EMPIRICAL BEST LINEAR UNBIASED PREDICTION FOR OUT OF SAMPLE AREAS

AYOUB SAEI, RAY CHAMBERS

ABSTRACT

Models for small area estimation based on a random effects specification typically assume population units in different areas are uncorrelated. However, they can be extended to account for the correlation between areas by assuming that area random effects are spatially correlated. In this paper we suggest a simple variance-covariance structure for such a spatial correlation structure within the context of a linear model for the population characteristic of interest, and derive estimates of parameters and components of variance using maximum likelihood and restricted maximum likelihood methods. This allows empirical best linear unbiased predictions for area totals to be computed for areas in sample as well as those that are not in sample. An expression for the mean cross-product error (MCPE) matrix of these predicted small area totals is derived, as is an estimator of this matrix. The estimation approach described in the paper is then evaluated by a simulation study, which compares the new method with other methods of small area estimation for this situation.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/03

Empirical Best Linear Unbiased Prediction for Out of Sample Areas

Ayoub Saei

Southampton Statistical Sciences Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Email: A.Saei@soton.ac.uk

Ray Chambers[†]

Southampton Statistical Sciences Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Email: R.Chambers@soton.ac.uk

Summary

Models for small area estimation based on a random effects specification typically assume population units in different areas are uncorrelated. However, they can be extended to account for the correlation between areas by assuming that area random effects are spatially correlated. In this paper we suggest a simple variance-covariance structure for such a spatial correlation structure within the context of a linear model for the population characteristic of interest, and derive estimates of parameters and components of variance using maximum likelihood and restricted maximum likelihood methods. This allows empirical best linear unbiased predictions for area totals to be computed for areas in sample as well as those that are not in sample. An expression for the mean cross-product error (MCPE) matrix of these predicted small area totals is derived, as is an estimator of this matrix. The estimation approach described in the paper is then evaluated by a simulation study, which compares the new method with other methods of small area estimation for this situation.

Key words Spatial correlation, Random effects, Maximum likelihood, REML, Simultaneous autoregressive model.

[†] Contact author for correspondence.

1. Introduction

Efficient estimation of population characteristics for sub-national domains is an important objective for statistical surveys. In particular, geographically defined domains, e.g. regions, states, counties, wards and metropolitan areas are often of interest. Estimates for these domains based on the classical design-based approach to survey sampling inference are often called direct estimates in the literature. However, sample sizes are typically small or even zero within the domains/areas of interest, leading to large sampling variability for these direct estimators. An alternative approach that is now widely used in small area estimation is the so-called indirect or model-based approach. This uses auxiliary information for the small areas of interest and has been characterized in the statistical literature as ″borrowing strength” from the relationship between the values of the response variables and the auxiliary information.

A flexible and popular way of borrowing strength is based on the application of mixed models with area specific random effects (Rao, 2003), with estimation and inferences typically carried out using empirical best linear unbiased prediction (EBLUP), see Prasad and Rao (1990), Singh *et al.* (1998) and You and Rao (2000). In many applications, however, there are no sample observations in some (often many) of the small areas of interest. Clearly, direct estimates cannot be calculated for such out of sample areas. In contrast, model-based estimates for such areas can be computed, but this is typically by making the clearly incorrect assumption that the random effects for these areas are zero. If random effects are uncorrelated between areas there seems to be no way around this problem because there is no area specific sample information about an out of sample area that can be used to estimate its effect. However, most small area boundaries are essentially 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. The implication of this observation is that correlation between small area effects should be the norm, rather than the exception. That is, small area models should allow for spatial correlation of area random effects. An immediate benefit of using such models is that prediction of random area effects for out of sample areas becomes straightforward. This paper therefore extends the EBLUP approach so that estimates for areas in sample as well as those that are not in sample are calculated in a consistent way. In order to do this, it assumes a linear mixed model with spatially correlated area random effects.

In section 2 we define the spatially correlated linear mixed model and its associated notation. Assuming the variance components of this model are known, we develop the corresponding best linear unbiased predictors (BLUPs) for in sample and out of sample areas in section 3. The corresponding EBLUPs are developed in section 4, based on use of maximum likelihood and restricted maximum likelihood methods for estimating the variance components. The mean cross-product errors matrix of the EBLUP estimator and an estimator of this quantity are developed in section 5. Results from a simulation study of the performance of the new method are then provided in section 6. Section 7 concludes the paper with a discussion of potential avenues for further research.

2. Model Specification

Let the vector $\mathbf{y} = \{y_{di}, i = 1,\ldots,N_d; d = 1,\ldots,D\}$ denote the *N*-vector of population values of the survey variable of interest *Y*, where the subscripts *d* and *i* represent area and unit respectively. Let $\mathbf{x}_{di}$ be a known vector of dimension *p* with first element equal to 1 that corresponds to the auxiliary information for population unit *i* in area *d*. The population values of this auxiliary are collected in the $N \times p$ matrix **X**. The objective then is to estimate the value of the vector-valued parameter $\boldsymbol{\theta} = \mathbf{A}\mathbf{y}$, where **A** is a known matrix. In order to do this, we use a linear mixed model to characterise the relationship between the population values of the survey variable and the auxiliary information. This is a model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients (including an intercept) and $\mathbf{u}$ is the *D*-vector of area random effects u_d. Here $\mathbf{Z}$ denotes the incidence matrix for the random component vector **u**, i.e. the matrix that “picks out” population units in different areas. The random vector **u** is assumed to be a realisation from a multivariate normal distribution with zero mean vector and variance-covariance matrix $\sigma_u^2\boldsymbol{\Omega}$ of the same order as the matrix $\mathbf{Z}$. Note that $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\lambda})$ is a function of a parameter $\boldsymbol{\lambda}$. Similarly, the *N*-vector $\mathbf{e}$ is assumed to be independent of $\mathbf{u}$ and normally distributed, with zero mean vector and variance-covariance matrix $\sigma^2\mathbf{W}$, where **W** is a known square matrix of order *N*. The covariance matrix of $\mathbf{y}$ is then

$$\sigma^2(\mathbf{W} + \varphi\mathbf{Z}\boldsymbol{\Omega}\mathbf{Z}') = \sigma^2\boldsymbol{\Sigma} \tag{2}$$

where $\varphi = \sigma_u^2 / \sigma^2$.

After the sample is observed, the vector $\mathbf{y}$ can be partitioned as $\mathbf{y} = [\mathbf{y}_s' \;\; \mathbf{y}_{rs}' \;\; \mathbf{y}_{rr}']'$, with the subscripts of *s* and *r* corresponding to sample and non-sample population units respectively. The subscripts *rs* and *rr* here refer to non-sample units in sample and non-sample areas respectively. The model (1) can then be conformably partitioned as

$$\begin{bmatrix}\mathbf{y}_s\\ \mathbf{y}_{rs}\\ \mathbf{y}_{rr}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_s\\ \mathbf{X}_{rs}\\ \mathbf{X}_{rr}\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\mathbf{Z}_s & \mathbf{0}\\ \mathbf{Z}_{rs} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{rr}\end{bmatrix}\begin{bmatrix}\mathbf{u}_s\\ \mathbf{u}_r\end{bmatrix}+\begin{bmatrix}\mathbf{e}_s\\ \mathbf{e}_{rs}\\ \mathbf{e}_{rr}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_s\boldsymbol{\beta}+\mathbf{Z}_s\mathbf{u}_s+\mathbf{e}_s\\ \mathbf{X}_{rs}\boldsymbol{\beta}+\mathbf{Z}_{rs}\mathbf{u}_s+\mathbf{e}_{rs}\\ \mathbf{X}_{rr}\boldsymbol{\beta}+\mathbf{Z}_{rr}\mathbf{u}_r+\mathbf{e}_{rr}\end{bmatrix} \tag{3}$$

Similarly, the matrix **A** can be partitioned $\mathbf{A}=\begin{bmatrix}\mathbf{A}_s & \mathbf{A}_{rs} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{rr}\end{bmatrix}$, allowing the vector-valued parameter of interest $\boldsymbol{\theta}=\mathbf{Ay}$ to be written

$$\boldsymbol{\theta}=\begin{bmatrix}\mathbf{A}_s\mathbf{y}_s+\mathbf{A}_{rs}\mathbf{y}_{rs}\\ \mathbf{A}_{rr}\mathbf{y}_{rr}\end{bmatrix}. \tag{4}$$

The term $\mathbf{A}_s\mathbf{y}_s$ in (4) depends only on the sample values and is known after the sample is observed. The terms $\mathbf{A}_{rs}\mathbf{y}_{rs}$ and $\mathbf{A}_{rr}\mathbf{y}_{rr}$ depend on non-sample values and are unknown. Our estimated or predicted value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, is therefore

$$\hat{\boldsymbol{\theta}}=\begin{bmatrix}\mathbf{A}_s\mathbf{y}_s+\mathbf{A}_{rs}\hat{\mathbf{y}}_{rs}\\ \mathbf{A}_{rr}\hat{\mathbf{y}}_{rr}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_s\mathbf{y}_s+\mathbf{A}_{rs}(\mathbf{X}_{rs}\hat{\boldsymbol{\beta}}+\mathbf{Z}_{rs}\hat{\mathbf{u}}_s)\\ \mathbf{A}_{rr}(\mathbf{X}_{rr}\hat{\boldsymbol{\beta}}+\mathbf{Z}_{rr}\hat{\mathbf{u}}_r)\end{bmatrix} \tag{5}$$

where a “hat” denotes an estimate of an unknown quantity.

3. Best Linear Unbiased Prediction

A widely used method for defining the estimates in (5) is via substitution of the corresponding best linear unbiased estimators (for unknown parameters) and best linear unbiased predictors (for unknown realisations of random variables). In what follows we do not distinguish between theses two sorts of estimates, referring to both as “BLUPs”. Put l_1 equal to the log-likelihood for $\boldsymbol{\beta}$ and σ^2 generated by $\mathbf{y}_s$ given the value of the random component vector $\mathbf{u}_s$, l_2 equal to the logarithm of the probability density of $\mathbf{u}_s$ given the value of the random component vector $\mathbf{u}_r$, l_3 equal to the logarithm of the probability density function of random component $\mathbf{u}_r$ and set $l=l_1+l_2+l_3$. The BLUPs of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$ are then the values of these quantities where l is maximised (Henderson, 1950).

In order to derive these BLUPs, let $\begin{bmatrix}\boldsymbol{\Omega}_s & \boldsymbol{\Omega}_{sr}\\ \boldsymbol{\Omega}_{rs} & \boldsymbol{\Omega}_r\end{bmatrix}$ be the partition of the variance-covariance matrix $\boldsymbol{\Omega}$ corresponding to the dimensions of the in sample area random effects $\mathbf{u}_s$ and the out of sample area random effects $\mathbf{u}_r$ respectively. Setting the partial derivatives of l with respect to $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$ to zero and solving for these quantities then leads to the BLUP estimating equations

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}}_s \\ \tilde{\mathbf{u}}_r \end{bmatrix} = \mathbf{V}^{-1} \begin{bmatrix} \mathbf{X}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{y}_s \\ \mathbf{Z}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{y}_s \\ \mathbf{0} \end{bmatrix} \tag{6}$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{X}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{X} & \mathbf{X}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{Z}_s & 0 \\ \mathbf{Z}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{X} & \mathbf{Z}'_{\mathbf{s}}\mathbf{W}_s^{-1}\mathbf{Z}_s + \varphi^{-1}\boldsymbol{\Lambda} & -\varphi^{-1}\boldsymbol{\Lambda}\boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_r^{-1} \\ 0 & -\varphi^{-1}\boldsymbol{\Omega}_r^{-1}\boldsymbol{\Omega}_{rs}\boldsymbol{\Lambda} & \varphi^{-1}(\boldsymbol{\Omega}_r^{-1} + \boldsymbol{\Omega}_r^{-1}\boldsymbol{\Omega}_{rs}\boldsymbol{\Lambda}\boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_r^{-1}) \end{bmatrix},$$

$\boldsymbol{\Lambda} = (\boldsymbol{\Omega}_{sr} - \boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_r^{-1}\boldsymbol{\Omega}_{rs})^{-1}$ and $\mathbf{W}_s$ denotes the restriction of **W** to the sampled units. The BLUP of $\boldsymbol{\theta}$ is then

$$\tilde{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{A}_s\mathbf{y}_s + \mathbf{A}_{rs}\tilde{\mathbf{y}}_{rs} \\ \mathbf{A}_{rr}\tilde{\mathbf{y}}_{rr} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_s\mathbf{y}_s + \mathbf{A}_{rs}(\mathbf{X}_{rs}\tilde{\boldsymbol{\beta}} + \mathbf{Z}_{rs}\tilde{\mathbf{u}}_s) \\ \mathbf{A}_{rr}(\mathbf{X}_{rr}\tilde{\boldsymbol{\beta}} + \mathbf{Z}_{rr}\tilde{\mathbf{u}}_r) \end{bmatrix}. \tag{7}$$

The estimator (7) assumes the variance components φ and $\boldsymbol{\lambda}$ are known. In practice of course, this is hardly ever the case. We therefore need to estimate these parameters from the sample data. Two standard ways of doing this are via maximum likelihood (ML) or via restricted maximum likelihood (REML). The latter approach is usually preferable for small to medium sample sizes.

4. Estimation of Variance Components

The variance components estimation method described below is based on that of Henderson (1963, 1975). The approach uses initial estimates of the variances components φ and $\boldsymbol{\lambda}$ to calculate the BLUP estimates (6). These estimates are used in turn as starting values for an iterative procedure that updates these initial estimates. This process is repeated to convergence. This interrelationship between BLUP, ML and REML is discussed in Harville (1977) and is investigated further in Thompson (1980), Fellner (1986, 1987) and Speed (1990). Let

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{bmatrix} \text{ and } \mathbf{V}^{-1} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \mathbf{T}_{13} \\ . & \mathbf{T}_{22} & \mathbf{T}_{23} \\ . & . & \mathbf{T}_{33} \end{bmatrix}$$

be the partitions of the matrix **V** and its inverse that correspond to the dimensions of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$. The iterative procedure used to obtain the ML estimates of φ, $\boldsymbol{\lambda}$ and σ^2 can be specified as follows:

1. Assign initial values to the variance components φ, $\boldsymbol{\lambda}$ and σ^2.

2. Using the current values for these variance components, calculate $\mathbf{\Omega}$.
3. Calculate $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{u}}_s$ and $\tilde{\mathbf{u}}_r$ using (6).
4. Update $\sigma^2 = n^{-1}\mathbf{y}_s'\mathbf{W}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\tilde{\boldsymbol{\beta}} - \mathbf{Z}_s\tilde{\mathbf{u}}_s)$.
5. Calculate $\mathbf{T}_s^* = \begin{bmatrix} \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{32} & \mathbf{V}_{33} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{T}_{22}^* & \mathbf{T}_{23}^* \\ \mathbf{T}_{32}^* & \mathbf{T}_{33}^* \end{bmatrix}$.
6. Update $\varphi = D_s^{-1}(\mathrm{tr}(\mathbf{T}_{22}^*\mathbf{\Omega}_s^{-1}) + \sigma^{-2}\tilde{\mathbf{u}}_s'\mathbf{\Omega}_s^{-1}\tilde{\mathbf{u}}_s)$.
7. Check for convergence of the different estimates. If not return to step 2.
8. Update $\boldsymbol{\lambda} = f(\boldsymbol{\lambda}, \varphi, \mathbf{T}_{22}^*, \sigma^2, \tilde{\mathbf{u}}_s)$ where *f* is the Newton-Raphson updating function for this parameter, i.e. a function whose specification depends on the parameterization of $\mathbf{\Omega}$, and where current values for variance components are used in the right hand side of this equation.
9. Return to step 2 and repeat the procedure until the values of the different parameters converge.

At convergence the ML-based empirical best linear unbiased predictor (EBLUP) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{A}_s\mathbf{y}_s + \mathbf{A}_{rs}\hat{\mathbf{y}}_{rs} \\ \mathbf{A}_{rr}\hat{\mathbf{y}}_{rr} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_s\mathbf{y}_s + \mathbf{A}_{rs}(\mathbf{X}_{rs}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{rs}\hat{\mathbf{u}}_s) \\ \mathbf{A}_{rr}(\mathbf{X}_{rr}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{rr}\hat{\mathbf{u}}_r) \end{bmatrix} \tag{8}$$

where $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}_s$ and $\hat{\mathbf{u}}_r$ are the final values of $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{u}}_s$ and $\tilde{\mathbf{u}}_r$ output by this iterative process. Note that replacing $\mathbf{T}_s^*$ by $\mathbf{T}_{22}$ above leads to the REML estimates of the variance components, and hence the REML-based EBLUP of $\boldsymbol{\theta}$.

5. Estimating the Mean Cross-Product Error (MCPE) Matrix

The EBLUP estimator (8), calculated either via ML or REML, has a prediction error of the form $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \begin{bmatrix} \mathbf{A}_{rs}(\hat{\mathbf{y}}_{rs} - \mathbf{y}_{rs}) \\ \mathbf{A}_{rr}(\hat{\mathbf{y}}_{rr} - \mathbf{y}_{rr}) \end{bmatrix}$. Its associated mean cross-product error (MCPE) matrix is $MCPE(\hat{\boldsymbol{\theta}}) = \mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})']$. Put $\mathbf{A}_r = \begin{bmatrix} \mathbf{A}_{rs} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{rr} \end{bmatrix}$, $\mathbf{X}_r' = [\mathbf{X}_{rs}' \ \ \mathbf{X}_{rs}']'$ and $\mathbf{Z}_r = \begin{bmatrix} \mathbf{Z}_{rs} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{rr} \end{bmatrix}$.

Without loss of generality we assume that the population values are ordered so that values from the D_s in sample areas precede the values from the $D_r = D - D_s$ out of sample areas. After some algebra, it can be shown that

$$MCPE(\hat{\boldsymbol{\theta}}) \cong M_{\beta}(\boldsymbol{\omega}) + M_{\beta u}(\boldsymbol{\omega}) + M_u(\boldsymbol{\omega}) + M_e(\sigma^2) + M_{\omega}(\boldsymbol{\omega}) \tag{9}$$

where $MCPE(\tilde{\boldsymbol{\theta}}) = M_\beta(\hat{\boldsymbol{\omega}}) + M_{\beta u}(\hat{\boldsymbol{\omega}}) + M_u(\hat{\boldsymbol{\omega}}) + M_e(\hat{\sigma}^2)$ is the corresponding mean cross-product error matrix of the BLUP estimator $\tilde{\boldsymbol{\theta}}$. Here $M_\beta(\boldsymbol{\omega})$ and $M_u(\boldsymbol{\omega})$ measure the uncertainty due to estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$; $M_{\beta u}(\boldsymbol{\omega})$ is the covariance between the estimators of $\boldsymbol{\beta}$ and $\mathbf{u}$; $M_\omega(\boldsymbol{\omega})$ measures the uncertainty due to estimation of the variance components $\boldsymbol{\omega} = (\sigma^2, \varphi, \boldsymbol{\lambda}')'$ and $M_e(\sigma^2)$ is the uncertainty due to estimation of the error term. The approximation (9) (without the $M_e(\sigma^2)$ term) is due to Kacker and Harville (1984) and is discussed in Prasad and Rao (1990) and Datta and Lahiri (2000). Put $\mathbf{X}_r^* = \mathbf{A}_r\mathbf{X}_r$, $\mathbf{Z}_r^* = \mathbf{A}_r\mathbf{Z}_r$, $\mathbf{X}_s^* = \mathbf{Z}_s\mathbf{W}_s^{-1}\mathbf{X}_s$ and $\mathbf{y}_s^* = \mathbf{Z}_s\mathbf{W}_s^{-1}\mathbf{y}_s$. The components of (9) are then

$$M_\beta(\boldsymbol{\omega}) = \sigma^2\mathbf{X}_r^*\mathbf{T}_{22}\mathbf{X}_r^{*\prime}\,,\ M_u(\boldsymbol{\omega}) = \sigma^2\mathbf{Z}_r^*\mathbf{T}_s^*\mathbf{Z}_r^{*\prime}\,,\ M_e(\sigma^2) = \sigma^2\mathbf{A}_r\mathbf{W}_r\mathbf{A}_r'$$

and

$$M_{\beta u}(\boldsymbol{\omega}) = -\sigma^2[\mathbf{X}_r^*(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s\mathbf{T}_s^*\mathbf{Z}_r^{*\prime} + \mathbf{Z}_r^*\mathbf{T}_s^*\mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_r^{*\prime}].$$

The final component $M_\omega(\boldsymbol{\omega})$ is a measure of the uncertainty due to estimation of the variance components $\boldsymbol{\omega}$ and is defined as follows. Put $\boldsymbol{\Delta} = \mathbf{Z}_r^*\mathbf{T}_s^* = [\boldsymbol{\Delta}_1', \boldsymbol{\Delta}_2', \ldots, \boldsymbol{\Delta}_D']'$ and let $\mathbf{Z}_\alpha^*$ be the α^{th} row of the matrix $\mathbf{Z}_r^*$, so that $\partial\boldsymbol{\Delta}_\alpha / \partial\boldsymbol{\gamma} = \partial(\mathbf{Z}_\alpha^*\mathbf{T}_s^*)/\partial\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = (\varphi\ \ \boldsymbol{\lambda}')'$. Then

$$M_\omega(\boldsymbol{\omega}) = \sigma^2[\mathrm{tr}(\boldsymbol{\nabla}_\alpha\boldsymbol{\Sigma}_s^*\boldsymbol{\nabla}_{\alpha'}'\mathbf{B})]$$

where $\boldsymbol{\Sigma}_s^* = \mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s + \varphi\mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s\boldsymbol{\Omega}_s\mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s$ and $\boldsymbol{\nabla}_\alpha$ is the first D_s columns of the matrix $\partial\boldsymbol{\Delta}_\alpha / \partial\boldsymbol{\gamma}$. Here $\mathbf{B}$ is the asymptotic variance-covariance matrix of the estimator of the variance components vector $\boldsymbol{\gamma}$. An estimate of MCPE matrix of the EBLUP $\hat{\boldsymbol{\theta}}$ is therefore

$$\widehat{MCPE}(\hat{\boldsymbol{\theta}}) = M_\beta(\hat{\boldsymbol{\omega}}) + M_{\beta u}(\hat{\boldsymbol{\omega}}) + M_u(\hat{\boldsymbol{\omega}}) + M_e(\hat{\sigma}^2) + 2M_\omega(\hat{\boldsymbol{\omega}}) \tag{10}$$

where $\hat{\boldsymbol{\omega}}$ is the ML/REML estimate of the variance components vector $\boldsymbol{\omega}$. Note that the multiplier of two for the last term on the right hand side of (10) follows because (see Datta and Lahiri, 2000)

$$E\big(M_u(\hat{\boldsymbol{\omega}})\big) = M_u(\boldsymbol{\omega}) - M_\omega(\boldsymbol{\omega}).$$

6. Simulation Results

In this section we present results from a simulation study of the EBLUP methodology outlined above that focuses on estimating the small area totals for a survey variable *Y*, so $\boldsymbol{\theta} = \mathbf{Ay}$ is the vector of these small area totals. The population values themselves were generated from a linear mixed model with spatially correlated area random effects, defined by

$$y_{di} = 0.5 + x_{di} + u_d + e_{di}\,. \tag{11}$$

The values e_{di} were independently generated from a normal distribution with zero mean and variance σ^2. The values $\mathbf{u} = [u_1, u_2, ..., u_{D_s}, ..., u_D]'$ were generated from a multivariate normal distribution with zero mean vector and variance-covariance matrix

$$\sigma_u^2 \mathbf{\Omega}(\lambda) = \sigma_u^2 [(\mathbf{I}_D - \lambda \mathbf{\Lambda})(\mathbf{I}_D - \lambda \mathbf{\Lambda}')]^{-1} \tag{12}$$

where $\mathbf{I}_D$ is an identity matrix of order *D* and $\mathbf{\Lambda}$ is a spatial weight matrix. This is the SAR or simultaneous autoregressive model (Cressie, 1993). The symmetric spatial weight matrix $\mathbf{\Lambda}$ was made up of ones and zeros with $\mathbf{\Lambda}_{ij} = 1$ if areas *i* and *j* are considered "spatial neighbours" and is zero otherwise. It was generated by randomly assigning "neighbours" to each area in the population and was kept fixed for all simulations. The x_{di} values were generated from a uniform distribution between 0 and 1 and were also kept fixed throughout the simulations. Values of y_{di} were generated for *D* = 30 and *D* = 100 areas with 90 population units per area. Random samples of size n_d were taken from each area, with n_d increasing with *d*. The first D_s = 25 (95) areas were taken to be in sample areas, with the remaining D_r = 5 areas considered as being out of sample areas. The sample data from the in sample areas were used to estimate the model parameters via REML, and then estimates of all 30 (100) area totals for the y_{di} were calculated. Note that under (12) the Newton-Raphson updating equation for the parameter λ is given by

$$\lambda_k \;=\; \lambda_{k-1} + b_1 b_2$$

where $b_1 = -0.5[\varphi^{-1}\sigma^{-2}\tilde{\mathbf{u}}'(\partial\mathbf{\Omega}^{-1}/\partial\lambda)\tilde{\mathbf{u}} + \varphi^{-1}\mathrm{tr}((\partial\mathbf{\Omega}^{-1}/\partial\lambda)\mathbf{T}_{22}^{*}) - \mathrm{tr}((\partial\mathbf{\Omega}^{-1}/\partial\lambda)\mathbf{\Omega})]$ and b_2 is the (3, 3) element of the information matrix of the estimators $\hat{\sigma}^2$, $\hat{\varphi}$ and $\hat{\lambda}$.

We considered four ways of defining the small area estimates:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{A}_s \mathbf{y}_s + \mathbf{A}_{rs}\mathbf{X}_{rs}\hat{\boldsymbol{\beta}} \\ \mathbf{A}_{rr}\mathbf{X}_{rr}\hat{\boldsymbol{\beta}} \end{bmatrix} \tag{13a}$$

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{A}_s \mathbf{y}_s + \mathbf{A}_{rs}(\mathbf{X}_{rs}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{rs}\hat{\mathbf{u}}_s) \\ \mathbf{A}_{rr}\mathbf{X}_{rr}\hat{\boldsymbol{\beta}} \end{bmatrix} \tag{13b}$$

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{A}_s \mathbf{y}_s + \mathbf{A}_{rs}(\mathbf{X}_{rs}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{rs}\hat{\mathbf{u}}_s) \\ \mathbf{A}_{rr}(\mathbf{X}_{rr}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{rr}\hat{\mathbf{u}}_r) \end{bmatrix} \tag{13c}$$

$$\hat{\boldsymbol{\theta}} = \mathbf{A}_s \mathbf{y}_s + \mathbf{A}_r(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}})\,. \tag{13d}$$

Here (13a) corresponds to a synthetic estimation procedure, where the mixed model defined by (11) and (12) is first fitted to the sample data, but then estimation is carried out on the basis that $u_d = 0$ in every small area. In contrast, (13b) fits the model (11) to the sample data, but forces $\lambda = 0$ in (12), i.e. this estimator assumes there is no spatial correlation among the area effects. The estimator defined by (13c) corresponds to the EBLUP procedure defined earlier in this paper, while (13d) serves as a benchmark since it assumes that sample data are available from every small area (and so works with a larger sample size than the preceding methods).

The process of generating population and sample data, estimation of model parameters and calculation of (13a) – (13d) was independently replicated 2000 times. For each set of estimates $\hat{\boldsymbol{\theta}}$ and each small area d we then calculated the actual and average estimated mean squared errors

$$ActMSE_d = diag_d\left(\sum_{k=1}^{2000}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)' / 2000\right)$$

$$EstMSE_d = diag_d\left(\sum_{k=1}^{2000}\widehat{MCPE}(\hat{\boldsymbol{\theta}}_k) / 2000\right)$$

where $diag_d(\mathbf{X})$ denotes the d^{th} element of the main diagonal of $\mathbf{X}$. The actual coefficient of variation

$$ActCV_d = 100 \times \frac{ActMSE_d}{\sum_{k=1}^{2000}\boldsymbol{\theta}_{dk} / 2000}$$

and the estimated coefficient of variation

$$EstCV_d = 100 \times \frac{EstMSE_d}{\sum_{k=1}^{2000}\hat{\boldsymbol{\theta}}_{dk} / 2000}$$

were then calculated, as was the average coverage of the area d total by the nominal 95% confidence intervals defined by these estimated mean squared errors.

Nine different combinations of overall sample sizes and parameter values in (11) and (12) were used in the simulations. These are denoted Par1 – Par9 and are set out in Table 1. Table 2 shows the average values of both the actual coefficient variation (*ActCV*) and estimated coefficient of variation (*EstCV*) for the estimators (13a) – (13d). These show that for Method (13a) in particular, estimated CVs are far from their actual values, irrespective of whether the areas concerned are in sample or out of sample. This problem persists, albeit in a somewhat reduced form, with Method (13b), where now it is out of sample areas whose estimated CVs tend to be far too optimistic. Both Method (13c) and Method (13d) – as one would expect – perform much better in this regard, with estimated and actual CVs for both in sample and out of sample areas under Method (13c) being very close. Note also that average values of *ActCV* for Methods (13b) and (13c) in Table 2 are very similar for small values of λ, but indicate substantial gains in efficiency for (13c) for large values of σ_u^2 and λ. As might be expected, these gains are more pronounced for large values of *D*.

Irrespective of potential increases in efficiency, an important gain from modelling the spatial correlation of the area random effects is better estimation of mean squared error. This is confirmed in Table 3 where we see that Method (13a) generally leads to severe undercoverage because it is based on conditionally biased synthetic estimators. In contrast, Method (13b) has good coverage for in sample areas, but poor coverage for out of sample areas (even when there is no spatial correlation), reflecting its use of conditionally biased synthetic estimators for these areas. There also seems to be some evidence that this coverage gets worse as this spatial correlation increases. On the other hand, Method (13c) records coverages very close to the nominal 95% level for in sample areas, and only slightly less for out of sample areas. Furthermore, this overall good performance holds across all sets of parameter values investigated, including where there is no spatial correlation. Note that larger values of *D* also lead to better coverage performance.

7. Summary and Discussion

In this paper we describe a method for constructing the EBLUP for a small area total or mean when there are no sample units in the area. In doing so, we assume a unit level linear model with spatially correlated area effects defined by the SAR model (12). Our simulations indicate that our proposed method has the potential to lead to substantial increases in prediction efficiency for these areas when there is strong spatial correlation in the data. They also show that the estimates of mean squared error calculated under the spatial model are much more accurate than those based on the usual synthetic estimates that are often used for prediction in out of sample areas. As a consequence, confidence intervals based on these estimates of mean squared error tend to be more accurate, in the sense of achieving their nominal level of coverage.

The analysis in this paper has been restricted to the linear mixed model (1) and assumes the availability of unit level data. Many applications, however, are based on area level data and/or non-linear mixed models, e.g. generalised linear mixed models. The methodology outlined in this paper can be extended to these situations, and results from this research will be published elsewhere.

References

Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley.

Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10**, 613 – 627.

Fellner, W.H. (1986). Robust estimation of variance components. *Technometrics* **28**, 51 – 60

Fellner, W.H. (1987). Spare matrices, and the estimation of variance components by likelihood methods. *Communication in Statistics and Simulation.* **16**, 439 – 463

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320-340.

Henderson, C.R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*. **21** 309 – 310.

Henderson, C.R. (1963) Selection index and expected genetic advance. *In Statistical Genetics and Plant Breeding* (W.D. Hanson and H.F. Robinson, eds.), 141-163. National Academy of Sciences and National Research Council Publication No. 982, Washington, D.C.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under selection model. *Biometrics* **31**, 423-447.

Kacker, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimations of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853-862.

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

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

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society Series B* **60**, 377-396.

Speed, T. (1991) Comment on Robinson: Estimation of random effect. *Statistical Science* **6**, 42 - 44.

Thompson , R. (1981). Maximum likelihood estimation of variance components. *Math. Operforch. Statist. Ser. Statist.* **11**, 125 – 131.

You, Y. and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, **26** 173-181.

Table 1. Parameter sets used in the simulations. Note D_s = number of in sample areas, D_r = number of out of sample areas and $\bar{n}$ is average sample size for in sample areas.

Set	Parameter Values					
	$\bar{n}$	σ^2	σ_u^2	λ	D_s	D_r
Par1	8.1	1	0.5	0.7	25	5
Par2	9.2	3	1.5	0.7	25	5
Par3	12.4	1	0.5	0.0	25	5
Par4	10.4	1	0.5	0.2	25	5
Par5	9.4	1	0.5	0.7	95	5
Par6	8.4	1	0.5	4	25	5
Par7	11.1	1	0.5	4	95	5
Par8	7.0	1	5	4	25	5
Par9	9.7	1	5	4	95	5

Table 2 Estimated coefficients of variation (*EstCV*) and actual coefficients of variation (*ActCV*) for different methods of estimation, averaged over the small areas. *Areas* denotes the small areas whose values are averaged, while *Set* denotes the set of parameter values used in the simulation (see Table 1 for their definition).

Areas	*Set*	Method							
		(13a)		(13b)		(13c)		(13d)	
		ActCV	*EstCV*	*ActCV*	*EstCV*	*ActCV*	*EstCV*	*ActCV*	*EstCV*
All	Par1	84.49	36.59	40.29	31.37	40.92	40.78	31.75	32.04
	Par2	139.8	61.04	67.98	52.87	69.07	69.67	53.08	53.57
	Par3	77.06	29.43	41.97	33.05	42.96	43.24	33.28	33.85
	Par4	73.33	28.66	36.02	27.25	36.55	36.95	27.67	27.91
	Par5	75.27	28.44	35.07	32.05	35.19	35.40	32.39	32.59
	Par6	159.7	16.88	57.42	38.33	39.27	41.69	30.20	31.11
	Par7	82.80	10.75	34.12	29.88	31.06	30.79	28.41	28.18
	Par8	591.69	36.7	122.3	63.72	74.41	75.62	44.41	44.82
	Par9	267.96	13.65	49.95	36.62	44.43	44.56	35.44	35.51
In sample	Par1	83.07	36.08	33.87	33.80	34.38	34.01	33.97	34.29
	Par2	136.62	59.92	57.01	56.86	57.89	57.61	57.23	57.76
	Par3	75.35	28.96	35.63	35.70	36.40	36.33	35.99	36.61
	Par4	71.89	28.23	28.99	29.05	29.29	29.39	29.18	29.45
	Par5	74.80	28.28	33.03	33.02	33.12	33.27	33.07	33.28
	Par6	160.34	16.53	37.75	37.84	32.86	33.55	32.01	33.06
	Par7	81.92	10.69	30.68	30.66	28.98	28.7	28.88	28.64
	Par8	613.06	36.74	50.16	51.79	49.07	53.57	48.76	49.17
	Par9	266.27	13.62	36.79	36.88	36.32	36.43	36.29	36.37
Out of sample	Par1	91.59	39.16	72.42	19.22	73.63	74.61	20.68	20.8
	Par2	155.72	66.62	122.82	32.94	124.99	129.99	32.31	32.57
	Par3	85.65	31.76	73.66	19.76	75.77	77.77	19.76	20.01
	Par4	80.57	30.79	71.19	18.27	72.86	74.75	20.15	20.21
	Par5	84.10	31.37	73.72	13.74	74.43	75.90	19.39	19.51
	Par6	156.54	18.63	155.76	40.79	71.33	82.38	21.16	21.37
	Par7	99.55	11.89	99.51	15.16	70.67	70.63	19.56	19.46
	Par8	484.82	36.51	483.00	123.36	201.11	185.89	22.69	23.05
	Par9	299.96	14.29	299.95	31.67	198.56	198.95	19.14	19.21

Table 3 Coverage of nominal 95% confidence intervals (*95%Coverage*) generated by different methods of estimation, averaged over the small areas. *Areas* denotes the set of small areas whose values are being averaged, while *Set* denotes the set of parameter values used in the simulation (see Table 1 for their definition).

Areas	*Set*	*95%Coverage*			
		(13a)	(13b)	(13c)	(13d)
All	Par1	46.02	85.59	93.87	94.9
	Par2	46.91	85.48	94.22	94.99
	Par3	41.92	85.56	94.05	95.03
	Par4	43.53	85.44	94.55	94.93
	Par5	37.16	91.68	94.93	95.08
	Par6	16.66	85.51	95.38	95.51
	Par7	19.95	91.31	94.77	94.75
	Par8	10.05	85.03	95.65	95.13
	Par9	8.09	91.11	95.04	95.03
In sample	Par1	46.18	94.77	94.20	94.87
	Par2	47.11	94.68	94.44	94.98
	Par3	42.20	94.70	94.30	94.97
	Par4	43.76	94.76	94.79	94.94
	Par5	37.17	94.92	94.97	95.06
	Par6	16.49	95.26	95.29	95.52
	Par7	20.01	94.9	94.78	94.76
	Par8	9.66	95.35	96.26	95.07
	Par9	8.12	95.04	95.07	95.04
Out of sample	Par1	45.25	39.66	92.23	95.04
	Par2	45.93	39.48	93.13	95.05
	Par3	40.56	39.86	92.77	95.37
	Par4	42.37	38.82	93.31	94.88
	Par5	37.05	30.03	94.16	95.44
	Par6	17.51	36.81	95.80	95.45
	Par7	18.96	22.96	94.64	94.60
	Par8	12.00	33.39	92.61	95.43
	Par9	7.54	16.49	94.39	94.86