and will be efficient if the model assumptions hold. However, a major difficulty in use of GLMM for SAE is that the likelihood function often involves high dimensional integrals (computed by integrating a product of discrete and normal densities, which has no analytical solution), which are difficult to evaluate numerically. Although computationally attractive alternatives to the likelihood method are available, they can suffer from inconsistency (Jiang, 1998). In context of SAE, mean squared error (MSE) estimation for EBP is an outstanding problem because the analytical form of MSE cannot be calculated explicitly (Manteiga *at el.*, 2007), although an approximate MSE of the EBP can be derived (Saei and Chambers, 2003). An option in this case is to use re-sampling methods, but these are computationally intensive.

An alternative is to ignore the deficiency of the LMM and proceed as if a linear model does hold. This option is relatively simple and cheap to implement. However, it sidesteps the issues that the LMM is incorrect. Given that the MBDE approach has been shown to be model-robust in a number of empirical applications (Chandra and Chambers 2005, 2009 and Chandra *et al.* 2007), it can be expected to produce reasonable results in this case.

This paper explores two model-based techniques of SAE for small area proportions, the empirical best predictor under a GLMM and the model-based direct estimator under a population level LMM. In particular, we examine the application of linear assumption based MBDE to binary data and compare its performance with the EBP via simulation studies using real data sets.

The rest of the paper is organised as follows. The next section introduces the linear mixed model and the generalised linear mixed model, associated estimators for small area proportions and their mean squared error estimators. In the section 3 we then report empirical results and provide a discussion. Finally, section 4 concludes the paper with major findings and further research prospects.

2. Small Area Estimation of Proportions

In this section we introduce the generalised linear mixed model (GLMM) and linear mixed model (LMM). We then describe related estimators for small area quantities based on these models and their mean squared error (MSE) estimation. In particular, we focus on a binary response variable with aim of estimating the population proportions for the variable of interest in small areas and as well as estimates for the MSEs of these estimated proportions.

2.1 The Empirical Best Predictor for the Small Areas

GLMMs are widely used for the development of indirect estimates for small areas when the response data are non-normal. Indirect estimators for small area quantities under GLMMs are often known as empirical best predictors (EBPs).

To start with, let us denote the finite population size by N and assume that it is partitioned into D non-overlapping sub-groups (or small areas), U_i each of sizes N_i with $i = 1,...,D$ such that $N = \sum_{i=1}^{D} N_i$. Let j and i respectively index units within small areas, y_{ij} is the survey variable of interest (typically a binary variable), known for sampled units, $\boldsymbol{x}_{ij}$ is the vector of auxiliary variables (including the intercept), known for the whole population. Let s_i and r_i respectively denotes the sample (of size n_i) and non-sample (of size $N_i - n_i$) in small area i. The objective is to make inference about the small area i population proportions, $p_i = N_i^{-1} \sum_{j \in U_i} y_j = N_i^{-1} \left\{ \sum_{j \in s_i} y_j + \sum_{j \in r_i} y_j \right\}$. Let π_{ij} be the probability that $y_{ij} = 1$. Let u_i denote the random area effect for the small area i, assumed to be normally distributed with mean zero and variance φ. We assume that u_i's are independent and $y_{ij} \mid u_i \sim Bin(1, \pi_{ij})$ with $E(y_{ij} \mid u_i) = \mu_{ij} = \pi_{ij}$ and $Var(y_{ij} \mid u_i) = \sigma_{ij} = \pi_{ij}(1 - \pi_{ij})$. A popular

model for this type of data is the GLMM with logistic link function, also referred as the linear logistic mixed model (LLMM), given by

$$logit(\pi_{ij}) = \log\left\{\pi_{ij}/(1-\pi_{ij})\right\} = \eta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i,\ j = 1,....,N_i; i = 1,...,D \qquad (1)$$

where $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of regression parameters.

In the small area estimation literature, it is common practice to express the model (1) at the population level as follows (Rao, 2003, Chapter 6). Let y_U be the $N \times 1$ vector of response variable with elements y_{ij} $(j = 1,....,N_i; i = 1,...,D)$, $\mathbf{X}_U$ be the $N \times p$ known design matrix with rows $\boldsymbol{x}_{ij}$, $\mathbf{G}_U = diag(\mathbf{1}_{N_i}; 1 \le i \le D)$ is the known matrix of order $N \times D$, $\mathbf{1}_{N_i}$ is a column vector of ones of size N_i, $\mathbf{u} = (u_1,...,u_D)'$ and $\boldsymbol{\eta}_U$ denotes the $N \times 1$ vector of linear predictors η_{ij} given by (1). We define $\boldsymbol{\mu} = E(\mathrm{y}_U \mid \mathbf{u})$ the conditional mean function of the response vector y_U given $\mathbf{u}$ with elements μ_{ij} and $Var(\mathrm{y}_U \mid \mathbf{u}) = diag\{\sigma_{ij}\}$ the conditional covariance matrix. Let $g(\cdot)$ be a monotonic link function (McCullagh and Nelder, 1989, page 27), such that $g(\boldsymbol{\mu})$ can be expressed in terms of a linear model of form

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}_U = \mathbf{X}_U\boldsymbol{\beta} + \mathrm{G}_U\mathbf{u}\,. \qquad (2)$$

The equation (2) then defines a GLMM if $\mathbf{y}_U$ given $\boldsymbol{\mu}$ are independent and belong to the exponential family of distributions. The vector of random area effects $\mathbf{u}$ has mean $\mathbf{0}$ and variance $\boldsymbol{\Omega}(\boldsymbol{\delta}) = \varphi\mathbf{I}_D$, where $\mathbf{I}_D$ is the identity matrix of order *D*. For a binary response, the link function $g(\cdot)$ is typically a *logit* function, see (1). The relationship between $\mathbf{y}_U$ and $\boldsymbol{\eta}_U$ is therefore represented through a known function $h(\cdot)$, defined by $E(\mathbf{y}_U \mid \mathbf{u}) = h(\boldsymbol{\eta}_U)$.

Suppose that our interest is in predicting the vector of linear parameters for small areas $\boldsymbol{\theta} = \mathbf{a}_U\mathbf{y}_U$, where $\mathbf{a}_U = diag\{\mathbf{a}_i', i = 1,..,D\}$ is a $D \times N$ matrix and $\mathbf{a}_i' = (a_{i1},...,a_{iN_i})$ is a vector of known elements. In particular, for estimation of a population proportion p_i for small area

i, $\mathbf{a}_i'$ denotes the population vector with value N_i^{-1} for each population unit in area *i* and zero elsewhere. Without loss of generality, we arrange the vector $\mathbf{y}_U$ so that its first *n* elements correspond to the sampled units, and then partition $\mathbf{a}_U, \mathbf{y}_U, \boldsymbol{\eta}_U, \mathbf{X}_U$ and G_U according to sample and non-sample units as

$$\mathbf{a}_U = \begin{bmatrix} \mathbf{a}_s \\ \mathbf{a}_r \end{bmatrix}, \ \mathbf{y}_U = \begin{bmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{bmatrix}, \ \boldsymbol{\eta}_U = \begin{bmatrix} \boldsymbol{\eta}_s \\ \boldsymbol{\eta}_r \end{bmatrix}, \ \mathbf{X}_U = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \text{ and } \mathbf{G}_U = \begin{bmatrix} \mathbf{G}_s \\ \mathbf{G}_r \end{bmatrix}.$$

Here a subscript of *s* denotes components defined by the *n* sample units while a subscript of *r* is used to denote components defined by the remaining $N - n$ non-sample units. We then write $E(\mathbf{y}_s \mid \mathbf{u}) = h(\boldsymbol{\eta}_s)$ and $E(\mathbf{y}_r \mid \mathbf{u}) = h(\boldsymbol{\eta}_r)$. Typically, $h(\cdot)$ is obtained as $g^{-1}(\cdot)$. The parameter of interest $\boldsymbol{\theta} = \mathbf{a}_U \mathbf{y}_U$ can be expressed as

$$\boldsymbol{\theta} = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r \mathbf{y}_r = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r \boldsymbol{\beta} + \mathrm{G}_r \mathbf{u}). \tag{3}$$

The vector $\mathbf{y}_s$ of sample values is known, whereas the second term in the right hand side of (3), which depends on the non-samples values $\mathbf{y}_r = h(\mathbf{X}_r \boldsymbol{\beta} + \mathrm{G}_r \mathbf{u})$, is unknown and can be predicted by fitting the model (3) to the sample data. In this paper $\mathbf{y}_s = \{y_{sij}\}$ denotes the vector of sample values of the binary survey variable *y*, e.g. $y = 1$ if the consumption expenditure per household is less than a poverty line, 0 otherwise. Similarly, $\mathbf{y}_r = \{y_{rij}\}$ represents the vector of non-samples values of the survey variable. The parameter of interest p_i for each small area can then be obtained by predicting each element of $\{y_{rij}\}$.

For known $\boldsymbol{\Omega}(\boldsymbol{\delta})$, the values of $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated from the sample data by Penalized Quasi-Likelihood (PQL) under model (3) (Breslow and Clayton, 1993). This gives the best linear unbiased estimate (BLUE) for $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) for $\mathbf{u}$. Using (3) we then obtain the BLUP-type estimator of $\boldsymbol{\theta}$. In practice $\boldsymbol{\Omega}(\boldsymbol{\delta})$ is unknown and the vector of variance components $\boldsymbol{\delta}$ is estimated from the sample data. Using

the estimated value $\hat{\boldsymbol{\delta}}$ of the $\boldsymbol{\delta}$ leads to the empirical BLUE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and the empirical BLUP $\hat{\mathbf{u}}$ for $\mathbf{u}$ and thus the empirical BLUP type estimator of $\boldsymbol{\theta}$, which is given by

$$\hat{\boldsymbol{\theta}} = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathrm{G}_r \hat{\mathbf{u}}) . \tag{4}$$

As mentioned in the previous section, fitting a GLMM involves evaluating a likelihood function that does not have close form analytical expression. Several approximations to this likelihood function and approximate maximum likelihood estimators have been proposed in the literature. In particular, the PQL approach is a popular estimation procedure for the GLMM that is based on a linear approximation to the non-normal response variable, which is then assumed to have an approximately normal distribution. This approach is reliably convergent but tends to underestimate variance components as well as fixed effect coefficients (Breslow and Clayton, 1993). McGilchrist (1994) introduced the idea of using BLUP to obtain approximate restricted maximum likelihood (REML) estimates for GLMMs. This link between BLUP and REML is described in Harville (1977) for the normal case. Saei and Chambers (2003) described an iterative procedure to obtained Maximum Penalized Quasi-Likelihood (MPQL) estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ for given $\boldsymbol{\Omega}$. At convergence, the MPQL estimate of $\boldsymbol{\theta}$ is obtained by substituting the converged values of $\boldsymbol{\beta}$ and $\mathbf{u}$. However, in practice the variance components parameters defining the matrix $\boldsymbol{\Omega}$ are unknown and have to be estimated from sample data. The MPQL estimates of these variance components are biased and so this approach is not recommended in the practice. Alternative estimates based on ML and REML can be defined. In particular, the bias in the REML estimates is typically small. An iterative procedure that combines the MPQL estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with REML estimation of $\boldsymbol{\Omega}$ is described in Saei and Chambers (2003). In the empirical results reported in section 3, we adopted this algorithm for parameter estimation.

Turning now to estimation of mean squared error of the EBLUP-type predictor (4) we put $\mathbf{H}_r = \mathbf{H}(\hat{\eta}_r) = \partial h(\eta_r)/\partial \eta_r\big|_{\eta_r=\hat{\eta}_r}$ and $\hat{\boldsymbol{B}}_s = \partial^2 l_1/\partial \eta_s \partial \eta_s'\big|_{\eta_s=\hat{\eta}_s}$, the matrix of second derivatives of l_1 (the log-likelihood function l_1 defined by the vector $\mathbf{y}_s$ given **u**) with respect to η_s at $\eta_s = \hat{\eta}_s$. Similarly, we put $\hat{\boldsymbol{B}}_r = \partial^2 l_1/\partial \eta_r \partial \eta_r'\big|_{\eta_r=\hat{\eta}_r}$. We write $\boldsymbol{X}_r^* = \mathbf{a}_r \mathbf{H}_r \boldsymbol{X}_r$ and $\boldsymbol{G}_r^* = \mathbf{a}_r \mathbf{H}_r \boldsymbol{G}_r$. An approximate estimate of the mean squared error for the EBLUP-type estimator (4) (see Saei and Chambers, 2003; Manteiga *et al.*, 2007) is then

$$mse(\hat{\boldsymbol{\theta}}) = m_1(\hat{\boldsymbol{\delta}}) + m_2(\hat{\boldsymbol{\delta}}) + 2m_3(\hat{\boldsymbol{\delta}}) + m_4(\hat{\boldsymbol{\delta}}) \tag{5}$$

where

$m_1(\hat{\boldsymbol{\delta}}) = \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s \boldsymbol{G}_r^{*\prime}$ with $\hat{\boldsymbol{T}}_s = (\hat{\boldsymbol{\Omega}}^{-1} + \boldsymbol{G}_s' \hat{\boldsymbol{B}}_s \boldsymbol{G}_s)^{-1}$,

$m_2(\hat{\boldsymbol{\delta}}) = \mathbf{C}_r \left(\boldsymbol{X}_s' \hat{\mathbf{B}}_s \boldsymbol{X}_s - \boldsymbol{X}_s' \hat{\mathbf{B}}_s \boldsymbol{G}_s \hat{\boldsymbol{T}}_s \boldsymbol{G}_s' \hat{\mathbf{B}}_s \boldsymbol{X}_s \right)^{-1} \mathbf{C}_r'$, with $\mathbf{C}_r = \left\{ \boldsymbol{X}_r^* - \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s \boldsymbol{G}_s' \hat{\mathbf{B}}_s \boldsymbol{X}_s \right\}$,

$m_3(\hat{\boldsymbol{\delta}}) = \left\{ tr\left((\hat{\nabla}_t \hat{\Sigma}_s \hat{\nabla}_k') v(\hat{\boldsymbol{\delta}}) \right) \right\}$, with $\hat{\Sigma}_s = \boldsymbol{G}_s' \hat{\mathbf{B}}_s \mathbf{G}_s + \phi \boldsymbol{G}_s' \hat{\mathbf{B}}_s \mathbf{G}_s \boldsymbol{G}_s' \hat{\mathbf{B}}_s \mathbf{G}_s$, and

$m_4(\hat{\boldsymbol{\delta}}) = \mathbf{a}_r \hat{\mathbf{B}}_r \mathbf{a}_r'$.

Let $\varsigma = \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s$ where $\boldsymbol{G}_{rt}^*$ is the t^{th} row of the matrix $\boldsymbol{G}_r^*$, then $\hat{\nabla}_t = \partial(\varsigma_t)/\partial\delta\big|_{\delta=\hat{\delta}} = \hat{\phi}^{-2} \boldsymbol{G}_{rt}^* \hat{\boldsymbol{T}}_s \hat{\boldsymbol{T}}_s$. Here $v(\hat{\boldsymbol{\delta}})$ is the asymptotic covariance matrix of estimates of variance components $\hat{\boldsymbol{\delta}}$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\boldsymbol{\delta}}$. This depends upon whether we are using ML or REML to estimate $\hat{\boldsymbol{\delta}}$. In this paper we used REML estimates for $\hat{\boldsymbol{\delta}}$. See Saei and Chambers (2003) for these expressions for both ML and REML estimates for $\hat{\boldsymbol{\delta}}$. Using (4) the empirical best predictor (EBP) for the small area *i* proportion p_i is then

$$\hat{p}_i^{EBP} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\mu}_{ij} \right\} \tag{6}$$

where $\hat{\mu}_{ij} = \exp(\hat{\eta}_{ij})\{1+\exp(\hat{\eta}_{ij})\}^{-1} = \hat{\pi}_{ij}$ and $\hat{\eta}_{ij} = \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{u}_i$. Similarly, replacing $\mathbf{a}'_i$ as above, we obtain the MSE estimator of (6) from (5).

2.2 The MBDE for Small Area Proportion

The model-based direct estimation (MBDE) approach to SAE investigated in Chandra and Chambers (2005, 2009) is effectively a linear estimation methodology and implicitly assumes that the variable of interest follows a LMM. Following the notation of Chandra *et al.* (2007), a brief description of MBDE is as follows. Suppose that the population values follow the linear mixed model

$$\mathbf{y}_U = \mathbf{X}_U\boldsymbol{\beta} + \boldsymbol{G}_U\mathbf{u} + \mathbf{e}_U \tag{7}$$

where $\mathbf{y}_U = (\mathbf{y}'_1,.....,\mathbf{y}'_D)'$, $\mathbf{X}_U = (\mathbf{X}'_1,.....,\mathbf{X}'_D)'$, $\boldsymbol{G}_U = diag(\boldsymbol{G}_i = \mathbf{1}_{N_i}; 1 \le i \le D)$, $\mathbf{u} = (u_1,...,u_D)'$ and $\mathbf{e}_U = (\mathbf{e}_1,...,\mathbf{e}_D)'$ denote partitioning into area components. The independence between small areas indicates the covariance matrix of $\mathbf{y}_U$ has block diagonal structure, $\mathbf{V}_U = diag(\mathbf{V}_i; 1 \le i \le D)$ with $\mathbf{V}_i = \varphi\mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2\mathbf{I}_{N_i}$. In practice the variance components that define $\mathbf{V}_U$ are unknown and can be estimated from the sample data using methods described, for example, in Harville (1977). We denote these estimates by $\hat{\boldsymbol{\delta}} = (\hat{\varphi}, \hat{\sigma}_e^2)'$ and put a 'hat' on any quantity where these estimates are substituted for actual values, e.g. $\hat{\mathbf{V}}_U = diag(\hat{\mathbf{V}}_i; 1 \le i \le D)$ and $\hat{\mathbf{V}}_i = \hat{\varphi}\mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \hat{\sigma}_e^2\mathbf{I}_{N_i}$. As with (2) we again consider the decomposition of different terms into sample and non-sample components and, from Royall (1976), we note that the sample weights that define the EBLUP for the population total of *y* under the population level linear mixed model (7) are then

$$\mathbf{w}_s^{EBLUP} = (w_j^{EBLUP}) = \mathbf{1}_s + \hat{\mathbf{H}}'\left(\mathbf{X}_U\mathbf{1}_N - \mathbf{X}'_s\mathbf{1}_s\right) + \left(\mathbf{I}_s - \hat{\mathbf{H}}'\mathbf{X}'_s\right)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \tag{8}$$

where $\hat{\mathbf{H}} = \left(\sum_i \mathbf{X}_{is}' \hat{\mathbf{V}}_{iss}^{-1} \mathbf{X}_{is}\right)^{-1} \left(\sum_i \mathbf{X}_{is}' \hat{\mathbf{V}}_{iss}^{-1}\right)$. The model-based direct estimator (MBDE) of proportion for small area *i* is then defined as

$$\hat{p}_i^{MBDE} = \sum_{j \in s} w_{ij}^{MBDE} y_j = (w_{is}^{MBDE})' y_s \tag{9}$$

where

$$w_{ij}^{MBDE} = \frac{I(j \in s_i) w_j^{EBLUP}}{\sum_{k \in s} I(k \in s_i) w_k^{EBLUP}}.$$

Here $I(j \in s_i)$ is the indicator function for unit *j* to be in the area *i* sample, and $w_s^{EBLUP} = (w_j^{EBLUP})$ is the vector of weights given by (8). A robust estimator of the MSE of the MBDE (9) (Chandra and Chambers, 2009; Royall and Cumberland, 1978) is

$$mse(\hat{p}_i^{MBDE}) = \hat{V}ar(\hat{p}_i^{MBDE}) + \left\{\hat{B}ias(\hat{p}_i^{MBDE})\right\}^2. \tag{10}$$

The first term on right hand side of (10) is the estimate of prediction variance of the MBDE (9), given by $\hat{V}ar(\hat{p}_i^{MBDE}) = N_i^{-2} \sum_{j \in s} \left\{a_{ij}^2 + (N_i - n_i) n^{-1}\right\} \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2$ with $a_{ij} = N_i w_{ij}^{MBDE} - I(j \in i)$, where $I(t)$ is the indicator function for condition *t*, and $j \in i$ corresponds to unit *j* coming from small area *i* and $\hat{\mu}_j$ is an unbiased linear estimator of the conditional expected value of y_j under (7), i.e. of $\mu_j = \mathbf{x}_j \beta + \boldsymbol{G}_j u_i; j \in s_i$. Under (7), $\hat{\mu}_j = \mathbf{x}_j \hat{\beta} + \boldsymbol{G}_j \hat{u}_i; j \in s_i$. Here $\hat{\lambda}_j$ is given by $\hat{\lambda}_j = 1 - 2\phi_{jj} + \sum_{k \in s} \phi_{kj}^2$, where the constants ϕ_{kj} are obtained from writing $\hat{\mu}_j$ in the form $\hat{\mu}_j = \sum_{k \in s} \phi_{kj} y_k$. The second term on the right hand side of (10), which is estimate of prediction bias of the MBDE (9), is $\hat{B}ias(\hat{p}_i^{MBDE}) = \sum_{j \in s} w_{ij}^{MBDE} \hat{\mu}_j - N_i^{-1} \sum_{j \in i} \hat{\mu}_j$. The MSE estimator (10) is called a *robust* model-based estimator because it does not depend on second order moments assumptions and is thus robust to misspecification of the second order moments of the working model. A more

detailed discussion of this approach to mean squared error estimation is set out in Chambers et al. (2007).

3. Empirical Evaluations

In this section we present simulation studies that illustrate the performance of the empirical best predictor (6) under the GLMM (2), denoted by EBP below, and the MBDE estimator (9) under the LMM (7), denoted by MBDE below.

3.1 Data Sets

We carried out design-based simulation studies using three real data sets. These data are from different types of surveys (agricultural, environmental and consumer expenditure), and allow us to evaluate the performance of these methods in the context of real populations and realistic sampling methods. The three data sets used in the simulations are as follows:

i) *The Australian Agricultural and Grazing Industries Survey (AAGIS) Data*. This is based on data collected from a sample of 1652 Australian broadacre farms spread across 29 regions of Australia. These regions are the small areas of interest. A population of N = 81,982 farms was generated by bootstrapping the original AAGIS sample. That is, the 1,652 farms in the original AAGIS sample were themselves sampled with replacement N times using selection probabilities proportional to a farm's AAGIS sample weight, where the sum of AAGIS sample weights is 81,982. Independent samples of n = 1,652 farms were then taken from this population using stratified random sampling, with regions are strata and with stratum sample allocations the same as in the original AAGIS sample. The y-variable of interest was a binary (0-1) variable, ZeroDebt, which takes the value 1 if farm debt is zero for the given farm and value zero otherwise. The total area of the farm in

hectares is used as the model covariate (x), and the target is estimation of the proportion of ZeroDebt farms in each region.

ii) *The Environmental Monitoring and Assessment Program (EMAP) Data*. This data set is based on data provided by Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University. It consists of a sample of 349 lakes in the North-Eastern states of the United States, grouped by 6-digit Hydrologic Unit Codes (HUC). The HUCs are the regions of interest. There were sample sizes equal to one in three of these HUCs, so these regions were combined with neighbouring regions. This resulted in 23 small areas, with sample sizes that varied from 2 to 45. We generated a population of size N = 21,028 by sampling N times with replacement from the above sample data and with probability proportional to a unit's sample weight; and then selected 1000 independently stratified random samples of the same size as the original sample from this (*fixed*) simulated population. HUC sample sizes were also fixed to be the same as in the original sample. The variable of interest y in this case takes value 1 if Acid Neutralizing Capacity (ANC) - an indicator of the acidification risk of water bodies - is less than 500 and 0 otherwise. The elevation of the lake is the auxiliary variable. We are interested in estimating the proportion of lakes in each HUC with ANC less than 500.

iii) *Albanian Living Standards Measurement Study (LSMS) Data.* These data are from a sample of 3591 households spread across 36 districts of Albania that participated in the World Bank's Living Standards Measurement Study conducted in 2002 in Albania. The survey provides information on a variety of issues related to the living conditions of the people in Albania, including details on income and non-income dimensions of poverty in the country, and forms the basis of poverty assessment in this country. We generated a population of N = 724,782 households by sampling N times with replacement from the above sample of 3,591 households and with probability proportional to a household's

sample weight. The simulation was then based on selecting 1000 independently stratified random samples of the same size as the original sample from this simulated population (fixed). District sample sizes were also fixed to be the same as in the original sample, varying from a low of 8 to a high of 688. The variable of interest *y* takes value 1 if equivalent income of household is below median income and is 0 otherwise. Our aim is to estimate the proportion of households below median equivalent income at District level, using the ownership of land, which is a strong indicator of poverty, and the presence of facilities in the dwelling (television and parabolic dish antenna) as covariates. Unlike the first and second data sets these covariates are binary.

3.2 Performance Measures

The performance of different small area estimators were evaluated with respect to three basic criteria: the relative bias (RB) and the relative root mean squared error (RRMSE), both expressed as percentages, of estimates of the small area proportions and the coverage rate of nominal 95 per cent confidence intervals for these proportions. In the evaluation of coverage performances, intervals are defined by the estimate of small area proportion plus or minus twice their standard error.

The relative bias was measured by *%AvRB* and *%MedRB*, where

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

with *%MedRB* defined similarly, but with the mean over the small areas replaced by the median. The root mean squared error was measured by *%AvRRMSE* and *%MedRRMSE*, where

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

with *%MedRRMSE* differing from *%AvRRMSE* only by the use of median rather than mean when averaging over the small areas. Coverage performance for prediction intervals was measured by *%AvCR* and *%MedCR*, where

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

and again *%MedCR* differs from *%AvCR* only by the use of median rather than mean when averaging over the small areas. Note that the subscript of *k* here indexes the *K* simulations, with m_{ik} denoting the value of the small area *i* mean in simulation *k* (this is a fixed population value in the design-based simulations considered here), and $\hat{m}_{ik}$, $\hat{M}_{ik}$ denoting the area *i* estimated value and corresponding estimated MSE in simulation *k*. The actual area *i* mean value (the average over the simulations) is denoted $M_i = K^{-1}\sum_{k=1}^{K} m_{ik}$.

3.3 Result and Discussion

In Table 1 we report the average (*AvRB*) and median (*MedRB*) relative bias, average (*AvRRMSE*) and median (*MedRRMSE*) relative root mean squared error and average (*AvCR*) and median (*MedCR*) coverage rate for nominal 95% intervals of the small area proportions generated by two small area estimation methods (EBP and MBDE) based on repeated sampling from the simulated AAGIS, EMAP and Albanian populations. All averages (and medians) are expressed as percentages and are over the small areas of interest. For the EMAP population the true small area proportions for regions 5 and 9 are zero. Consequently, average (and median) results for EMAP data in Table 1 are based on the remaining 21 areas. The region-specific performance measures for the AAGIS, EMAP and Albanian data are shown in Figures 1-3 respectively.

The results in Table 1 show that the average (and median) relative bias of MBDE is smaller than that of EBP. The region-specific relative biases given in Figures 1-3 also show

that MBDE has consistently better bias behaviour than EBP. In particular, EBP is badly biased in some regions, e.g. region 6 and 10 for the AAGIS data (Figure 1), region 1, 3 and 13 for the EMAP data (Figure 2) and region 1, 3 and 14 for the Albanian data (Figure 3). Overall in term of relative bias, MBDE appears to dominate EBP for these populations.

In contrast, the two methods are comparable in terms of relative root mean squared error (i.e. efficiency), with neither approach dominating the other. However, in many areas MBDE approach seems preferable, e.g. region 1 and 6 for AAGIS data (Figure 1). In two regions (1 and 6) where EBP fails, inspection of the population and sample data indicated that this is because of a few outlying estimates. Similarly, in Figure 2 the unstable performance of the EBP in regions 3 and 6 is noteworthy. These unstable results are due mainly to the fact that there is little or no variability in the data in these two regions. In contrast, the MBDE method appears unaffected by such behaviour. Further, in these cases the EBP produces overestimates for the small area proportions.

The MBDE has marginally better coverage performance for the AAGIS and the Albania data, while both methods show overcoverage for the EMAP data. In Figure 2 we observe overcoverage in a number of regions. This is because the MSE for the MBDE is being significantly overestimated. This is particularly puzzling for regions 1-6, 9, 16 and 17. A critical examination of results revealed that in these regions true small area population proportion is either 1 (regions 1-4, 6, 16 and 17) or 0 (regions 5 and 9). In these regions the estimated area proportions via MBDE are same as true values so true MSEs are zero. However, the estimates of these MSEs are not zero. This leads to overestimated MBDE mean squared errors. Although the true MSE is not exactly zero for the EBP method in these cases (since it is an indirect estimator), similar problems exist with its MSE estimator in such regions.

These empirical results clearly show that MBDE performs well when applied to binary data. In contrast, under the true model, EBP is expected to be more efficient than the MBDE. However, in practice, the true model in always unknown and so we deal with working models. In this case MBDE can be expected to perform reasonably well. In particular, our results indicate that under a misspecified model (e.g. data with less variability) the MBDE approach provides more robust small area estimates that are easy to implement. In contrast, the EBP is a computationally intensive method based on approximations that seems less robust.

4. Concluding Remarks

We have investigated two model-based methods of small area estimation for small area proportions, the empirical best predictor (EBP) under a generalized linear mixed model and the model-based direct estimator (MBDE) under a population level linear mixed model. In particular, we examine an application of linear assumption based MBDE to binary data. The empirical evaluations based on three real data from different types of survey (agricultural, environmental and consumer and expenditure) show that both MBDE and EBP methods perform well. No efficiency loss was observed in MBDE due to linear assumption. The EBP is a computationally intensive method, whereas the MBDE is easy to implement. In case of model misspecification, the MBDE also appears to be more robust. In addition, MSE estimation of MBDE is simple and straightforward, which is in contrast to complicated MSE estimation for the EBP.

Our results also indicate that there is a need for research to be carried out on a suitable methodology for small area estimation of proportions when the area sample is all either 1 or 0. There is some theory (see Jovanovic and Levy, 1997) that attempts to address this problem. However, this needs to be explored in the context of small area estimation.

References

Breslow, N.E. and Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Model. *Journal of the American Statistical Association*, **88**, pp. 9-25.

Chambers, R., Chandra, H. and Tzavidis, N. (2007). On Bias-Robust Mean Squared Error Estimation for Linear Predictors for Domains. Working Papers, 09-08, The University of Wollongong, Australia. (Available from: http://cssm.uow.edu.au/publications).

Chandra, H. and Chambers, R. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation. *Statistics in Transition*, **7**, 637-648.

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

Chandra, H. and Chambers, R. (2009). Multipurpose Weighting for Small Area Estimation. *Journal of Official Statistics*, to appear.

Harville, D.A. (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, **72**, pp. 320–338.

Jiang, J. (1998) Consistent Estimators in Generalized Linear Mixed Models, *Journal of the American Statistical Association* **93**, pp. 720-729.

Jovanovic, B.D. and Levy**,** P.S. **(**1997**)**. A look at the Rule of Three. *American Statistician,* **51 (2)**, pp 137-139.

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

McGilchrst, C.A. (1994). Estimation in Generalized Mixed Models, *Journal of the Royal Statistical Society Series B*, **56**, pp. 61-69.

Manteiga, G. W., Lombardìa, M.J., Molina, I., Morales, D. and Santamarìa, L. (2007) Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters

under a Logistic Mixed Model, *Computational Statistics & Data Analysis*, **51**, pp. 2720-2733.

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. *Journal of the American Statistical Association*, **71**, 657-664.

Royall, R.M. and Cumberland, W.G. (1978) Variance Estimation in Finite Population Sampling, *Journal of the American Statistical Association*, **73**, pp. 351-358.

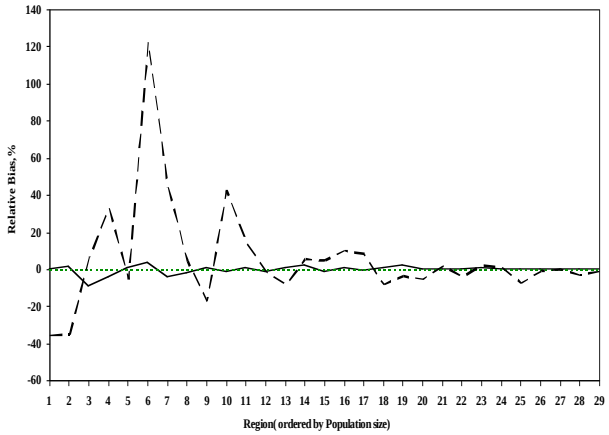
Saei, A. and Chambers, R. (2003) Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects, *Methodology Working Paper- M03/15*, University of Southampton, United Kingdom.

Table 1. Average (*AvRB*) and median (*MedRB*) relative bias, average (*AvRRMSE*) and median (*MedRRMSE*) relative RMSE and average (*AvCR*) and median (*MedCR*) coverage rate generated by EBP and MBDE. All averages are expressed as percentages and are over the small areas of interest.

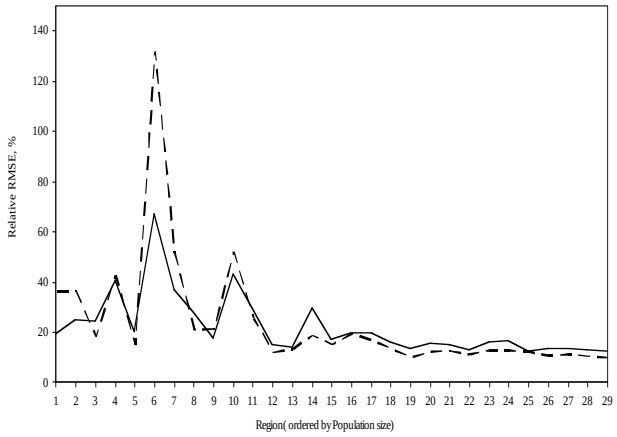
Criterion	AAGIS		EMAP		Albania	
	EBP	MBDE	EBP	MBDE	EBP	MBDE
AvRB	6.13	-0.32	1.22	-0.25	1.02	-0.03
MedRB	0.46	0.24	-0.35	0.00	0.08	-0.04
AvRRMSE	23.89	21.76	17.50	18.05	11.05	12.64
MedRMSE	15.01	17.06	8.43	7.92	10.23	11.23
AvCR	88	93	96	98	93	94
MedCR	96	94	97	99	95	95

Figure 1. Regional performance of EBP (dashed line) and MBDE (solid line) for the AAGIS data.

Relative Bias (%)



Relative RMSE (%)



Coverage Rate

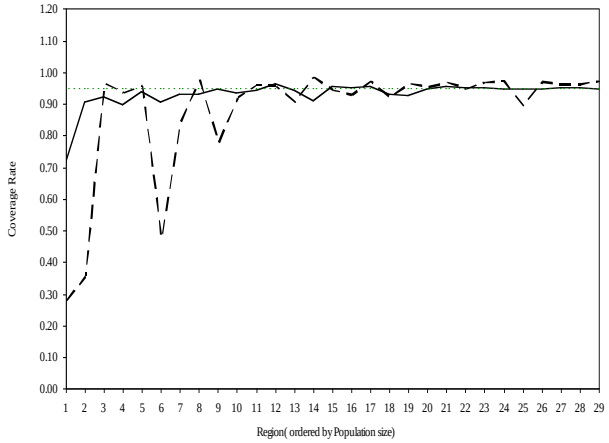
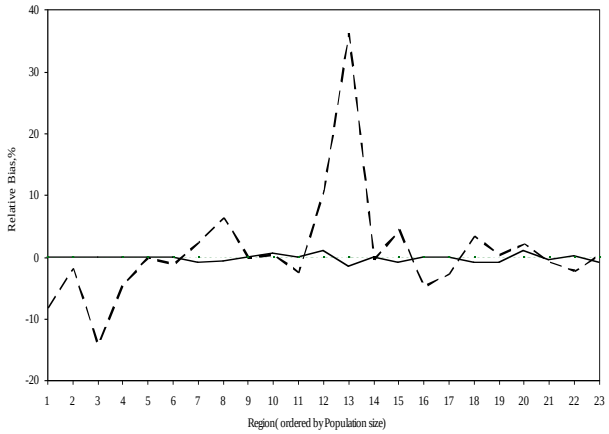
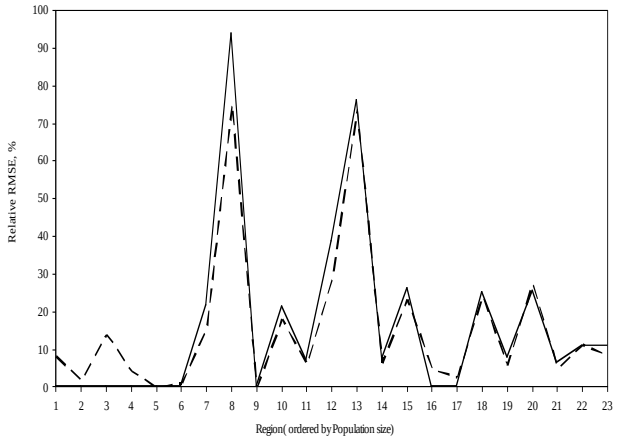


Figure 2. Regional performance of EBP (dashed line) and MBDE (solid line) for the EMAP data.

Relative Bias (%)



Relative RMSE (%)



Coverage Rate

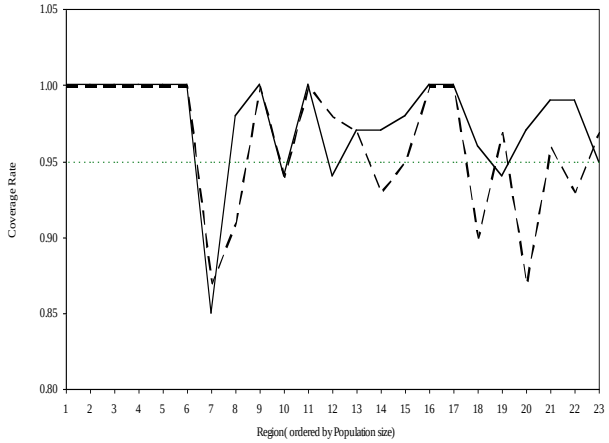


Figure 3. Regional performance of EBP (dashed line) and MBDE (solid line) for the Albania data.