Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(3) 2010 367-374

Small Area Estimation with Binary Variables

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi

Received 02 June 2008; Revised 08 September 2009; Accepted 25 March 2010

SUMMARY

Estimates of proportion for small areas are often required for many policy planning and resource allocation. This paper examines an application of linear assumption based model-based direct (MBD) approach (Chandra and Chambers 2005, 2009) of small area estimation (SAE) to estimate small area proportions. The empirical performance of the MBD approach is compared with indirect method of SAE, the empirical best predictor (EBP) under a generalised linear mixed model (Rao 2003, chapter 5 and Saei and Chambers 2003). The empirical results based on two real data show that both the MBD and EBP perform well. The EBP is a computation intensive method, in contrast, MBD is easy to implement. In case of model misspecifications, the MBD appears to be more robust.

Keywords: Binary variable, Small area estimation, MBD approach, GLMM, Empirical best predictor.

1. INTRODUCTION

The linear mixed models (LMMs) are widely used in small area estimation (SAE). See Rao (2003). These models assume that the relationship between the mean of the variable of interest Y and the fixed and the random effects can be modelled as a linear function, the variance is not a function of mean, and that the random effects follow a normal distribution. However, any or all these assumptions may be questionable for certain data. For example, for binary data, assumption of linear relationship, normality and constant variance is questionable. A generalised linear mixed model (GLMM) gives extra flexibility for developing an appropriate model for such data. GLMMs have attracted considerable attention over the years. However, unlike linear mixed model, estimation of GLMM by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) is not straightforward. This is because the likelihood is computed by integrating a product of discrete and normal densities, which has no analytical solution. There are two classes of approximation to tackle this problem, marginal and penalised quasi-likelihood (Breslow and Clayton 1993 and Goldstein and Rasbash 1996). These are approximation methods and both methods yield parameter estimates that are biased (McCullagh and Nelder 1989, Jiang 1998, Lin and Breslow 1996). In small area estimation, we require estimates for both small area quantities and their mean squared error (MSE) therefore problem becomes more complex.

An alternative is to ignore the deficiency of the LMM and proceed as if a linear model does hold. These options have the appeal that they are relatively simple to implement. Given the robustness of the estimation procedures, they can be expected to produce reasonable results. Chandra and Chambers (2005, 2009) proposed the model-based direct (MBD) approach for SAE. The MBD estimation is effectively linear estimator and implicitly assumes that the variables of interest follow a LMM. By definition, MBD is a direct estimator and so enjoyed the robustness property of this class of estimator. Weights used to define the MBD 'borrow strength' via a model that explicitly allows for small

E-mail address: hchandra@iasri.res.in

area effects. The empirical evidence shows in case of model misspecification, MBD provides a robust set of small area estimates (Chandra and Chambers 2005, 2009, Chandra *et al.* 2007).

This paper explores an application of MBD approach to estimate small area proportions. The performance of MBD estimation has been compared with the empirical best predictor under generalised linear mixed model via empirical studies using two real data sets. The rest of the paper is organised as follows. Section 2 describes the LMM and defines MBD approach of SAE and their MSE estimation. Section 3 elaborates GLMM and defines the empirical best predictor for small area proportions. Section 4 introduces data used for empirical studies and reports the empirical results. Finally, Section 5 summarizes the results and indicates some further research.

2. ESTIMATION BASED ON A LINEAR MIXED MODEL

Let $\mathbf{Y}_i$ ($N_i \times 1$) be the vector of values of variable of interest y and $\mathbf{X}_i$ ($N_i \times p$) be the matrix of values of the auxiliary variables x in small area i(i = 1,...,m), where m is the number of small areas and N_i is the population size for small area i such that $N = \sum_{i=1}^{m} N_i$. Here N is the population size. We consider the following population level linear mixed model

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

where $\mathbf{Y} = (\mathbf{Y}_1',..,\mathbf{Y}_m')'$, $\mathbf{X} = (\mathbf{X}_1',...\mathbf{X}_m')'$, $\boldsymbol{Z} = diag$ $(\mathbf{Z}_i = \mathbf{1}_{N_i};\ 1 \le i \le m)$, $\mathbf{u} = (u_1',\ldots,u_m')'$ and $\boldsymbol{e} = (\boldsymbol{e}_1',\ldots,\boldsymbol{e}_m')'$ partitioned to small area components. Here β($p \times 1$) is a vector of fixed effects, $\mathbf{Z}_i$ ($N_i \times q$) is a matrix of known covariates characterising differences between small areas, u_i is a random area effect associated with small area i and e_i ($N_i \times 1$) is a vector of individual level random errors. The random variables u_i and $\boldsymbol{e}_i$ are assumed to be independently distributed, with zero means and with variances $Var(u_i) = \sigma_u^2$ and $Var(\boldsymbol{e}_i) = \sigma_e^2\mathbf{I}_{N_i}$ respectively. We assume that two random variables follow the normality. The covariance matrix of $\mathbf{Y}$ is $\mathbf{V} = diag\ (\mathbf{V}_i; 1 \le i \le m)$ with $\mathbf{V}_i = Var\ (\mathbf{Y}_i) = \sigma_e^2\mathbf{I}_{N_i} + \sigma_u^2\mathbf{1}_{N_i}\mathbf{1}'_{N_i}$. In practice the variance components, $\sigma^2 = (\sigma_u^2, \sigma_e^2)$ defining $\mathbf{V}$ are unknown and they are estimated from sample data under the model (1), see Harville (1977). We use a "hat" to denote an estimate and put $\hat{\mathbf{V}} = diag\ (\hat{\mathbf{V}}_i;\ 1 \le i \le m)$ with $\hat{\mathbf{V}}_i = \hat{\sigma}_e^2\mathbf{I}_{N_i} + \hat{\sigma}_u^2\mathbf{1}_{N_i}\mathbf{1}'_{N_i}$. Further, it is assumed that sampling is uninformative given the values of auxiliary variables, so the sample data also follow the population model (1).

Without loss of generality, we consider the sample and non-sample decomposition of $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{V}$ so that $\mathbf{X}_s$ is the $n \times p$ matrix of sample values of the auxiliary variables, $\mathbf{Z}_s$ is the corresponding $n \times q$ matrix of sample components of $\mathbf{Z}$ and $\mathbf{V}_{ss}$ is the $n \times n$ covariance matrix associated with the n sample units that make up the $n \times 1$ sample vector $\mathbf{Y}_s$. That is a subscript of s is used to denote the quantities associated with n sample units of the population. A subscript of r is used to denote corresponding quantities defined by the $N - n$ non-sample units, with $\mathbf{V}_{rs}$ denoting the $(N - n) \times n$ matrix defined by Cov($\mathbf{Y}_r$, $\mathbf{Y}_s$). In what follows we denote $\mathbf{1}_N$, $\mathbf{1}_n$ and $\mathbf{1}_r$ as vectors of 1's and $\mathbf{I}_N$, $\mathbf{I}_n$ and $\mathbf{I}_r$ as identity matrices of order N, n and $N - n$ respectively. We use similar notation at the small area level by introducing an extra subscript i to denote small area. For example, we denote by s_i the set of n_i sample units in area i, r_i the corresponding $N_i - n_i$ non-sampled units and put

$$\mathbf{V}_{iss} = \sigma_e^2\mathbf{I}_{n_i} + \sigma_u^2\mathbf{Z}_{is}\mathbf{Z}'_{is} \text{ and } \mathbf{V}_{isr} = \sigma_u^2\mathbf{Z}_{is}\mathbf{Z}'_{ir}$$

Under model (1), the sample weights that define the empirical best linear unbiased predictor (EBLUP) for the population total of y (Royall 1976) are given by

$$\boldsymbol{w}_{EBLUP} = \mathbf{1}_n + \hat{\mathbf{H}}'(\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{1}_n) + (\mathbf{I}_n - \hat{\mathbf{H}}\mathbf{X}'_s)\,\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \qquad (2)$$

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 MBD estimator (Chandra and Chambers 2005, 2009) for small area i mean of y is then defined as

$$\hat{\bar{Y}}_{i,MBD} = \frac{\sum_{j \in s_i} w_j y_j}{\sum_{j \in s_i} w_j} \qquad (3)$$

The weights in (2) are those associated with the sample units in area i in (3). A robust estimator (Chandra and Chambers 2005, 2009; Royall and Cumberland 1978) of the mean squared error of (3) is

$$mse(\hat{\bar{Y}}_{i,MBD}) = v(\hat{\bar{Y}}_{i,MBD}) + \left\{b(\hat{\bar{Y}}_{i,MBD})\right\}^2 \qquad (4)$$

where

$$v(\hat{\bar{Y}}_{i,MBD}) = \sum_{s_i} \lambda_j (y_j - \boldsymbol{x}_j\hat{\boldsymbol{\beta}})^2$$

$$\lambda_j = N_i^{-2}\{a_j^2 + (N_i - n_i)(n_i - 1)^{-1}\}$$

$$a_j = \left(\sum_{s_i} w_j\right)^{-1}\left(N_i w_j - \sum_{s_i} w_j\right)$$

and $$b(\hat{\bar{Y}}_{i,MBD}) = (\hat{\bar{\mathbf{X}}}_{i,MBD} - \bar{\mathbf{X}}_i)\hat{\boldsymbol{\beta}}$$

with $$\bar{\mathbf{X}}_i = N_i^{-1}\sum_j^{N_i} x_j$$

and $$\hat{\bar{\mathbf{X}}}_{i,MBD} = \left(\sum_{j\in s_i} w_j\right)^{-1}\left(\sum_{j\in s_i} w_j \boldsymbol{x}_j\right)$$

3. ESTIMATION BASED ON GENERALISED LINEAR MIXED MODEL

Like LMM (1), a GLMM (Breslow and Clayton 1993, McCulloch and Searle 2001) includes fixed effects, $\boldsymbol{\beta}$, random effects, $\mathbf{u}$, design matrices $\mathbf{X}$ and $\mathbf{Z}$, and a vector of observations, y, for which conditional distribution given the random effects has mean function, $E(\mathbf{Y} \mid \mathbf{u}) = \pi$. In addition, a GLMM includes a linear predictor, $\boldsymbol{\eta}$, and a link and /or inverse link function. The conditional mean, $E(\mathbf{Y} \mid \mathbf{u})$, depends on linear predictor through an link function $h(\cdot)$. Under this type of model, distribution of values of the variable of interest $\boldsymbol{y}$ is assumed to depend on $\boldsymbol{\eta}$ that is related to regression covariates and random component through the model of form

$$\boldsymbol{\eta} = g(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \qquad (5)$$

That is linear predictor $\boldsymbol{\eta}$ is connected to $\boldsymbol{y}$ via a known function h (inverse of g) as $E(\mathbf{Y} \mid \mathbf{u}) = \boldsymbol{\pi} = h(\boldsymbol{\eta})$. The model (5) fitted using sample data yields the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. The predicted values of non-samples $\boldsymbol{y}$ are given as $\hat{y}_r = h(\hat{\eta}_r)$, where

$$\hat{\eta}_r = \mathbf{X}_r\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

For a binary variable, variable of interest y takes value 1 and 0, with distinct population values of y independently distributed, the function $g(\cdot)$ is *logit* or *logistic* function of the probability, $\pi_j = (y_j = 1, j \in i)$. That is

$$\log it(\boldsymbol{\pi}) = \log\{\boldsymbol{\pi}/(1 + \boldsymbol{\pi})\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \qquad (6)$$

The predicted value of population mean (which is proportion here) of y for small area i is

$$\hat{\bar{\mathbf{Y}}}_i = f_i\bar{\mathbf{Y}}_{is} + N_i^{-1}\sum_{r_i}\hat{y}_j$$

$$= f_i\bar{\mathbf{Y}}_{is} + N_i^{-1}\left(\sum_{r_i}\frac{\exp(\boldsymbol{x}_j\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_i)}{1 + \exp(\boldsymbol{x}_j\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_i)}\right) \qquad (7)$$

where $f_i = n_i/N_i$. The estimator (7) is usually referred as the empirical best predictor (EBP). The analytical expression for MSE of (7) is very complex (Saei and Chambers 2003) which is in contrast to simple MSE estimation of MBD (3). As an alternative, we can use resampling methods for MSE estimation of (7). Some related references are Rao (2001, 2003), Maiti (2004), Jiang *et al*. (2002) and Lahiri (2003). The empirical studies in Section 4 uses nonparametric bootstrap method for MSE of EBP. See for example, Barber and Thompson (2000), and Efron and Tibshirani (1993).

4. EMPIRICAL EVALUATIONS

This section illustrates the empirical performance of two SAE methods to estimate the small area proportions. These are the model based direct estimator under a linear mixed model (denoted by MBD) and the empirical best predictor (denoted by EBP) derived under a generalised linear mixed model. The empirical studies are based two real data sets.

The first data comes from Environmental Monitoring and Assessment Program (EMAP) survey: a sample of 349 plots in the lakes from the North-eastern states of the U.S. The survey is based on a population of 21,028 lakes from which 334 lakes were surveyed, some of which were visited, in different plots, several times during the study period (1991-1996). The 349 plot are the result of their grouping by lake and by 6-digit Hydrologic Unit Codes (HUC). The HUCs are considered as small areas or regions of interest. In original sample data there were 26 small areas but sample sizes for three areas were 1's. I combined these areas with their similar areas. So finally I left with 23 regions or small areas. The sample sizes in these 23 small areas vary from 2 to 45. A population of size $N = 21{,}028$ was generated by sampling N times with 8 replacement from the above sample of 349 plots (units) and with probability proportional to a unit's sample weight; and then 1,000 independently stratified random samples of the same size as the original sample were selected from this (fixed) simulated population. Small area sample sizes were also fixed to be the same as in the original sample. The variable of interest y takes value 1 if Acid Neutralizing Capacity (ANC) - an indicator of the acidification risk of water bodies in water resource surveys is less than 500 and 0 otherwise. The elevation of the lake is the auxiliary variable x, is

Table 1. Percentage relative biases (RB%), percentage relative root mean squared errors (RRMSE%), coverage rates (CR) and small area proportions for the EMAP or Lake data. Regions are arranged in order of increasing population size.

Region	RB%		RRMSE%		CR		Small area proportions		
							Estimated		
	MBD	EBP	MBD	EBP	MBD	EBP	MBD	EBP	True
1	0.00	–5.60	0.00	5.75	1.00	0.01	1.00	0.94	1.00
2	0.00	–1.21	0.00	1.29	1.00	0.05	1.00	0.99	1.00
3	0.00	–8.95	0.00	9.03	1.00	0.01	1.00	0.91	1.00
4	0.00	–2.56	0.00	2.60	1.00	0.03	1.00	0.97	1.00
5	—	—	—	—	1.00	0.30	0.00	0.16	0.00
6	0.00	–0.66	0.00	0.69	1.00	0.02	1.00	0.99	1.00
7	–0.86	1.90	21.62	16.55	0.85	0.86	0.75	0.77	0.76
8	–0.68	3.87	93.71	82.29	0.98	0.59	0.28	0.29	0.28
9	—	—	—	—	1.00	0.02	0.00	0.07	0.00
10	0.47	0.96	21.33	18.61	0.94	0.91	0.64	0.64	0.63
11	–0.18	–1.35	7.12	6.13	1.00	0.65	0.93	0.92	0.93
12	0.85	6.44	38.78	28.08	0.94	0.94	0.44	0.47	0.44
13	–1.48	20.15	76.01	69.87	0.97	0.75	0.22	0.26	0.22
14	–0.16	–0.10	7.40	6.62	0.97	0.91	0.87	0.87	0.87
15	–0.91	2.53	26.21	23.32	0.98	0.94	0.36	0.37	0.36
16	0.00	–2.83	0.00	2.89	1.00	0.03	1.00	0.97	1.00
17	0.00	–1.61	0.00	1.65	1.00	0.04	1.00	0.98	1.00
18	–0.89	1.71	25.27	24.79	0.96	0.90	0.55	0.56	0.55
19	–0.97	0.27	7.92	5.90	0.94	0.94	0.79	0.80	0.79
20	0.94	1.44	25.31	28.19	0.97	0.91	0.47	0.47	0.46
21	–0.45	–0.32	6.44	5.23	0.99	0.90	0.88	0.88	0.88
22	0.20	–1.25	11.06	11.50	0.99	0.86	0.83	0.81	0.82
23	–1.04	0.21	10.98	8.38	0.95	0.93	0.64	0.65	0.65
Min	–1.48	–8.95	0.00	0.69	0.85	0.01	0.00	0.07	0.00
Q1	–0.86	–1.35	0.00	5.23	0.97	0.04	0.46	0.47	0.45
Median	0.00	–0.10	7.92	8.38	0.99	0.75	0.79	0.80	0.79
Mean	–0.25	0.62	18.06	17.11	0.98	0.54	0.68	0.68	0.68
Q3	0.00	1.71	25.27	23.32	1.00	0.91	1.00	0.93	1.00
Max	0.94	20.15	93.71	82.29	1.00	0.94	1.00	0.99	1.00

continuous variable. Here, the interest lies in the estimation of small area proportion of plots for which ANC less than 500. The results generated from this data are set out in Table 1.

The second data is from a sample of 3,591 households spread across 36 districts of Albania. A population of N =724,782 households was generated by sampling N times with replacement from the above sample of 3,591 households and with probability proportional to a household's sample weight; and then 1,000 independently stratified random samples of the same size as the original sample were selected from this simulated population (fixed). District sample sizes were also fixed to be the same as in the original sample, and it varies from very low 8 to very high values 688. In

Table 2. Percentage relative biases (RB%), percentage relative root mean squared errors (RRMSE%), coverage rates (CR) and small area proportions for the Albania data. Regions are arranged in order of increasing population size.

Region	RB%		RRMSE%		CR		Small area proportions		
							Estimated		
	MBD	EBP	MBD	EBP	MBD	EBP	MBD	EBP	True
1	–0.18	–17.94	17.81	20.00	0.99	0.62	0.79	0.65	0.79
2	–0.16	0.68	11.99	10.11	0.96	0.95	0.58	0.59	0.59
3	0.02	14.28	31.79	25.93	0.97	0.92	0.37	0.43	0.37
4	–0.16	–3.65	17.89	11.96	0.95	0.93	0.66	0.64	0.66
5	0.27	4.41	21.38	14.82	0.97	0.93	0.55	0.58	0.55
6	–0.49	9.14	26.12	19.49	0.96	0.91	0.46	0.51	0.47
7	0.59	–2.96	17.57	11.47	0.95	0.96	0.64	0.62	0.64
8	–0.20	–1.74	6.65	6.15	0.96	0.95	0.72	0.70	0.72
9	–0.68	0.15	14.66	11.05	0.94	0.95	0.59	0.59	0.59
10	0.29	–1.79	4.35	4.19	0.98	0.94	0.82	0.81	0.82
11	–0.72	4.29	20.38	15.98	0.97	0.94	0.51	0.53	0.51
12	–0.76	–3.38	13.74	10.14	0.95	0.94	0.69	0.67	0.69
13	–0.11	–0.07	15.18	11.30	0.96	0.95	0.63	0.63	0.63
14	–1.37	20.52	28.40	31.40	0.99	0.90	0.28	0.34	0.28
15	0.18	–0.99	13.85	11.17	0.96	0.95	0.62	0.61	0.62
16	0.14	–0.31	6.39	6.01	0.97	0.96	0.64	0.64	0.64
17	0.91	9.03	19.76	18.91	0.96	0.93	0.34	0.37	0.34
18	–0.10	–6.80	7.47	8.83	0.99	0.89	0.84	0.79	0.84
19	1.19	6.06	19.17	16.18	0.96	0.93	0.47	0.49	0.47
20	–0.20	–2.48	11.01	9.12	0.96	0.95	0.69	0.67	0.69
21	–0.06	–1.20	3.70	3.57	0.97	0.95	0.80	0.79	0.80
22	0.49	–1.50	8.70	7.89	0.98	0.95	0.67	0.66	0.67
23	–0.03	0.13	10.55	9.07	0.95	0.95	0.59	0.60	0.59
24	0.00	–1.63	2.46	2.70	0.99	0.92	0.90	0.88	0.90
25	–0.34	0.19	12.24	10.32	0.96	0.95	0.58	0.59	0.58
26	0.09	–0.58	3.80	3.67	0.97	0.95	0.75	0.75	0.75
27	–0.33	0.88	10.90	10.42	0.96	0.94	0.49	0.49	0.49
28	0.50	3.74	11.44	11.51	0.97	0.94	0.39	0.41	0.39
29	0.29	0.04	6.11	5.66	0.95	0.95	0.64	0.64	0.64
30	–0.21	–0.42	7.45	6.87	0.95	0.95	0.58	0.58	0.58
31	0.14	2.77	10.60	10.68	0.97	0.95	0.37	0.38	0.37
32	–0.02	1.12	9.18	8.81	0.96	0.95	0.42	0.43	0.42
33	0.55	5.12	12.47	13.18	0.98	0.94	0.32	0.33	0.31
34	–0.35	1.56	8.35	8.21	0.98	0.96	0.38	0.39	0.38
35	0.01	–0.70	5.95	5.40	0.94	0.95	0.66	0.66	0.66
36	–0.13	0.70	5.65	5.60	0.97	0.96	0.32	0.32	0.32
Min	–1.37	–17.94	2.46	2.70	0.94	0.62	0.28	0.32	0.28
Q1	–0.20	–1.53	7.25	6.69	0.96	0.93	0.45	0.48	0.46
Median	–0.05	0.09	11.23	10.23	0.96	0.95	0.59	0.60	0.59
Mean	–0.03	1.02	12.64	11.05	0.97	0.93	0.58	0.58	0.58
Q3	0.20	3.01	17.63	12.27	0.97	0.95	0.68	0.66	0.68
Max	1.19	20.52	31.79	31.40	0.99	0.96	0.90	0.88	0.90

this data a binary variable was created which takes value 1 if income of individual is below median income and 0 otherwise. Here, aim is to estimate the proportion of person below median income at the district level, three household level covariates are used. These are ownership of land, television and parabolic dish antenna (as presence of facilities in the dwelling). These covariates take values 0 or 1. Unlike the first data in this case the covariates are binary. Table 2 presents the results from Albanian data.

Three measures of estimation performance were computed using the estimates generated in the simulation studies. These are the relative bias (RB) and the relative root mean squared error (RRMSE), both expressed as percentages, of small area mean estimates and the coverage rate of nominal 95 per cent confidence intervals for small area means. In the evaluation of coverage performances intervals are defined by the small area mean estimates plus or minus twice their standard error.

The relative bias (RB) was measured as

$$\text{RB} = \left\{ M_i^{-1} \left(K^{-1} \sum_{k=1}^{K} \hat{m}_{ik} \right) - 1 \right\} \times 100$$

The relative root mean squared error (RRMSE) was measured as

$$\text{RRMSE} = \left[M_i^{-1} \left\{ \sqrt{K^{-1} \sum_{k=1}^{K} (\hat{m}_{ik} - m_{ik})^2} \right\} \right] \times 100$$

Coverage performance for prediction intervals was measured as

$$\text{CR} = \left\{ K^{-1} \sum_{k=1}^{K} I\left(| \hat{m}_{ik} - m_{ik} | \leq 2 \hat{M}_{ik}^{1/2} \right) \right\} \times 100$$

Note that the subscript of k here indexes the K simulations, with m_{ik} denoting the value of the small area i proportion in simulation k (this is a fixed population value in the design-based simulations considered in this paper), and $\hat{m}_{ik}$, $\hat{M}_{ik}$ denoting the area i estimated value and corresponding estimated MSE in simulation k. The actual area i proportion value (the average over the simulations) is denoted by $M_i = K^{-1} \sum_{k=1}^{K} m_{ik}$.

For the linear mixed model (1), parameters were estimated using REML method. '*lme*' function in *nlme* library in R has been used to derive the estimates for parameters of model (1). In contrast, penalized quasi-likelihood (PQL) method for estimation of parameters for the generalised linear mixed model (5) using '*glmmPQL*' function in *MASS* library in R has been applied. All the simulation results reported in this article are produced using codes written in software R. See http://www.r-project.org.

Table 1 shows the values of relative bias (RB), relative root mean squared error (RRMSE), coverage rate (CR) of small area proportions and values (both true and estimated) of small area proportions for EMAP or Lake data. Analogous results for Albanian data are presented in Table 2. Both Tables also report the mean and the fivepoint summary (Min-minimum, Q1-first quartile, Median- median, Q3-third quartile and Max - maximum) of the distribution of area-specific values of relative bias, relative root mean squared error, coverage rate and values of small area proportions over the small areas. Recall that in Albanian data covariates are binary while in Lake data they are continuous. This gives an opportunity to explore these two methods for two different types of population.

In Table 1 missing values for areas 5 and 9 are noteworthy. In these two areas true area proportions (i.e. population proportions) are zero. Therefore, the relative bias and relative root mean square error for these two areas could not be calculated. The summary statistics in Table 1 are based on values of remaining 21 areas. Two things stand out in Tables 1 and 2. The first is that the MBD offers substantial gains over the EBP in terms of lower average relative bias. Secondly, on average the EBP has a clear efficiency advantage over the MBD. However, this gain in efficiency is marginal only. The average coverage rate for MBD is higher than the EBP. In terms of median values of these measures (i.e., relative bias, relative root mean squared error and coverage rate), the results in Tables 1 and 2 lead to the identical conclusions as based on average values except that median RRMSE of MBD is marginally smaller than the EBP for Lake data. In Tables 1 and 2, the average values of estimated small area proportions generated by two methods (i.e. MBD and EBP) are same as true small area proportion, indicating neither method dominates the other for the population data (Lake and Albanian data) considered in this study.

In Table 1 it can be noted that in few areas (e.g., area 8, 13 and 15) all estimators are unstable. This may be due to the fact that there is little or no variability in the population values of y in these areas. The MBD

appears unaffected, recording RB values that are consistently low. This is in contrast to the EBP, which is quite biased in a number of areas. In Table 1, further it can be seen that in few areas the MBD shows an over coverage. It seems clear that the mean squared error (MSE) for the MBD is being significantly overestimated. This is particularly puzzling for areas 1-6, 9, 16 and 17. A critical examination of results revealed that in these areas true area population proportion is either 1 (e.g., areas 1- 4, 6, 16 and 17) or 0 (e.g., areas 5 and 9). In these cases the estimated small area proportions by MBD estimator are same as population proportion so true MSE turn out to be zero. However, estimated MSE is not zero. This resulted in an overestimation of MSE and over coverage rate of confidence interval in several areas. Although true MSE is not exactly zero for the EBP since they are indirect estimator but similar problem exist with MSE estimation with these methods too. From Table 2, it is clear that such problems do not occur in case of Albania data. For this data both MBD and EBP have shown good coverage rates, see Table 2.

If one looks at the five-point summary given in Table 1 and 2, these results clearly indicates that in terms of bias the MBD dominates EBP. In contrast to MBD, the minimum and maximum values of biases for EBP are very large for both data sets. Description summary statistics for RRMSE reveals that neither method dominates the other in terms of efficiency for the population data considered in this paper.

5. SUMMARY AND FURTHER RESEARCH

An interesting point to note in this paper is an application of linear assumption based MBD approach of SAE for the binary variable. In this context an appropriate method of SAE is the empirical best predictor described in Section 2. However, empirical studies using two real populations clearly show MBD method performs well with no efficiency loss. Although under true model, the EBP should be superior to MBD. In practice true model is unknown and we use working models. In such cases, the MBD performs reasonably well. Overall, for the population data considered in this study MBD dominates the EBP in terms of bias but no method dominates the other in terms of efficiency. Under misspecified models (e.g. area with lack of variability) MBD provides more robust small area estimates. Further, MBD has ease of implementation. In contrast, EBP is a computation intensive method and based on approximations too. Hence, MBD is a real alternative to EBP.

These results further indicate that the EBP breaks down if there is little or no within area variability, while the MBD remains 'stable'. In contrast the estimated MSE for the MBD blows up in these cases. The problem of estimating area proportions when the entire sample is either 1 or 0 is an interesting research area that needs further attention. Author is currently working on this issue.

ACKNOWLEDGEMENT

The author wishes to acknowledge Professor Ray Chambers for his constructive and insightful comments on empirical results, which helped greatly to improve the manuscript. The comments from referee are gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original.

REFERENCES

Barber, J.A. and Thompson, S.G. (2000). Analysis of cost data in randomized trials: An application of the non-parametric bootstrap. *Stat. Medicine,* **19(23)**, 3219-3236.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *J. Amer. Statist. Assoc.*, **88**, 9-25.

Chandra, H. and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Stat. Trans.*, **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. *Stat. Trans.,* **8**, 887-906.

Chandra, H. and Chambers, R. (2009). Multipurpose weighting for small area estimation. *J. Official Stat.,* **25(3)**, 379-395.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Stat. Soc.,* **A159(3)**, 505-513.

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

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.,* **93**, 720-729.

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

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small area estimation. *Statist. Sci.*, **18(2),** 199-210.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Amer. Statist. Assoc.,* **91**, 1007-1016.

Maiti, T. (2004). Applying jackknife method of mean squared prediction error estimation in SAIPE. *Stat. Trans.,* **6**, 685-695.

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

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

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

Rao, J.N.K. (2001). On measuring the quality of indirect small area estimates. *Proceedings of Statistics Canada Symposium 2001 on Achieving Data Quality in a Statistical Agency: A Methodological Perspective.*

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

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

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*, Southampton Statistical Sciences Research Institute,University of Southampton, United Kingdom.