OUT OF SAMPLE ESTIMATION FOR SMALL AREAS USING AREA LEVEL DATA

AYOUB SAEI, RAY CHAMBERS

ABSTRACT

A Fay-Herriot type model with independent area effects is often assumed when small area estimates based on area level data are required. However, under this approach out of sample areas are limited to synthetic estimates. In this paper we relax the independent area effects assumption, allowing area random effects to be spatially correlated. Empirical best linear unbiased predictors are then developed for areas in sample as well as those that are not in sample, with variance components estimated via maximum likelihood and residual (restricted) maximum likelihood. An expression for the mean cross-product error (MCPE) matrix of the small area estimators 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 M05/11

Out of Sample Estimation for Small Areas using Area Level Data

Ayoub Saei and **Ray Chambers**†

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

Email: A.Saei@soton.ac.uk and R.Chambers@soton.ac.uk

† Contact author for correspondence.

Summary

A Fay-Herriot type model with independent area effects is often assumed when small area estimates based on area level data are required. However, under this approach out of sample areas are limited to synthetic estimates. In this paper we relax the independent area effects assumption, allowing area random effects to be spatially correlated. Empirical best linear unbiased predictors are then developed for areas in sample as well as those that are not in sample, with variance components estimated via maximum likelihood and residual (restricted) maximum likelihood. An expression for the mean cross-product error (MCPE) matrix of the small area estimators 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.

1. Introduction

Large sample surveys are usually designed to produce reliable estimates of national or large area characteristics. However, there is a growing demand for similar estimates for smaller areas or domains. Such estimates are now routinely calculated using 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.

In many applications, however, only area level aggregate data are available. In such cases, small area estimation is usually based on the Fay-Herriot mixed model with independent area specific random effects (Rao, 2003). See Fay and Herriot (1979), Ghosh and Rao (1994), Prasad and Rao (1990), Datta and Lahiri (2000) and Butar and Lahiri (2003). Furthermore, there can be some (often many) small areas of interest where there is no sample, and consequently, no data. Model-based estimates for such areas can be computed, but this is typically by making the clearly incorrect assumption of no random effects for these areas. If random effects are uncorrelated between areas there seems to be no way around this problem because there is no 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. In an earlier paper (Saei and Chambers, 2005) the authors developed this approach for the case where unit level data are available from the in-sample areas. This paper extends this approach to the important aggregate data situation. In particular, the empirical best linear unbiased predictor (EBLUP) for an area level characteristic for areas in sample as well as those that are not in sample is derived for the case of aggregate area level data, assuming a Fay-Herriot-type linear mixed model with spatially correlated area random effects.

In section 2 we define this model and its associated notation. The EBLUP is developed in section 3, based on use of either maximum likelihood or restricted maximum likelihood methods for estimating the variance components of the model. The mean cross-product errors matrix of the EBLUP estimator and an estimator of this quantity are developed in section 4. Results from a simulation study of the performance of the new method are then provided in section 5. Section 6 concludes the paper with a discussion of potential avenues for further research.

2. Model Specification

We assume that the population of interest is made up of D areas and the aim is to estimate the conditional expectation of the area mean of a survey variable Y for each of them. The D-vector of these conditional expectations is denoted by θ and we note that the conditioning is with respect to the values of auxiliary variables as well as random area effects. We assume a standard unit level mixed effects model for Y, defined in terms of a fixed effect contribution $X'\beta$, where X is a vector X of unit level

characteristics, a random area effect $Z'\mathbf{u}$, where Z is a vector of area level characteristics, and an individual random effect. This induces a linear model for $\boldsymbol{\theta}$, given by $\boldsymbol{\theta} = \bar{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, where $\bar{\mathbf{X}}$ is the matrix of area means of X, $\mathbf{Z}$ is the matrix of (area level) values of Z, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients (including an intercept) and $\mathbf{u}$ is an unknown vector of random area effects. Following standard practice, we assume that $\mathbf{u}$ is a realisation from a multivariate normal distribution with zero mean vector and variance-covariance matrix $\sigma_u^2\boldsymbol{\Omega}$ of order D. Furthermore, $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\lambda})$ is a function of an unknown parameter $\boldsymbol{\lambda}$.

Not all of the areas will be represented in sample. An area d will be denoted as in sample if a direct survey estimate for the characteristic θ_d of that area is available. Otherwise the area is out of sample. Let D_s denote the number of small areas in sample, with $D_r = D - D_s$ denoting the number of out of sample areas. We can then partition the matrices $\bar{\mathbf{X}}$ and $\mathbf{Z}$ into $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_s' \; \bar{\mathbf{X}}_r']'$ and $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_r \end{bmatrix}$ corresponding to sample and non-sample areas. Similarly we can partition the random area effects vector as $\mathbf{u} = [\mathbf{u}_s' \; \mathbf{u}_r']'$. The objective is to predict the value of

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_s \\ \boldsymbol{\theta}_r \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{X}}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}_s \\ \bar{\mathbf{X}}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u}_r \end{bmatrix}. \tag{1}$$

Let $\tilde{\boldsymbol{\theta}}_s = \{\tilde{\theta}_d, \; d = 1, 2, ..., D_s\}$ denote the vector of direct estimates for the in sample areas. Generally, these direct estimates will be weighted averages of unit level data obtained in samples taken from each of the in sample areas. Consequently application of

the underlying unit level model that led to (1) implies that these direct estimates will follow the closely related model

$$\tilde{\boldsymbol{\theta}}_s = \boldsymbol{\theta}_{sw} + \mathbf{e}_s \tag{2}$$

where $\boldsymbol{\theta}_{sw} = \bar{\mathbf{X}}_{sw}\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}_s$ and $\mathbf{e}_s$ represents estimation error. Here $\bar{\mathbf{X}}_{sw}$ denotes the weighted estimate of $\bar{\mathbf{X}}_s$ obtained from the sample data in the in sample areas. Note that $\bar{\mathbf{X}}_{sw}$ can be replaced by $\bar{\mathbf{X}}_s$ when $\bar{\mathbf{X}}_{sw}$ is not available.

The model (2) is often referred to as a Fay-Herriot model. Under this model the sampling error vector $\mathbf{e}_s$ is assumed to be independent of **u** and normally distributed, with zero mean vector and variance-covariance matrix $\sigma^2\mathbf{W}_s$, where $\mathbf{W}_s$ is a known square matrix of order *n* (the overall sample size). Combining this assumption with (1), we see that the implied model for $\tilde{\boldsymbol{\theta}}_s$ is

$$\tilde{\boldsymbol{\theta}}_s = \boldsymbol{\theta}_{sw} + \mathbf{e}_s = \bar{\mathbf{X}}_{sw}\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}_s + \mathbf{e}_s = \bar{\mathbf{X}}_{sw}\boldsymbol{\beta} + \boldsymbol{\zeta}_s \tag{3}$$

where $Var(\boldsymbol{\zeta}_s) = \sigma^2(\mathbf{W}_s + \varphi\mathbf{Z}_s\boldsymbol{\Omega}\mathbf{Z}_s') = \sigma^2\boldsymbol{\Sigma}_s$ and $\varphi = \sigma_u^2 / \sigma^2$.

3. Empirical Best Linear Unbiased Prediction

In this section we describe an algorithm for computing the empirical best linear unbiased estimate of $\boldsymbol{\beta}$ and the corresponding predictions of $\mathbf{u}_s$ and $\mathbf{u}_r$ given the area level data $\tilde{\boldsymbol{\theta}}_s$ and $\bar{\mathbf{X}}_{sw}$. This assumes that the parameter σ^2 in (3) is known (or that a good estimate of it is available), and is based on combining the ideas of Henderson (1950) and Harville (1977).

Put l_1 equal to the log-likelihood for $\boldsymbol{\beta}$ generated by $\tilde{\boldsymbol{\theta}}_s$ 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 best linear unbiased predictors (BLUPs) of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$ are then the values of these quantities where l is maximised (Henderson, 1950). Of course, these BLUPs depend on the variance components φ and $\boldsymbol{\lambda}$, so the empirical best linear unbiased predictors (EBLUPs) are obtained by substituting estimates for these parameters. In what follows, we describe an algorithm that calculates these EBLUPs by combining the above method for calculating the BLUPs of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$ with ML and REML estimation of the variance components.

Let $\begin{bmatrix} \boldsymbol{\Omega}_{ss} & \boldsymbol{\Omega}_{sr} \\ \boldsymbol{\Omega}_{rs} & \boldsymbol{\Omega}_{rr} \end{bmatrix}$ denote the partition of the variance-covariance matrix $\boldsymbol{\Omega}$ corresponding to the in sample and out of sample components of $\mathbf{u}$. Similarly, put

$$\mathbf{T}^* = \begin{bmatrix} \mathbf{T}^*_{ss} & \mathbf{T}^*_{sr} \\ \mathbf{T}^*_{rs} & \mathbf{T}^*_{rr} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s + \varphi^{-1}\boldsymbol{\Lambda}_{s|r} & -\varphi^{-1}\boldsymbol{\Lambda}_{s|r}\boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_{rr}^{-1} \\ -\varphi^{-1}\boldsymbol{\Omega}_{rr}^{-1}\boldsymbol{\Omega}_{rs}\boldsymbol{\Lambda}_{s|r} & \varphi^{-1}(\boldsymbol{\Omega}_{rr}^{-1} + \boldsymbol{\Omega}_{rr}^{-1}\boldsymbol{\Omega}_{rs}\boldsymbol{\Lambda}_{s|r}\boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_{rr}^{-1}) \end{bmatrix}^{-1}$$

where $\boldsymbol{\Lambda}_{s|r} = (\boldsymbol{\Omega}_{ss} - \boldsymbol{\Omega}_{sr}\boldsymbol{\Omega}_{rr}^{-1}\boldsymbol{\Omega}_{rs})^{-1}$. An iterative procedure for obtaining the ML estimates of φ and $\boldsymbol{\lambda}$ for given σ^2 is then:

1. Assign initial values to the variance components φ and $\boldsymbol{\lambda}$.
2. Using the current values for these variance components, calculate $\boldsymbol{\Omega}$.
3. Update $\boldsymbol{\beta} = (\bar{\mathbf{X}}_{sw}'\boldsymbol{\Sigma}_s^{*-1}\bar{\mathbf{X}}_{sw})^{-1}\bar{\mathbf{X}}_{sw}\boldsymbol{\Sigma}_s^{*-1}\tilde{\boldsymbol{\theta}}$

where $\boldsymbol{\Sigma}_s^{*-1} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{Z}_s\mathbf{T}_{ss}^{*}\mathbf{Z}'\mathbf{W}^{-1}{}_s$.

4. Update $\mathbf{u}_s = \mathbf{T}_{ss}^{*}\mathbf{Z}_s'\mathbf{W}_s^{-1}(\tilde{\boldsymbol{\theta}} - \bar{\mathbf{X}}_s\boldsymbol{\beta})$.
5. Update $\mathbf{u}_r = \mathbf{T}_{sr}^{*}\mathbf{Z}_s'\mathbf{W}_s^{-1}(\tilde{\boldsymbol{\theta}} - \bar{\mathbf{X}}_s\boldsymbol{\beta})$.
6. Update $\varphi = D_s^{-1}(\mathrm{tr}(\mathbf{T}_{ss}^{*}\boldsymbol{\Omega}_s^{-1}) + \sigma^{-2}\mathbf{u}_s'\boldsymbol{\Omega}_s^{-1}\mathbf{u}_s)$.
7. Check for convergence of the different estimates. If not return to step 2.
8. Update $\boldsymbol{\lambda} = f(\boldsymbol{\lambda}, \varphi, \mathbf{T}_{ss}^{*}, \sigma^2, \tilde{\mathbf{u}}_s)$ where f is the Fisher score or Newton-Raphson updating function for this parameter, i.e. a function whose specification depends on the parameterization of $\boldsymbol{\Omega}$, 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.

We denote the final values of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$ output by the above iterative process by $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}_s$ and $\hat{\mathbf{u}}_r$ respectively. These estimates are then substituted in (1) to give the ML-based EBLUP $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\boldsymbol{\theta}}_s \\ \hat{\boldsymbol{\theta}}_r \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{X}}_s\hat{\boldsymbol{\beta}} + \mathbf{Z}_s\hat{\mathbf{u}}_s \\ \bar{\mathbf{X}}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}_r \end{bmatrix}. \tag{4}$$

In order to define the REML-based EBLUP, we modify this iterative procedure to give the REML estimates of φ and $\boldsymbol{\lambda}$. Define

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

and let $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{bmatrix}$ and $\mathbf{V}^{-1} = \mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \mathbf{T}_{13} \\ . & \mathbf{T}_{22} & \mathbf{T}_{23} \\ . & . & \mathbf{T}_{33} \end{bmatrix}$ be the partitions of the matrix $\mathbf{V}$ and its inverse that correspond to the dimensions of $\boldsymbol{\beta}$, $\mathbf{u}_s$ and $\mathbf{u}_r$. Replacing $\mathbf{T}_{ss}^{*}$ by $\mathbf{T}_{22}$ in the iterative algorithm above leads to the REML estimates of the variance components, and hence to the REML-based EBLUP of $\boldsymbol{\theta}$.

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

We first obtain the mean cross-product errors matrix (MCPE) of the ML-based EBLUP estimator (4). This has prediction error $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \begin{bmatrix} \hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s \\ \hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r \end{bmatrix}$, with

$\mathrm{MCPE}(\hat{\boldsymbol{\theta}}) = \mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})']$. 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. Put $\boldsymbol{\omega} = (\varphi, \boldsymbol{\lambda}')'$. Then, following Prasad and Rao (1990) and after some algebra, we have

$$\mathrm{MCPE}(\hat{\boldsymbol{\theta}}) \cong \mathrm{MCPE}(\boldsymbol{\theta}_{BLUP}) + M_{\omega}(\sigma^2, \boldsymbol{\omega}) \tag{5}$$

where

$\mathrm{MCPE}(\boldsymbol{\theta}_{BLUP}) = \mathrm{E}[(\boldsymbol{\theta}_{BLUP} - \boldsymbol{\theta})(\boldsymbol{\theta}_{BLUP} - \boldsymbol{\theta})'] = M_{\beta}(\sigma^2, \boldsymbol{\omega}) + M_{\beta u}(\sigma^2, \boldsymbol{\omega}) + M_{u}(\sigma^2, \boldsymbol{\omega})$.

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 $\mathbf{u}$ and $M_{\omega}(\boldsymbol{\omega})$ measures the uncertainty due to estimation of the variance components $\boldsymbol{\omega}$. Using the general results set out in Henderson (1975), the first three components of (5) are given by

$$M_{\beta}(\sigma^2, \boldsymbol{\omega}) = \sigma^2 \bar{\mathbf{X}} \mathbf{T}_{11} \bar{\mathbf{X}}'$$

$$M_{u}(\sigma^2, \boldsymbol{\omega}) = \sigma^2 \mathbf{Z} \mathbf{T}^* \mathbf{Z}'$$

and

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

The final component $M_{\omega}(\sigma^2, \boldsymbol{\omega})$ is a measure of the uncertainty due to estimation of the variance components $\boldsymbol{\omega} = (\varphi \ \boldsymbol{\lambda}')'$ and is defined as follows. Put $\boldsymbol{\Delta} = \mathbf{Z} \mathbf{T}^* = [\boldsymbol{\Delta}'_1, \boldsymbol{\Delta}'_2, ..., \boldsymbol{\Delta}'_D]'$ and let $\mathbf{Z}_{\alpha}$ be the α^{th} row of the matrix $\mathbf{Z}$, so that $\partial \boldsymbol{\Delta}_{\alpha} / \partial \boldsymbol{\omega} = \partial (\mathbf{Z}_{\alpha} \mathbf{T}^*) / \partial \boldsymbol{\omega}$. Then

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

where $\boldsymbol{\Sigma}_s^* = \mathbf{Z}'_s \mathbf{W}_s^{-1} \mathbf{Z}_s + \varphi \mathbf{Z}'_s \mathbf{W}_s^{-1} \mathbf{Z}_s ½_s \mathbf{Z}'_s \mathbf{W}_s^{-1} \mathbf{Z}_s$ and ∇_{α} is the first D_s columns of the matrix $\partial \boldsymbol{\Delta}_{\alpha} / \partial \boldsymbol{\omega}$. Here **B** is the asymptotic variance-covariance matrix of the estimator of the variance components vector $\boldsymbol{\omega}$. An estimator of the MCPE matrix of the EBLUP $\hat{\boldsymbol{\theta}}$ is therefore

$$\widehat{MCPE}(\hat{\boldsymbol{\theta}}) = M_{\beta}(\sigma^2, \hat{\boldsymbol{\omega}}) + M_{\beta u}(\sigma^2, \hat{\boldsymbol{\omega}}) + M_{u}(\sigma^2, \hat{\boldsymbol{\omega}}) + 2M_{\omega}(\sigma^2, \hat{\boldsymbol{\omega}}) \qquad (6)$$

where $\hat{\boldsymbol{\omega}}$ is the ML estimate of the variance components vector $\boldsymbol{\omega}$.

In order to define the corresponding estimator of the MCPE matrix of the REML-based EBLUP, we replace the ML estimate $\hat{\boldsymbol{\omega}}$ and its asymptotic variance-covariance matrix, **B** by corresponding REML values, and evaluate $\mathbf{T}^*$ at the REML estimates of the various parameters. Substitution in (6) then yields the REML-based estimator of the MCPE matrix.

5. Simulation Results

In this section we use simulation to demonstrate how the new methodology outlined above provides improved results compared with existing EBLUP methods for small area estimation when not all areas of interest are represented in sample. The population values were generated from a linear mixed model with spatially correlated area random effects, defined by

$$y_{di} = 0.5 + x_{di} + u_d + e_{di}\,. \tag{7}$$

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

$$\sigma_u^2 \mathbf{\Omega}(\lambda) = \sigma_u^2 [(\mathbf{I}_D - \lambda\Psi)(\mathbf{I}_D - \lambda\Psi')]^{-1} \tag{8}$$

where $\mathbf{I}_D$ is an identity matrix of order D and Ψ is a known square matrix of the same order containing strictly positive weights. This is the SAR or simultaneous autoregressive model (Cressie, 1993). Let $\Psi_0 = [\psi_{0ij}]$ to be a $D \times D$ matrix with ψ_{0ij} = 1 if areas i and j are considered "spatial neighbours" and is zero otherwise. The weight matrix Ψ was then constructed by scaling the elements of Ψ_0 to have row sums equal to one. Two different versions for Ψ_0 were used, defined by the regional maps for Ghana and Nigeria provided in Cliff and Ord (1973). The x_{di} values were generated from a uniform distribution between 0 and 1 and were kept fixed throughout the simulations. Values of y_{di} were generated for D = 40 (Ghana) and D = 50 (Nigeria) with 90 population units per area. The first D_s = 30 (Ghana) and 40 (Nigeria) areas were taken to be in sample areas, with the remaining D_r = 10 areas considered as being out of

sample. Random samples of size n_d were taken from each in sample area, with n_d increasing with d. The population data were aggregated to yield area level means for both in sample and out of sample areas. The sample data from the in sample areas were also aggregated to produce corresponding sample means, which were then used to estimate the model parameters via REML. Note that under the SAR model (8) the updating equation for the parameter λ under the Fisher scoring method (see step 8 of the EBLUP estimation procedure described in section 3) is given by

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

where $b_1 = -0.5[\varphi^{-1}\sigma^{-2}\mathbf{u}_s'(\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\mathbf{u}_s + \varphi^{-1}\mathrm{tr}((\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\mathbf{T}_{22}^{*}) - \mathrm{tr}((\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\boldsymbol{\Omega}_{ss}^{-1})]$ and b_2 is the (2, 2) element of inverse of the information matrix for the estimators $\hat{\varphi}$ and $\hat{\lambda}$. Note that we fixed σ^2 at its true value throughout the estimation process. The information matrix of the REML estimators $\hat{\varphi}$ and $\hat{\lambda}$ is given by

$$\mathbf{I}_{REML} = \frac{1}{2}\begin{bmatrix} \varphi^{-2}(\nu - 2r_1) + \varphi^{-4}r_{11} & \varphi^{-1}(2k_1 - v_1 - \varphi^{-1}k_{11}) \\ . & (\varphi^{-2}k_{11}^{(11)} + v_{11} - 2k_1^{(11)}) \end{bmatrix}$$

where

$$r_1 = \varphi^{1}\mathrm{tr}(\boldsymbol{\Omega}_{ss}^{-1}\mathbf{T}_{22}),$$

$$r_{11} = \mathrm{tr}(\boldsymbol{\Omega}_{ss}^{-1}\mathbf{T}_{22}\boldsymbol{\Omega}_{ss}^{-1}\mathbf{T}_{22}),$$

$$k_1 = \varphi^{-1}\mathrm{tr}((\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\mathbf{T}_{22}),$$

$$v_1 = \mathrm{tr}((\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\boldsymbol{\Omega}_{ss}),$$

$$k_{11} = \varphi^{-1}\mathrm{tr}(\mathbf{T}_{22}(\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\mathbf{T}_{22}\boldsymbol{\Omega}_{ss}^{-1}),$$

$$v_{11} = \mathrm{tr}((\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\boldsymbol{\Omega}_{ss}(\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\boldsymbol{\Omega}_{ss}),$$

$$k_{11}^{(11)} = \mathrm{tr}(\mathbf{T}_{22}(\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda)\mathbf{T}_{22}(\partial\boldsymbol{\Omega}_{ss}^{-1}/\partial\lambda))$$

$$k_1^{(11)} = \varphi^{-1}\mathrm{tr}[(\partial\boldsymbol{\Omega}_{ss}^{-1} / \partial\lambda)\mathbf{T}_{22}(\partial\boldsymbol{\Omega}_{ss}^{-1} / \partial\lambda)\boldsymbol{\Omega}_{ss}] .$$

Given the above set up, we considered four ways of defining the small area estimates. The first corresponded to a synthetic estimation procedure, where the mixed model defined by (7) and (8) is first fitted to the sample data, but then estimation is carried out on the basis that $u_d = 0$ in every small area. We refer to this as method A below. The second also fits (7) and (8) to the sample data, but forces $\lambda = 0$ in (8), i.e. this method assumes there is no spatial correlation among the area effects. Estimation then incorporates predicted area effects for in sample areas, but sets these to zero for out of sample areas. This is denoted method B in what follows. In contrast, the third method takes account of the correlation between areas when estimating the model parameters. However, it still sets the predicted value of u_d to zero for out of sample areas. We refer to this as method C. Finally, the fourth method, denoted D, corresponds to the EBLUP procedure defined earlier in this paper.

The process of generating population and sample data, estimation of model parameters and calculation of (A) – (D) was independently replicated 3000 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

$$ActMSE_d = diag_d\left(\sum_{k=1}^{3000}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)' / 3000\right)$$

$$EstMSE_d = diag_d\left(\sum_{k=1}^{3000}\widehat{MCPE}(\hat{\boldsymbol{\theta}}_k) / 3000\right)$$

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

$$ActCV_d = 100 \times \frac{ActMSE_d}{\sum_{k=1}^{3000} \theta_{dk} / 3000}$$

and the estimated coefficient of variation

$$EstCV_d = 100 \times \frac{EstMSE_d}{\sum_{k=1}^{3000} \hat{\theta}_{dk} / 3000}$$

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.

Information about the various simulation scenarios considered, including average sample sizes and true values of the variance components, is provided in Table 1. Thirteen different combinations of overall sample sizes and parameter values (Par1 – Par13) in (7) and (8) were used in the simulations. Table 2 shows the average values of both the actual coefficient variation (*ActCV*) and estimated coefficient of variation (*EstCV*) for the four estimation methods we considered. These show that for Method A 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 B and Method C, where now it is out of sample areas whose estimated CVs tend to be far too optimistic. The results also indicate that Method C performs better than Method B for both in sample and out of sample areas. Method D – as one would expect – performs much better in this regard, with estimated and actual CVs for both in sample and out of sample areas being very close. Note also

that average values of *ActCV* for Methods B, C and D in Table 2 are very similar for small values of λ, but use of Method D leads to substantial gains in efficiency 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 prediction intervals generated under Method A generally lead to severe undercoverage because they are based on conditionally biased synthetic estimators. In contrast, intervals generated under Method B has good coverage for in sample areas, but a very poor coverage for out of sample areas (even when there is no spatial correlation), reflecting this method's use of conditionally biased synthetic estimators for out of sample areas. There also seems to be some evidence that this coverage gets worse as this spatial correlation increases. The same pattern applies for the results generated by Method C, with somewhat better coverage for out of sample areas. On the other hand, Method D 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.

Table 3 also reports average confidence interval widths under Methods A – D. As expected, there is very little difference between Methods C and D for in sample areas. For such areas, and in particular for large values of σ_u^2 and λ, Method A results in very

wide confidence intervals. This reflects large standard error estimates under Method A for these values of σ_u^2 and λ. However, in spite of these large estimated standard errors, prediction interval coverages are very far from the nominal 95% level because of the biased estimators under this method. Small differences in covariate values for in sample and out of sample areas leads to insignificant differences in average confidence interval width for Method A. However, the three other methods considered, especially Method D, lead to significant differences in average confidence interval width between in and out of sample areas in such situations.

Finally, in Figure 1 we show the variation in relative bias for Methods A, B and D by area for both in and out of sample areas for two parameter sets (par3 and par 11). These confirm the overall superiority of Method D.

6. Summary and Discussion

In this article we develop EBLUP estimates for small area means under a Fay-Herriot type model when there are no sample units in the area. The model assumes spatially correlated area effects defined by the SAR model (8). 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 out of sample area prediction. 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. Note these conclusions are based on

simulation results just by leaving 10 areas out of sample, out of a total of 40 (Ghana scenario) and 50 (Nigeria scenario). We anticipate that the better performance of Method D relative to that of Methods A – C will become even more clear when the proportion of out of sample areas increases.

Note that our method makes the usual area level data assumption (e.g. Rao and Yu, 1994) that σ^2 is known. Recently, Wang and Fuller (2003) have investigated methods for fitting area level models without this assumption, and work is underway to see whether their ideas can be applied here.

The analysis in this paper has been restricted to the area level linear mixed model. Many applications, however, are based on 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. Application to other spatial correlation models, e.g. the conditional autogressive (CAR) model, is also of interest.

References

Butar, F.B. and Lahiri, P., 2003. On measures of uncertainty of empirical Bays small-area estimators. Journal of Statistical Planning and Inference 112, 63 – 76.

Cliff, A.D. and Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.

Cressie, N., 1993. Statistics for Spatial Data. John Wiley, New York.

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.

Fay, R.E. and Herriot, R.A., 1979. Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74, 269 – 277.

Ghosh, M. and Rao, J.N.K., 1994. Small area estimation: an appraisal. Statistical Science 9, 55 - 93.

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., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423-447.

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

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

Rao, J.N.K. and Yu, M., 1994. Small area estimation by combining time-series and cross-section data. The Canadian Journal of Statistics 22, 511 - 528.

Saei, A. and Chambers, R., 2005. Empirical best linear unbiased prediction for out of sample areas. Working Paper M05/03, Southampton Statistical Sciences Research Insitute, University of Southampton, UK.

Wang, J. and Fuller, W.A., 2003. The mean squared error of small area predictors constructed with estimated area variances. Journal of the American Statistical Association 98, 716 – 723.

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	7.93	1	5.0	0.00	30	10
Par2	7.93	1	5.0	0.60	30	10
Par3	10.67	1	5.0	0.92	30	10
Par4	8.53	1	0.5	0.60	30	10
Par5	7.37	1	0.5	0.92	30	10
Par6	7.33	1	1.5	0.60	30	10
Par7	7.17	1	1.5	0.92	30	10
Par8	8.05	1	1.5	0.60	40	10
Par9	8.05	1	1.5	0.92	40	10
Par10	9.13	1	5.0	0.60	40	10
Par11	8.05	1	5.0	0.92	40	10
Par12	7.18	1	0.5	0.60	40	10
Par13	8.43	1	0.5	0.92	40	10

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							
		A		B		C		D	
		ActCV	*EstCV*	*ActCV*	*EstCV*	*ActCV*	*EstCV*	*ActCV*	*EstCV*
All	Par1	226.02	57.09	100.56	56.25	101.45	58.60	101.93	102.31
	Par2	273.12	86.04	120.95	59.88	116.23	65.55	114.87	116.01
	Par3	689.76	161.75	322.43	90.84	253.10	94.24	205.05	216.89
	Par4	87.11	28.89	51.48	32.58	49.88	34.21	49.56	50.68
	Par5	182.53	47.66	91.67	44.32	74.26	43.59	66.58	72.67
	Par6	149.42	47.49	75.90	43.09	73.10	46.03	72.46	73.98
	Par7	280.52	69.05	127.58	52.64	101.72	50.79	87.51	94.80
	Par8	143.70	40.12	65.92	40.92	64.13	42.56	62.29	61.29
	Par9	286.52	63.53	99.59	51.15	88.28	47.92	65.00	64.99
	Par10	268.63	73.65	98.61	51.45	94.59	54.14	90.83	89.14
	Par11	486.06	107.44	154.11	71.45	131.80	63.35	94.62	94.32
	Par12	83.00	24.37	48.17	34.51	46.91	35.00	46.02	45.48
	Par13	160.30	39.15	68.01	40.73	60.53	38.94	49.14	49.36
In sample	Par1	221.52	56.66	53.88	53.81	54.60	53.09	54.62	53.09
	Par2	267.41	85.95	59.10	58.61	54.71	53.75	54.67	53.65
	Par3	488.26	115.57	72.89	70.28	46.98	47.82	45.92	46.03
	Par4	85.74	28.90	37.54	37.62	36.11	35.99	36.13	35.93
	Par5	186.34	48.05	50.01	49.78	41.98	42.62	41.44	41.79
	Par6	147.31	47.55	47.62	47.96	45.54	45.59	45.56	45.50
	Par7	283.30	69.40	56.81	56.13	44.90	45.05	44.17	44.16
	Par8	142.99	40.15	45.19	45.07	43.53	43.20	43.54	43.14
	Par9	291.54	63.75	54.64	53.87	43.74	44.24	42.86	42.90
	Par10	267.18	73.58	53.29	53.13	49.63	49.20	49.66	49.11
	Par11	495.03	108.01	73.57	72.30	52.21	52.90	51.02	50.96
	Par12	83.07	24.45	39.30	39.31	37.95	37.73	37.94	37.68
	Par13	163.06	39.28	45.06	44.93	38.34	39.01	37.65	37.93
Out of sample	Par1	230.1	55.17	228.77	42.94	230.10	55.17	232.33	234.00
	Par2	285.43	84.29	292.23	49.02	285.43	84.29	280.30	288.56
	Par3	483.00	115.24	594.05	74.66	483.00	115.24	391.86	439.70
	Par4	91.19	28.86	93.31	17.49	91.19	28.86	89.85	94.91
	Par5	171.08	46.51	216.66	27.93	171.08	46.51	142.00	165.32
	Par6	155.75	47.34	160.73	28.47	155.75	47.34	153.18	159.40
	Par7	272.2	68.02	339.88	42.17	272.2	68.02	217.54	246.71
	Par8	146.55	39.99	148.87	24.31	146.55	39.99	137.32	133.89
	Par9	266.44	62.65	279.39	40.27	266.44	62.65	153.56	153.37
	Par10	274.43	73.94	279.91	44.71	274.43	73.94	255.54	249.28
	Par11	450.16	105.15	476.29	68.05	450.16	105.15	269.02	267.76
	Par12	82.76	24.07	83.64	15.30	82.76	24.07	78.35	76.67
	Par13	149.27	38.64	159.80	23.90	149.27	38.64	95.09	95.06

Table 3 Confidence interval width and 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*	*Confidence interval width*				*95%Coverage*			
		A	B	C	D	A	B	C	D
All	Par1	2.16	2.13	2.21	3.93	36.48	78.89	80.10	94.76
	Par2	3.19	2.26	2.47	4.37	43.77	77.50	81.73	94.29
	Par3	4.81	2.62	2.73	5.87	36.82	74.41	80.15	94.23
	Par4	1.09	1.26	1.32	1.94	46.33	78.26	82.22	94.23
	Par5	1.71	1.59	1.58	2.64	40.10	75.94	81.73	94.69
	Par6	1.78	1.65	1.76	2.82	44.54	77.89	81.91	94.29
	Par7	2.77	2.06	2.04	3.86	37.23	75.66	80.73	94.73
	Par8	1.53	1.59	1.65	2.39	39.85	80.79	83.65	94.43
	Par9	2.50	1.98	1.90	2.59	33.39	79.90	83.06	94.71
	Par10	2.74	1.91	2.02	3.35	39.13	80.74	83.68	94.43
	Par11	4.28	2.70	2.51	3.80	33.29	79.76	82.98	94.64
	Par12	0.94	1.36	1.38	1.80	41.81	81.39	84.11	94.33
	Par13	1.49	1.55	1.49	1.90	36.80	80.44	83.90	94.76
In sample	Par1	2.16	2.03	2.01	2.01	36.73	94.71	94.3.0	94.29
	Par2	3.22	2.16	2.00	2.00	44.86	94.64	94.47	94.44
	Par3	4.86	2.67	1.94	1.88	35.79	94.19	94.95	94.84
	Par4	1.09	1.46	1.40	1.39	46.83	94.87	94.68	94.62
	Par5	1.71	1.79	1.54	1.51	39.56	94.81	95.08	94.96
	Par6	1.78	1.83	1.75	1.74	45.02	95.00	94.85	94.78
	Par7	2.77	2.18	1.80	1.77	36.97	94.72	94.98	94.98
	Par8	1.53	1.75	1.68	1.68	39.97	94.83	94.73	94.71
	Par9	2.50	2.08	1.75	1.70	32.96	94.63	95.05	94.89
	Par10	2.74	1.97	1.84	1.84	39.06	94.83	94.75	94.70
	Par11	4.28	2.69	2.07	2.01	32.89	94.47	95.00	94.81
	Par12	0.94	1.55	1.49	1.48	41.79	94.84	94.66	94.64
	Par13	1.49	1.71	1.49	1.46	36.28	94.94	95.16	94.99
Out of sample	Par1	2.09	1.66	2.09	9.01	35.57	28.91	35.57	94.18
	Par2	3.17	1.88	3.17	10.89	42.32	25.66	42.32	93.18
	Par3	4.87	3.10	4.87	18.24	36.24	18.75	36.24	93.95
	Par4	1.09	0.67	1.09	3.57	44.85	28.42	44.85	93.07
	Par5	1.71	1.02	1.71	6.03	41.71	19.31	41.71	93.86
	Par6	1.79	1.10	1.79	6.05	43.10	26.55	43.10	92.83
	Par7	2.77	1.69	2.77	10.15	38.00	18.51	38.00	93.99
	Par8	1.53	0.94	1.53	5.20	39.36	24.65	39.36	93.28
	Par9	2.49	1.58	2.49	6.17	35.11	21.01	35.11	93.98
	Par10	2.74	1.68	2.74	9.43	39.40	24.36	39.40	93.35
	Par11	4.28	2.72	4.28	11.00	34.87	20.95	34.87	93.94
	Par12	0.95	0.61	0.95	3.05	41.88	27.58	41.88	93.09
	Par13	1.49	0.91	1.49	3.67	38.85	22.46	38.85	93.83

Figure 1 Relative bias by area for two parameter sets (top = Par3, bottom = Par11). Dot-dash line is Method A, dashed line is Method B and solid line is Method D.