Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 267-276

Estimation of Small Area Proportions Under Unit Level Spatial Models

Hukum Chandra*
Indian Agricultural Statistics Research Institute, New Delhi

(Received: March 2009, Revised: September 2009, Accepted: October 2009)

SUMMARY

Generalized linear mixed models (GLMMs) containing fixed and random area-specific effects are often used for small area estimation (SAE) of discrete variables (McGilchrst 1994 and Rao 2003). In GLMM, the random area effects take account for between areas variation beyond that is explained by auxiliary variables included in the model. These area effects are generally assumed to be independent in SAE. However, in practice area effects are correlated with neighbouring areas and the correlation decays to zero as distance increases. In this paper we investigate SAE based on GLMM with spatially correlated random area effects where the neighbourhood structure is described by a contiguity matrix. We use simulation studies to compare the performances of empirical best predictor for small area proportions under such models with and without spatially correlated area effects. The simulation studies are based on two real data sets. Our empirical results show only marginal gains when spatial dependence between small areas is incorporated into the SAE model.

Key words : Cost function, Domain estimation, Optimal sample design, Probability of item response.

1. INTRODUCTION

The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of statistical methods from both the theoretical and empirical point of view. The traditional estimators (i.e. design-based direct estimators) for small area quantities based on survey data alone are often unstable because of sample size limitations. In this perspective the model-based methodologies allow for the construction of efficient estimators by borrowing the strength through use of a suitable small model. Such estimators are often referred as the indirect estimators, see Rao (2003).

Commonly used model for small area estimation (SAE) of discrete or non-normal data (e.g., binary or count data) is a generalized linear mixed model (GLMM) containing fixed and random effects, see Rao (2003) and McGilchrst (1994). The indirect estimators for small areas under GLMM are the EBLUP type estimators, often known as the empirical best predictors (EBP) for small area quantities, see Saei and Chambers (2003) and Manteiga *et al.* (2007). The mean squared error (MSE) estimation of the EBP is also described in Manteiga *et al.* (2007). These authors have also shown the performance of MSE estimator for the EBP. The area-specific random effects in GLMM take account for the between area dissimilarities beyond that is explained by auxiliary variables included in the fixed part of the model. Although it is customary to assume that these random area effects are independent, in practice most small area boundaries are arbitrary and there appears to be no good reason why population units just one side of such a boundary should not generally be correlated with population units just on the other side. In particular, it is often reasonable to assume that the effects of neighbouring areas (defined, for example, by a contiguity criterion) are correlated, with the correlation decaying to zero as the distance between these areas increases (Pratesi and Salvati 2008, 2009, and Petrucci and Salvati 2006). That is, small area models should allow for spatial correlation of area

* *Corresponding author* : Hukum Chandra
E-mail address : hchandra@iasri.res.in

random effects, See Cressie (1991). Such models allow efficient use of spatial auxiliary information (Chandra *et al.* 2007; Pratesi and Salvati 2008, 2009; Petrucci and Salvati 2006; and Singh *et al.* 2005).

In this paper we consider unit level generalized linear mixed models (Rao 2003, chapter 5 and Manteiga *et al.* 2007) and we extend the EBP for SAE (Saei and Chambers 2003 and Manteiga *et al.* 2007) to account for spatial correlation between the small areas where the neighbourhood structure is described by a contiguity matrix. We then use simulation studies to compare the performances of EBP under such models with and without spatially correlated area effects to examine the gains by incorporating the spatial dependence between the areas. The rest of the paper is organised as follows. In section 2 we review the EBP for SAE under a GLMM with spatially independent small area effects (Saei and Chambers 2003, and Manteiga *et al.* 2007) and discuss the extension of EBP for SAE to account for spatial dependence between the areas. We define the resulting estimator for the small area proportions and their mean squared error estimator. In section 3 we describe the design of our simulation studies and present empirical results and their discussion. In simulation studies we use two real data sets. The first data comes from consumer expenditure survey of the National Sample Survey Organisation (NSSO) for rural areas of state of the Uttar Pradesh in India and the second data from the Environmental Monitoring and Assessment Program (EMAP) survey of lakes in the north-east of the USA. It is noteworthy that two data are from two different real life surveys (i.e., social survey and environmental survey) and very different from each other. This clearly gives us an opportunity to examine the performance of proposed approach of SAE in two different life situations. Finally, in section 4 we provide some concluding remarks and identify further research prospects.

2. THE EMPIRICAL BEST PREDICTOR FOR THE SMALL AREAS

2.1 Models with Spatially Independent Random Area Effects

To start, let us consider a finite population U of size N and assumed to partitioned into D non-overlapping sub-groups (or small areas or small domains) 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 the unit j within small area i, y_{ij} is the survey variable of interest and known for sampled units, $\mathbf{x}_{ij}$ 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. We assume that y_{ij} is typically a binary variable. Let π_{ij} be the probability that a unit j in area i assumes value 1. Let u_i denotes the random area effect for the small area i and assumed to be normally distributed with mean zero and variance ϕ. We assume that u_i's are independent and $y_{ij}|u_i \sim Bin(1, \pi_{ij})$ with $E(y_{ij}|u_i) = \mu_{ij} = \pi_{ij}$ and $Var(y_{ij} \mid u_i) = \sigma_{ij} = \pi_{ij}(1 - \pi_{ij})$. A popular model for this type of data is the logistic linear mixed model of the form

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

where $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of regression parameters. For estimation of unknown model parameters, it is common practice to express model (1) at the population level (Rao 2003, chapter 6). What follows next, we aggregate model (1) and write a population level version of this model as below.

Let $\mathbf{y}_U$ 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 $\mathbf{x}_{ij}$, $\mathbf{G}_U = diag(\mathbf{1}_{N_i};\ 1 \le i \le D)$ is the known matrix of order $N \times D$, $\mathbf{1}_k$ is a column vector of ones of size k, $\mathbf{u} = (u_1, ..., u_D)'$ and $\boldsymbol{\eta}_U$ denotes the $N \times 1$ vector of linear predictors η_{ij} given by (1). We define $\boldsymbol{\mu} = E(\mathbf{y}_U \mid \mathbf{u})$ the conditional mean function of $\mathbf{y}_U$ given $\mathbf{u}$ with elements μ_{ij} and $Var(\mathbf{y}_U \mid \mathbf{u}) = \text{diag}\ \{\sigma_{ij}\}$ the conditional covariance matrix. Let $g(\cdot)$ be a monotonic function, the link function (McCullagh and Nelder 1989, page 27), such that $g(\boldsymbol{\mu})$ can be expressed as the linear model of form

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

The model (2) defines a GLMM, if $\mathbf{y}_U$ given $\boldsymbol{\mu}$ are independent and belong to the exponential family of distribution. Evidently, the vector of random area effects $\mathbf{u}$ has mean $\mathbf{0}$ and variance $\Omega(\delta) = \phi\mathbf{I}_D$, where $\mathbf{I}_D$ is the identity matrix of order D. For binomial response variable the link function $g(\cdot)$ is a logit function, see equation (1). We note that the logistic linear mixed model (1) is a special case of GLMM for logit link. The relationship among $\mathbf{y}_U$ and $\boldsymbol{\eta}_U$ is represented through a known function $h()$, defined by $E(\mathbf{y}_U \mid \mathbf{u}) = h(\boldsymbol{\eta}_U)$. Suppose that our interest is to predict the vector of linear parameters for small areas $\theta = \mathbf{a}_U\mathbf{y}_U$, 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. For example, when y_{ij} is a binary variable and our aim is to estimate proportion for small area i,

$$p_i = N_i^{-1}\sum_{j\in U_i} y_j = N_i^{-1}\left\{\sum_{j\in s_i} y_j + \sum_{j\in r_i} y_j\right\}$$

then $\mathbf{a}_i'$ (i = 1, ..., D) denote the population vector with value N_i^{-1} for each population unit in area i. The estimation of parameter of interest θ is carried out as follows.

Without loss of generality, we arrange the vector $\mathbf{y}_U$ so that its first n elements correspond to the sample units, and then partition $\mathbf{a}_U$, $\mathbf{y}_U$, $\boldsymbol{\eta}_U$, $\mathbf{X}_U$ and $\mathbf{G}_U$ according to sample and non-sample units as

$$\mathbf{a}_U, = \begin{bmatrix}\mathbf{a}_s\\ \mathbf{a}_r\end{bmatrix}, y_U = \begin{bmatrix}\mathbf{y}_s\\ \mathbf{y}_r\end{bmatrix}, \boldsymbol{\eta}_U = \begin{bmatrix}\boldsymbol{\eta}_s\\ \boldsymbol{\eta}_r\end{bmatrix}$$

$$\mathbf{X}_U = \begin{bmatrix}\boldsymbol{X}_s\\ \boldsymbol{X}_r\end{bmatrix} \text{ and } \boldsymbol{G}_U = \begin{bmatrix}\boldsymbol{G}_s\\ \boldsymbol{G}_r\end{bmatrix}$$

Here a subscript s denotes components defined by the n sample units while a subscript r is used to denote corresponding 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, $h()$ is obtained as $g^{-1}()$. Using sample and non-sample deposition of various quantities, parameter of interest $\theta = \mathbf{a}_U\mathbf{y}_U$ can be expressed as

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

Here $\mathbf{y}_s$ the vector of sample values is known, whereas the second term of (3), which depends on the non-samples values $\mathbf{y}_r = h(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{G}_r\mathbf{u})$ is unknown and can be predicted by fitting model (3) for sample data. In our case $\mathbf{y}_s = \{y_{sij}\}$ denotes the vector of sample values of the binary survey variable which takes value 1 or 0. Similarly, $\mathbf{y}_r = \{y_{rij}\}$ represents the vector of non-samples values of the survey variable. It is obvious that the parameter of interest p_i for each small area can be obtained by using as prediction of each element $\{y_{rij}\}$. The problem then reduced to prediction of y_{rij} under model (2) which has two unknown components $\boldsymbol{\beta}$ and $\mathbf{u}$. A major difficulty in use of GLMM for SAE is the estimation of unknown model parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ since the likelihood function for GLMM 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 of inconsistency (Jiang 1998).

For known $\boldsymbol{\Omega}(\boldsymbol{\delta})$, the values of $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated by Penalized Quasi Likelihood (PQL) under model (3) fitted for sample data (Breslow and Clayton 1993). The PQL approach is most popular estimation procedure for the GLMM and it constructs a linear approximation of the distribution of non-normal response variable and assumes the linearised dependent variable is approximately normal. This approach is reliably convergent but it has been noticed that the PQL tends to underestimate variance components as well as fixed effect coefficients (Breslow and Clayton 1993). McGilchrst (1994) introduced the idea to use BLUP to obtain approximate restricted maximum likelihood (REML) estimates for GLMMs. This link between BLUP and REML is illustrated in Harville (1977) for the normal case. For given $\boldsymbol{\Omega}(\boldsymbol{\delta})$, an iterative procedure to obtain maximum Penalized Quasi Likelihood (MPQL) estimate of $\boldsymbol{\beta}$ and $\mathbf{u}$ is described in Saei and Chambers (2003) as below.

1. Assign initial values to $\boldsymbol{\beta}$ and $\mathbf{u}$.
2. Update these values via

$$\begin{bmatrix}\boldsymbol{\beta}_{new}\\ \mathbf{u}_{new}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}_{old}\\ \mathbf{u}_{old}\end{bmatrix} + \mathbf{V}_s^{-1}\begin{bmatrix}\mathbf{X}_s'\\ \mathbf{G}_s'\end{bmatrix}\left(\frac{\partial l_1}{\partial \eta_s}\bigg|_{\boldsymbol{\beta}_{old},\mathbf{u}_{old}}\right) - \mathbf{V}_s^{-1}\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\Omega}^{-1}\mathbf{u}_{old}\end{bmatrix}$$

where $\mathbf{V}_s = \begin{bmatrix}\mathbf{X}_s'\\ \mathbf{G}_s'\end{bmatrix}\left(\frac{\partial^2 l_1}{\partial\eta_s\partial\eta_s'}\bigg|_{\boldsymbol{\beta}_{old},\mathbf{u}_{old}}\right)\begin{bmatrix}\mathbf{X}_s & \mathbf{G}_s\end{bmatrix} + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}^{-1}\end{pmatrix}$

and $\frac{\partial l_1}{\partial \eta_s}$, $\frac{\partial^2 l_1}{\partial\eta_s\partial\eta_s'}$ are first and second derivatives of l_i with respect to η_s.

3. Return to step 2.

At convergence, this gives the best linear unbiased estimate (BLUE) for $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) for $\mathbf{u}$. Hence, using (3) we obtain the BLUP type estimator of θ (i.e, the MPQL estimate of θ).

In practice the variance components parameters defining the matrix $\boldsymbol{\Omega}(\delta)$ are unknown and have to be estimated from sample data. Following Saei and Chambers (2003) an iterative procedure that combines the MPQL estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with ML estimation of $\boldsymbol{\Omega}$ is:

1. Assign initial values to $\boldsymbol{\beta}$, $\mathbf{u}$ and δ
2. Update $\boldsymbol{\Omega}$
3. Update $\boldsymbol{\beta}$ and $\mathbf{u}$ using the iterative PL estimation procedure described in above para
4. Update $\boldsymbol{\eta}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{G}_s\mathbf{u}$
5. Update $\mathbf{B}_s = -(\partial^2 l_1 / \partial\boldsymbol{\eta}_s\partial\boldsymbol{\eta}_s')$
6. Update $\mathbf{T}_s = (\boldsymbol{\Omega}^{-1} + \mathbf{G}_s'\mathbf{B}_s\mathbf{G}_s)^{-1}$
7. Update $\boldsymbol{\delta}$
8. Return to step 2 and repeat the procedure until the values of the different parameters converges.

The corresponding iterative procedure used to obtain the REML estimators is exactly the same except that the role of $\mathbf{T}_s$ in step 6 of ML algorithm is replaced by the $\mathbf{T}_{22}$ submatrix of $\mathbf{T}$ defined in section 6.1 in Saei and Chambers (2003). In our empirical results reported in section 3, we adopted the REML algorithm for parameters estimation.

Using estimated value $\hat{\boldsymbol{\delta}}$ of the $\boldsymbol{\delta}$ leads to the empirical BLUE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and the empirical BLUP (EBLUP) $\hat{\mathbf{u}}$ for $\mathbf{u}$ and the EBLUP type estimator (i.e., empirical best predictor (EBP)) of θ is

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

Turning now to estimation of mean squared error of the EBLUP type predictor or EBP (4) we define

$$\mathbf{H}_r = \mathbf{H}(\hat{\eta}_r) = \partial h(\eta_r)/\partial\eta_r\big|_{\eta_r=\hat{\eta}_r}$$

and

$$\hat{\mathbf{B}}_s = \partial^2 l_1/\partial\eta_s\partial\eta_s'\big|_{\eta_s=\hat{\eta}_s}$$

the matrix of second derivatives of l_1 (the log-likelihood function l_1 defined by the vector $\mathbf{y}_s$ given $\mathbf{u}$) with respect to η_s at $\eta_s = \hat{\eta}_s$. Similarly, $\hat{\mathbf{B}}_r = \partial^2 l_1/\partial\eta_r\partial\eta_r'\big|_{\eta_r=\hat{\eta}_r}$.

We put $\mathbf{X}_r^* = \mathbf{a}_r\mathbf{H}_r\mathbf{X}_r$ and $\mathbf{G}_r^* = \mathbf{a}_r\mathbf{H}_r\mathbf{G}_r$. Then an approximate estimate of the mean squared error for the EBP (4) (see Saei and Chambers 2003; Manteiga *et al.* 2007) is

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

where

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

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

with $\mathbf{C}_r = \left\{\mathbf{X}_r^* - \mathbf{G}_r^*\hat{\mathbf{T}}_s\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{X}_s\right\}$

$$m_3(\hat{\boldsymbol{\delta}}) = \left\{tr\left((\hat{\nabla}_t\hat{\Sigma}_s\hat{\nabla}_k')v(\hat{\boldsymbol{\delta}})\right)\right\}$$

with $\hat{\Sigma}_s = \mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s + \phi\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s$

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

Let $\varsigma = \mathbf{G}_r^*\hat{\mathbf{T}}_s$ and $\mathbf{G}_{rt}^*$ be the t^{th} row of the matrix $\mathbf{G}_r^*$, then $\hat{\nabla}_t = \partial(\varsigma_t)/\partial\delta\big|_{\delta=\hat{\delta}} = \hat{\phi}^{-2}\mathbf{G}_{rt}^*\hat{\mathbf{T}}_s\hat{\mathbf{T}}_s$. 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 $\hat{\boldsymbol{\delta}}$, see Saei and Chambers (2003). Manteiga *et al.* (2007) described the EBP (4) for small area proportion estimation and estimates of their mean squared error. They have studied the empirical performance of MSE estimator (5). However, they have not taken account of spatial dependence between the small areas, which is main objective of this article.

2.2 Models with Spatial Dependence Random Area Effects

In many situations the physical location of the small areas is so relevant that the assumption of spatial independence of the small area models becomes questionable. That is, small area data exhibit a spatial structure and therefore use of spatial models becomes essential. Spatial dependency is the extent to which the value of an attribute in one location depends on the value of the attribute in nearby locations or small areas. Recently the problem has been addressed by introducing a common autocorrelation parameter among small areas extending the linear mixed model through

the Simultaneously Autoregressive (SAR) process (Pratesi and Salvati 2008, 2009; Singh *et al.* 2005; Petrucci and Salvati 2006, Chandra *et al.* 2007). These extensions of small area models (e.g., area level models described in Pratesi and Salvati 2008, 2009; Singh *et al.* 2005 and unit level models discussed in Chandra *et al.* 2007) to spatial small area models are the special case of linear mixed models. The focus here is on the introduction of the SAR process in the generalised linear mixed models (GLMMs) where the vector of random area effects $\mathbf{v} = (v_i)$ satisfies

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u} \qquad (6)$$

where ρ is spatial autoregressive coefficient which determines the degree of spatial dependency of the model, $\mathbf{W}$ is proximity or contiguous matrix of order D. This matrix is symmetric and encapsulates the relative spatial arrangement (i.e. neighbourhood structure) of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. The simplest way to define such a matrix is as simple contiguity: the elements of $\mathbf{W} = \{w_{jk}\}$ take non-zero values only for those pairs of areas that are contiguous to each other. Generally, for ease interpretation, the general spatial weight matrix is defined in row-standardized form; in this case ρ is called spatial autocorrelation parameter (Banerjee *et al.* 2004). In row-standardised form this becomes

$$w_{jk} = \begin{cases} d_j^{-1} & \text{if } j \text{ and } k \text{ are contiguous} \\ 0 & \text{otherwise} \end{cases}$$

where d_j is the total number of areas that share an edge with area j (including area j itself). Contiguity is the simplest but not necessarily the best specification of a spatial interaction matrix. It may be more informative to express this interaction in a more detailed way, e.g. as some function of the length of shared border between neighbouring areas or as a function of the distance between certain locations in each area. Furthermore, the concept of neighbours of a particular area can be defined not just in terms of contiguous areas, but also in terms of all areas within a certain radius of the area of interest. In the empirical evaluations reported later in this paper, however, we used simple contiguity (row-standardized) to define the spatial interaction between different areas. Here

$E(\mathbf{u}) = \mathbf{0}$ and $\text{Var}(\mathbf{u}) = \phi\mathbf{I}_D$

$E(\mathbf{v}) = \mathbf{0}$ and $Var(\mathbf{v}) = \mathbf{\Omega}(\phi\ \rho)$

$= \phi[(\mathbf{I}_D - \rho\mathbf{W})\ (\mathbf{I}_D - \rho\mathbf{W}^T)^{-1}$

where $\mathbf{\Omega}(\phi, \rho) = \mathbf{\Omega}(\delta)$ is the SAR dispersion matrix. To define the EBP under spatially correlated area effects or spatial-EBP (denoted by SEBP), the linear predictor η_U is expressed as

$$\eta_U = \mathbf{X}_U\beta + \mathbf{G}_U\mathbf{v} \qquad (7)$$

where the vector $\mathbf{v}$ is an D-vector of spatially correlated area effects that satisfies SAR model (7). For estimation of unknown model parameters we adopt an iterative procedure similar to one described earlier in this section. However, variance components are now $\delta = (\phi, \rho)$ and $\hat{\mathbf{u}}$ is replaced by $\hat{\mathbf{v}}$. This leads to the spatial EBP of θ (i.e., SEBP) as

$$\hat{\theta} = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r\hat{\beta} + \mathbf{G}_r\hat{\mathbf{v}}). \qquad (8)$$

The MSE of the SEBP (8) are followed from (5) using the variance components $\delta = (\phi, \rho)$ and $\hat{\mathbf{v}}$ in place of $\hat{\mathbf{u}}$.

3. EMPIRICAL EVALUATIONS

In this section we present simulation studies to contrast the performance of the two SAE methods: (i) the empirical best predictor (4) under GLMM with spatially independent area effects, denoted by EBP (see Saei and Chambers 2003 and Manteiga *et al.* 2007) and (ii) the proposed empirical best predictor (8) under GLMM with spatially dependent area effects, denoted by SEBP. The empirical evaluations are based on design-based simulation studies using two real data sets. This evaluates the performance of these methods in the context of real population and realistic sampling methods. The two data sets used in the design-based simulations are from two different types of surveys and are very different to each other. They are

i) The National Sample Survey Organisation (NSSO) Consumer Expenditure Survey: The basis data comes from the survey that underpins the empirical results reported in Sud *et al.* (2008). I used the 61st round survey of NSSO (July 2004-June 2005), the quinquennial series of consumer expenditure survey for rural areas of the state of Uttar Pradesh in India. From this survey, I consider a sample of 307 household from $D = 10$ selected districts (districts are the small area of interest) of state of Uttar Pradesh. The selected districts are all from eastern region of the state so that reasonable neighbourhood can be constructed. This sample of 307 households was bootstrapped to create a realistic population of $N = 76{,}062$

households by sampling with replacement with probability proportional to a household's sample weight. However, in doing so we divided the survey weights in original sample data by 10 to reduce the overall population size, keeping in mind the computation intensity. Note that this does not change the original layout of the survey data except the population sizes used in SAE. A total of $K = 1000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample (varies from 16 to 45). The Y variable of interest takes value 1 if the Household's Monthly Per Capita Expenditure (MPCE) is less than median MPCE of these 10 districts and 0 otherwise. We used the household holding (hectares) of the household as the auxiliary variable. The aim is here to estimate the proportion of households below median MPCE class in each district. The results from this simulation are presented in Table 1.

ii) *The Environmental Monitoring and Assessment Program (EMAP) Survey*: The data consist of 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 (1991-1995). The total number of measurements is 551. The 349 plots are the result of their grouping by lake and by 6-digit Hydrologic Unit Codes (HUC). Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University supplied this data set, developed by EMAP. The HUCs are considered as regions of interest. These areas were having sample sizes as 1 only. Therefore we decided to combine these regions with their similar regions. Consequently, we left with 23 small areas. Sample sizes in these 23 areas vary from 2 to 45. We generated a population of size $N = 21,028$ by sampling N times with replacement from the above sample of 349 plots (units) and with probability proportional to a unit's sample weight; and then $K = 1000$ independently stratified random samples of the same size as the original sample were selected 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

Table 1. District-wise performance measures for the NSSO data. Districts are arranged in order of increasing population size.

Districts	Relative Bias, %		Relative RMSE, %		Coverage rates		Mean squared error			
							EBP		SEBP	
	EBP	SEBP	EBP	SEBP	EBP	SEBP	True	Estimated	True	Estimated
1	90.24	60.68	107.95	78.59	0.67	0.90	0.020	0.008	0.015	0.008
2	−4.42	1.94	21.18	19.59	0.97	0.98	0.009	0.011	0.008	0.011
3	−6.06	−3.07	16.63	14.65	0.96	0.97	0.009	0.011	0.007	0.009
4	1.28	0.04	15.37	15.59	0.98	0.97	0.005	0.008	0.005	0.007
5	−0.35	0.91	18.18	18.07	0.97	0.96	0.007	0.010	0.007	0.008
6	4.92	4.80	18.85	19.30	0.98	0.98	0.005	0.009	0.006	0.007
7	–6.70	−6.48	11.25	10.99	0.93	0.91	0.007	0.005	0.007	0.005
8	−1.26	−0.91	12.41	13.68	0.98	0.96	0.004	0.005	0.005	0.005
9	13.10	16.11	29.52	31.24	0.96	0.92	0.005	0.004	0.004	0.004
10	−2.94	−4.57	9.03	9.63	0.96	0.94	0.004	0.004	0.004	0.004
Average	8.78	6.94	26.04	23.13	0.94	0.95	0.0075	0.0075	0.0068	0.0068

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. We are interested in estimation of small area proportion of plots for which ANC less than 500. Results from this simulation experiment are set out in Table 2.

The performance of different small area estimators were evaluated with respect to three basic criteria –the relative bias and the relative root mean squared error 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

Table 2. Region-wise performance measures for the EMAP data. Regions are arranged in order of increasing population size.

Regions	Relative Bias, %		Relative RMSE,%		Coverage rates		Mean squared error			
							EBP		SEBP	
	EBP	SEBP	EBP	SEBP	EBP	SEBP	True	Estimated	True	Estimated
1	−8.13	−9.16	8.27	9.47	0.99	0.99	0.0068	0.0199	0.0172	0.0168
2	−1.72	−0.66	1.82	0.79	0.99	0.99	0.0003	0.0016	0.0096	0.0137
3	−14.08	−18.18	14.15	18.65	0.99	0.99	0.0200	0.0290	0.0152	0.0088
4	−4.23	−3.86	4.28	3.95	1.00	0.99	0.0018	0.0026	0.0143	0.0280
5	—	—	—	—	—	—	0.0639	0.0460	0.0201	0.0124
6	−1.06	−2.06	1.10	2.20	1.00	1.00	0.0001	0.0003	0.0087	0.0061
7	2.41	2.25	15.83	15.42	0.87	0.87	0.0143	0.0193	0.0129	0.0119
8	6.43	0.29	75.18	71.60	0.91	0.87	0.0442	0.0430	0.0042	0.0062
9	—	—	—	—	—	—	0.0133	0.0185	0.0083	0.0059
10	0.50	1.10	18.06	17.85	0.94	0.94	0.0131	0.0142	0.0074	0.0033
11	−2.40	−0.81	6.16	5.71	1.00	0.94	0.0033	0.0047	0.0054	0.0057
12	10.84	15.66	28.92	32.03	0.98	0.96	0.0161	0.0236	0.0083	0.0107
13	36.37	28.01	73.68	68.63	0.97	0.97	0.0263	0.0258	0.0093	0.0069
14	−0.35	−0.62	6.45	6.42	0.93	0.94	0.0031	0.0031	0.0040	0.0027
15	4.53	2.96	23.48	22.97	0.95	0.96	0.0071	0.0076	0.0075	0.0099
16	−4.65	−5.03	4.71	5.12	1.00	1.00	0.0022	0.0032	0.0034	0.0026
17	−2.64	−2.60	2.69	2.66	1.00	1.00	0.0007	0.0011	0.0057	0.0058
18	3.48	8.45	24.27	26.52	0.90	0.86	0.0180	0.0134	0.0038	0.0048
19	0.44	0.14	5.91	5.87	0.97	0.97	0.0022	0.0027	0.0052	0.0055
20	2.21	3.69	27.50	27.66	0.87	0.87	0.0163	0.0106	0.0045	0.0051
21	−0.72	−0.55	5.20	5.10	0.96	0.96	0.0021	0.0027	0.0035	0.0045
22	−2.17	−1.39	11.35	11.08	0.93	0.92	0.0087	0.0076	0.0041	0.0048
23	0.52	−0.45	8.43	8.39	0.97	0.97	0.0030	0.0038	0.0040	0.0044
Average	1.22	0.82	17.50	17.53	0.96	0.95	0.0125	0.0132	0.0081	0.0081

defined by the estimate of small area proportion plus or minus twice their standard error. The relative bias was measured by *%AvRB*, 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 average over the small areas. The root mean squared error was measured by *%AvRRMSE*, where

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

Coverage performance for prediction intervals was measured by *%AvCR*, where

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

Note that the subscript 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 (averaged over the simulations) is denoted by $M_i = K^{-1}\sum_{k=1}^{K} m_{ik}$.

In Table 1 we report the relative bias (RB) and relative root mean squared error (RRMSE), coverage rates (CR) for nominal 95% intervals for small area proportions and the mean squared error (both true and estimated) of small area proportion estimates for two methods of small area estimation (i.e., EBP and SEBP) based on repeated sampling from the simulated NSSO population. Analogous results for repeated sampling from the simulated EMAP population are presented in Table 2.

The results in Table 1 show that the average relative bias (*%AvRB*) and average relative root mean squared error (*%AvRRMSE*) of the proposed estimator (i.e., SEBP) is smaller than the EBP. Looking at the region specific results in Table 1 we note that relative biases in 7 out of 10 and relative root mean squared errors in 5 out of 10 regions are smaller for SEBP than the EBP. It seems advantageous to include spatial effects in EBP, with a marginal gain. The average coverage rates (*%AvCR*) are slightly underestimated if spatial effects are ignored in small area models, which again show an advantage of including spatial structure. Table 1 clearly shows a consistently good performance of MSE estimate (5) for both SEBP and EBP estimators. We further note that the average value of true MSE of small area proportions for SEBP is slightly lower than the EBP. In 8 out of 10 districts the values of true MSE of SEBP are either smaller or equal to that of true MSE of EBP. This again indicates the gain in small area estimation by incorporating the spatial dependence between the areas.

In Table 2 we noticed that results for regions 5 and 9 are missing. In these two regions true small area proportions (i.e. population proportions for small areas) are zero. Consequently, we could not calculate the relative performance measures (i.e. relative bias and relative root mean square error) since denominators were zero in these cases. The average results in Table 2 therefore are based on the average of remaining 21 regions. In terms of relative biases and relative RMSEs the conclusions from Table 2 are almost identical to results of NSSO data reported in Table 1. In contrast, ignoring the spatial structure in EMAP data leads to overestimation of coverage rates. From the results in Table 2 too we observed only marginal gain in SAE by incorporating spatial effects in estimation. Overall gain by incorporating spatial effects (when neighbourhood structure is described by a contiguity matrix) in small models for binary variable is marginal. The results in Table 2 show that the MSE estimator (5) performs very well for the EMAP data too. The comparative performance of this estimator for the EBP and SEBP is identical to that of NSSO data.

Overall empirical results reveal that MSE estimator performs well. Only a marginal gain can be achieved by including spatial structure in small area estimation of proportions. It is noteworthy that relatively the gains in SEBP are better for NSSO data than the EMAP data. A critical examination of original sample data reflects that the NSSO data has marginally higher degree of spatial dependence between areas than

the EMAP data. The relative gains in NSSO data are therefore more evident than in the EMAP data.

4. CONCLUDING REMARKS

This paper describes SAE of proportions under the GLMM with spatially correlated random area effects where the neighbourhood structure is defined by a contiguity matrix. The empirical results, based on two real data indicate that the gains from inclusion of spatial structure in SAE do not appear to be large. Note that the spatial models considered in this paper are based on neighbourhoods defined by contiguous areas. It is easy to see that this is just one way of introducing spatial dependence between area effects, and several other options remain to be investigated, e.g. geographical weighted regression etc.

There are many issues that still need to be explored in the context of using unit level models with spatially distributed area effects in SAE of discrete data. The most important of these is identification of situations where inclusion of spatial information does have an impact, and the most appropriate way of then including this spatial information in the small area modelling process. An important practical issue in this regard relates to the computational burden in fitting spatial models to survey data. With the large data sets common in survey applications it can be extremely difficult to fit spatial models without access to high-end computational facilities. Although spatial information is becoming increasingly available in environmental, epidemiological and economic applications, there has been comparatively little work carried out on how to efficiently use this information. A further issue relates to the link between the survey data and the spatial information (Chandra *et al.* 2007).

The development in this paper assumes that the sampling method used is uninformative for the population values of Y given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence, same model applies at both sample and population level. However, many often survey data comes from complex sampling designs (e.g., NSSO data illustrated in section 3). There are approaches to incorporate the complex sampling designs for SAE of continuous data (e.g., Pseudo EBLUP under a linear mixed model). However, to my knowledge no such parallel work has been reported for estimation with discrete data. This can be a future research work.

ACKNOWLEDGEMENTS

The constructive and insightful comments from referee are gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original. The author gratefully acknowledges Dr Nicola Salvati for his help in writing R code for empirical evalutions.

REFERENCES

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, New York.

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

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.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada,* 93-105.

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

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

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. *Comput. Statist. Data Anal.*, **51**, 2720-2733.

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

McGilchrst, C.A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc.*, **B56**, 61-69.

Pratesi, M. and Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *J. Off. Statist.*, **25 (1)**, 37–53.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator with autoregressive random area effects. *Statist. Methods Appl.*, **17**, 113-141.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Ag. Biol. Environ. Stat.*, **11**, 169-182.

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-M03/15*, University of Southampton, United Kingdom.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183-195.

Sud U.C., Bathla, H.V.L., Chandra, H. and Singh, J. (2008). Small area estimation - Some application to National Sample Survey data. *Proceedings of the National Seminar on NSS 62th Round Survey Results*, New Delhi, September, 2008.