F35

A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation – An Application to Crop Yield Estimation

Hukum Chandra
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi, India
Email: hchandra12@gmail.com

Nicola Salvati
Dipartimento di Economia e Management
University of Pisa
Pisa, Italy

Ray Chambers
University of Wollongong
Wollongong, Australia
National Institute for Applied Statistical Research Australia
University of Wollongong, NSW, 2522, Australia

U.C. Sud
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi, India

DOI: 10.1481/icasVII.2016.f35

ABSTRACT

In agricultural survey data, relationship between study variable (e.g., crop yield) and covariates may not be same over the study area, this phenomenon is referred to as spatial nonstationarity. Small area estimates based on the widely-used area-level model proposed in Fay and Herriot (1979) assume that the area level direct estimates are spatially nonstationary. We propose an extension to the Fay-Herriot model that accounts for the presence of spatial nonstationarity in the area level data. We refer to the predictor based on this extended model as the nonstationary empirical best linear unbiased predictor (NSEBLUP). We also develop two different estimators for the MSE of the NSEBLUP. The first estimator uses approximations similar to those in Opsomer et al. (2008). The second estimator is based on the parametric bootstrapping approach. Results from simulation studies using spatially nonstationary data indicate that the NSEBLUP compares favourably with alternative area-level predictors that ignore this spatial nonstationarity. Both of the proposed methods of MSE estimation for the NSEBLUP seem to perform adequately. Developed small area estimation method is applied to produce district level estimates of crop yield in the State of Uttar Pradesh using the data on crop cutting experiments supervised under Improvement of Crop Statistics scheme (data collected with much reduced sample size, however, the quality of data is very high) and the secondary data from the Census. The results show a considerable gain in precision in estimates produced applying small area estimation. These estimates will provide invaluable information to policy-analysts and decision-makers.

Keywords: District level estimates, small area estimation, spatial nonstationarity, geographical weighted regression, Census

1. Introduction

Sample surveys are usually designed so that direct estimators for larger domains lead to reliable estimates, where by direct estimators here we mean estimators that use only the sample data from the domain of interest. However, direct estimation is typically inefficient for smaller domains where sample sizes can be small, and cannot be used when there are no sample units in the domain. Following standard practice, we refer to these smaller domains as 'small areas' or just 'areas' from now on. Indirect (i.e. model-based) small area estimation (SAE) techniques are now widely employed to produce estimates and measures of precision for these small areas. In this context, we differentiate between SAE methods based on unit-level models and those based on area-level models. In the former case these models are for the individual survey measurements and include area effects, while in the latter case these models are used to smooth out the variability in the unstable area-level direct estimates. Area-level modelling is typically used when unit-level data are unavailable, or, as is often the case, where model covariates (e.g. census variables) are only available in aggregate form. Fay and Herriot (1979) proposed an area-level SAE model (hereafter the FH model) that relates small area direct survey estimates to area-level covariates. The FH model is widely used because of its flexibility in combining different sources of information with different error structures, and can be described as follows. Let *i* index the *m* areas of interest and let y_i be an unbiased direct survey estimator of an unobservable population parameter (for example, the population mean) Y_i of a variable of interest *y* for area *i*. Let $\mathbf{z}_i$ be a vector of *q* auxiliary variables for area *i* that are related to the population mean Y_i. These variables are typically obtained from administrative and census records. The FH model is then defined by the two equations

$$y_i - Y_i = e_i \text{ and } Y_i - \theta - \mathbf{z}_i^T \boldsymbol{\lambda} = u_i, \qquad (1)$$

where the first equation models the prediction error of the observed survey estimate y_i of the true area *i* population mean Y_i, while the second models the unobservable Y_i in terms of an overall mean θ and a linear combination of the components of the vector $\mathbf{z}_i$, and is such that the area effects u_i satisfy $E(u_i | \mathbf{z}_i) = 0$. Put $\mathbf{x}_i^T = (1, \mathbf{z}_i^T)$ with $p = q + 1$ equal to the dimension of $\mathbf{x}_i$. Combining these two equations then leads to an area level linear mixed model of form

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i\, ;\ i = 1, \ldots, m\, . \qquad (2)$$

Here $\boldsymbol{\beta} = (\theta, \boldsymbol{\lambda}^T)^T$ is a *p*-vector of unknown fixed effect parameters, the area effects u_i are independently and identically distributed, with $E(u_i|\mathbf{z}_i)=0$ and $\text{Var}(u_i|\mathbf{z}_i)=\sigma_u^2$, and the prediction errors e_i are independently distributed, with $E(e_i|\mathbf{z}_i)=0$ and $\text{Var}(e_i|\mathbf{z}_i)=\sigma_{ei}^2$. The area effects and the prediction errors are assumed to be independent of each other within and across areas. An important additional assumption that is usually made is that the prediction variances σ_{ei}^2 are known.

Since the parameters $\boldsymbol{\beta}$ and σ_u^2 are the same for every area, they can be estimated using the data from all *m* areas. This is usually accomplished by 'stacking' the area level direct estimates to produce an overall area level mixed model of the form

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}+\mathbf{e}, \tag{3}$$

where $\mathbf{y}=(y_1,...,y_m)^T$ is the $m\times 1$ vector of direct survey estimates, $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_m)^T$ is the $m\times p$ matrix whose *i*-th row is given by $\mathbf{x}_i^T$, $\mathbf{u}=(u_1,...,u_m)^T$ is the *m*-vector of random area effects and $\mathbf{e}=(e_1,...,e_m)^T$ is the *m*-vector of prediction errors. This model can be generalised by replacing **u** in (3) by **Du**, where **D** is diagonal matrix of dimension $m\times m$ of area-specific covariates that can be used to characterise heteroskedasticity in the area effects. In the interests of avoiding unnecessary notational complexity, we ignore this generalisation here. It is assumed that the vector of area effects **u** is distributed independently of the prediction errors $\mathbf{e}$, so that the covariance matrix of the vector $\mathbf{y}$ is given by $\text{Var}(\mathbf{y})=\mathbf{V}=\sigma_u^2\mathbf{I}_m+\boldsymbol{\Sigma}_e$, where $\mathbf{I}_m$ is the identity matrix of order *m* and $\boldsymbol{\Sigma}_e=diag\{\sigma_{ei}^2;1\le i\le m\}$ is the known matrix of prediction variances. The parameters σ_u^2 and $\boldsymbol{\Sigma}_e$ are sometimes referred to as the variance components of (3). Under the assumption that **u** is Gaussian, σ_u^2 can be estimated using maximum likelihood (ML) or restricted maximum likelihood (REML). Let $\hat{\sigma}_u^2$ denote the resulting estimator of σ_u^2 and define the plug-in estimator $\hat{\mathbf{V}}=\hat{\sigma}_u^2\mathbf{I}_m+\boldsymbol{\Sigma}_e=diag\{\hat{\sigma}_u^2+\sigma_{ei}^2;i=1,\ldots,m\}$ of the covariance matrix $\mathbf{V}$. Under (3), the empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$ and the empirical best linear unbiased predictor (EBLUP) of $\mathbf{u}$ are then

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y} \tag{4}$$

and

$$\hat{\mathbf{u}}=\hat{\sigma}_u^2\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}), \tag{5}$$

respectively. Under (2), the EBLUP estimate of Y_i is (Henderson, 1975; Fay & Herriot, 1979)

$$\hat{Y}_i=\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\boldsymbol{\delta}_i^T\hat{\sigma}_u^2\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\hat{u}_i, \tag{6}$$

where $\boldsymbol{\delta}_i^T$ denotes the i^{th} row of $\mathbf{I}_m$. Note that the corresponding EBLUP of the area effect u_i is therefore $\hat{u}_i=\hat{\gamma}_i(y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$, where $\hat{\gamma}_i=\hat{\sigma}_u^2\left(\hat{\sigma}_u^2+\sigma_{ei}^2\right)^{-1}$ defines the shrinkage effect for area *i*. See Rao (2003, chapter 5) for further details.

In practice areas are unplanned domains, and so many of them have zero sample sizes. These areas are referred to as non-sampled areas. The conventional approach for estimating area means in this case is synthetic estimation (Rao, 2003, pp. 46), based on a suitable model fitted to the data from the sampled areas. Let $\mathbf{z}_{j,out}$ denote the vector of covariates associated with non-sampled area *j*, and put $\mathbf{x}_{j,out}^T=(1,\mathbf{z}_{j,out}^T)$. Under model (2), the synthetic EBLUP predictor for the unknown population value $Y_{j,out}$ of area *j* is then

$$\hat{Y}_{j,out}=\mathbf{x}_{j,out}^T\hat{\boldsymbol{\beta}} \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ is given by (4). We refer to this predictor as SYN in what follows.

Model (2) implicitly assumes that direct estimates from different areas are uncorrelated. However the boundaries that define an area are typically arbitrarily set, and there appears to be no good reason why neighbouring areas should not be correlated. This can be the case, for example, with agricultural, environmental, economic and epidemiological data. It is therefore often reasonable to assume that the effects of neighbouring areas, defined via a contiguity criterion, are correlated. Cressie (1991), Singh *et al.* (2005) and Pratesi and Salvati (2008) extend the mixed model (3) to allow for spatially correlated random effects using conditional autoregressive (CAR) and simultaneous autoregressive (SAR) specifications for **u** (Anselin, 1992). These models allow for spatial correlation in the area effects, while keeping the fixed effects parameters spatially invariant. Under the area level version of this spatial mixed model, Singh *et al.* (2005) and Pratesi and Salvati (2008) define the spatial empirical best linear unbiased predictor (SEBLUP) for a small area mean and also derive an approximately unbiased estimator of the MSE of the SEBLUP.

An alternative approach to incorporating spatial information in SAE is to assume that the model for $E(Y_i|\mathbf{z}_i)$ varies spatially. There are currently two approaches to specifying such a model. The first uses a spatially varying surface to model the mean structure of (2). For example, Giusti *et al.* (2012) extend the unit level nonparametric spatial spline approximation of Opsomer *et al.* (2008) in order to define a spatially non-linear area level model. These authors then develop the corresponding area-level nonparametric empirical best linear unbiased predictor (NPEBLUP) for a small area mean. A key feature of this approach is that it assumes that the regression parameters associated with the model do not vary spatially. Instead, spatial variability is accommodated by adding spatially varying covariates to the model specification. There are situations, however, where this assumption is inappropriate, a phenomenon referred to as spatial nonstationarity, see for example Brunsdon *et al.* (1996) and the references therein. The second approach therefore replaces the global regression model (2) by one where the regression specification varies locally. Such a model can be fitted using geographically weighted regression (GWR), a method that is widely used for data exhibiting spatial nonstationarity (Brunsdon *et al*., 1996, Fotheringham *et al*., 2002). Note that the model underpinning GWR is a local linear model, i.e. a linear model for the conditional expectation of y given **z** at a specified location. Under GWR the data are assumed to follow a location specific linear regression function, with geographically defined weights used to estimate the parameters of this local regression function. We use the GWR concept to extend the FH model (2) to spatially nonstationary area level data. We refer to this extended model as a spatially nonstationary FH model, and investigate its suitability for SAE with area level data that exhibit spatial nonstationarity. Note that we have not provided simulation results, bootstrap approach of MSE estimation and many other analytical details in this paper. Readers are suggested to refer Chandra *et al*. (2015) for these details. The rest of paper is organized as follows. The nonstationary version of the area level linear mixed model and the estimator of a small area mean under this model are described in Section 2. Section 3 presents the theoretical expression for mean squared error of this predictor and an estimator for this mean squared error. Given that the approach is for the spatially nonstationary situation, a bootstrap procedure to test for the presence of spatial nonstationarity is also proposed in this Section. Section 4 presents an application of proposed method in a real data from agriculture survey to produce the crop yield estimates at small area level. Finally, Section 5 discusses concluding remarks.

2. A Spatially Nonstationary Area Level Model

Under (1), the parameters making up the vector $\boldsymbol{\lambda}$ are spatially invariant, i.e. the expected value of Y_i given $\mathbf{z}_i$ is the same at any two points in the study area that have the same set of values for this covariate. However, there are situations, for example in agricultural and environmental data, where this relationship is not constant, i.e. where there is spatial nonstationarity in the area level population parameters. In order to accommodate this situation, we now define a spatially nonstationary version of the FH model. To start, let the spatial location of area *i* correspond to the coordinates of an arbitrarily defined spatial location in the area, e.g. its centroid, which we denote by loc_i. Let $d(loc_i, loc_j)$ be an appropriate measure of the distance between the spatial locations of areas *i* and *j*, and define the spatial contiguity of these two locations to be $\omega_{ij} = \left(1 + d(loc_i, loc_j)\right)^{-1}$. Let $\mathbf{\Omega} = \left[\omega_{ij}\right]$ denote the positive definite $m \times m$ matrix of spatial contiguities defined by the loc_i. This matrix is assumed to be known. We consider a spatially nonstationary extension of the FH model (1) for area *i* of the form

$$y_i - Y_i = e_i \quad \text{and} \quad Y_i - \theta(loc_i) - \mathbf{z}_i^T \boldsymbol{\lambda}(loc_i) = u_i \tag{8}$$

where $\left(\theta(loc_i), \boldsymbol{\lambda}^T(loc_i)\right)^T = \left(\theta, \boldsymbol{\lambda}^T\right)^T + \boldsymbol{\gamma}(loc_i)$ and $\boldsymbol{\gamma}(loc) = \left(\gamma_k(loc); k = 1, \ldots, p\right)$ is a spatially varying multivariate random process of dimension p. Put $\mathbf{y} = \left(y_1, \ldots, y_m\right)^T$, $\mathbf{z} = \left(\mathbf{z}_1^T, \ldots, \mathbf{z}_m^T\right)^T$, $\mathbf{\Gamma} = \left(\boldsymbol{\gamma}^T(loc_1), \ldots, \boldsymbol{\gamma}^T(loc_m)\right)^T$ and $\mathbf{loc} = \left\{loc_1, \ldots, loc_m\right\}$, i.e. the set of locations for the *m* areas. Then

$$\mathrm{E}\left(\mathbf{\Gamma} \middle| \mathbf{z}, \mathbf{loc}\right) = \mathbf{0}_{pm\times 1}.$$

Since a general specification for the covariance structure of $\boldsymbol{\gamma}(loc)$ is complex, we follow Datta *et al.* (1998) and assume a separable working model for the second order spatial moments of $\boldsymbol{\gamma}(loc)$. This is a model of the form

$$\mathrm{Var}\left(\Gamma \middle| \mathbf{z}, \mathbf{loc}\right) = \mathbf{\Sigma} = \mathbf{\Omega} \otimes \mathbf{C} \tag{9}$$

where $\mathbf{C} = \left[c_{kl}\right]$ is a $p \times p$ covariance matrix that characterises the correlations between the components of $\boldsymbol{\gamma}$ at an arbitrary location loc and $\otimes$ denotes Kronecker product. Under (8) and (9),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{\Psi}\mathbf{\Gamma} + \mathbf{u} + \mathbf{e} \tag{10}$$

where $\boldsymbol{\beta} = \left(\theta, \boldsymbol{\lambda}^T\right)^T$ and $\mathbf{\Psi} = diag\left(\mathbf{x}_i^T; i = 1, \ldots, m\right)$. Then

$$\mathrm{E}\left(\mathbf{y} \middle| \mathbf{z}, \mathbf{loc}\right) = \mathbf{X}\boldsymbol{\beta}$$

$$Var\left(\mathbf{y} \middle| \mathbf{z}, \mathbf{loc}\right) = \mathbf{V} = \mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Psi}^T + \sigma_u^2 \mathbf{I}_m + diag\left\{\sigma_{ei}^2; i = 1, \ldots, m\right\} \tag{11}$$

$$\mathrm{Cov}\left(Y_i, \mathbf{y} \middle| \mathbf{z}, \mathbf{loc}\right) = \mathbf{\Psi}_i \mathbf{\Sigma}\mathbf{\Psi}^T + \sigma_u^2 \boldsymbol{\delta}_i^T \tag{12}$$

where $\mathbf{\Psi}_i$ is the i^{th} row of $\mathbf{\Psi}$. The minimum mean squared error (MMSE) predictor of Y_i under (8) - (12) is the expected value of Y_i given **y**, **z** and **loc**. Under a Gaussian errors assumption, and assuming the inverse of **V** exists, this is

$$\tilde{Y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathrm{Cov}\left(Y_i, \mathbf{y} \middle| \mathbf{z}, \mathbf{loc}\right) \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \tag{13}$$

The vector of MMSE predictors for the population values in **Y** is therefore

$$\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \left(\mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Psi}^T + \sigma_u^2 \mathbf{I}_m\right) \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \tag{14}$$

where **V** is defined by (11). The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is then

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, \tag{15}$$

and the best linear unbiased predictors (BLUPs) of $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}(loc_1)^T, \ldots, \boldsymbol{\gamma}(loc_m)^T\right)^T$ and $\mathbf{u}$ are

$$\tilde{\boldsymbol{\gamma}} = \mathbf{\Sigma}\mathbf{\Psi}^T \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) \tag{16}$$

$$\tilde{\mathbf{u}} = \sigma_u^2 \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right). \tag{17}$$

The corresponding BLUP of the vector $\mathbf{Y}$ is $\tilde{\mathbf{Y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\Psi}\tilde{\boldsymbol{\gamma}} + \tilde{\mathbf{u}}$. In particular, for small area *i*, the BLUP of Y_i , which we refer to below as the nonstationary BLUP, or NSBLUP, is

$$\tilde{Y}_i = \mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + \left(\boldsymbol{\Psi}_i\boldsymbol{\Sigma}\boldsymbol{\Psi}^T + \sigma_u^2\boldsymbol{\delta}_i^T\right)\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) = \mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + \boldsymbol{\Psi}_i\tilde{\boldsymbol{\gamma}} + \tilde{u}_i. \tag{18}$$

In practice the parameters σ_u^2 and $\mathbf{C}$ are unknown and have to be estimated from the data. Replacing these unknown parameters by their estimated values ^{2}
plug-in estimators by a 'hat', we obtain the empirical BLUE (EBLUE) of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right]^{-1}\left[\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}\right]$ and the empirical BLUP (EBLUP) of Y_i as

$$\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}_i\hat{\boldsymbol{\gamma}} + \hat{u}_i. \tag{19}$$

We refer to (19) as the nonstationary EBLUP, or NSEBLUP, of Y_i.

We now consider synthetic prediction for non-sampled areas under (10). Suppose that there are m_{out} of these non-sampled areas, indexed by *j*. We assume that the covariate vectors $\mathbf{z}_{out,j}$ and the spatial locations $loc_{out,j}$ (e.g. the centroids) of these areas are known. Let $\mathbf{z}_{out}$ and $\mathbf{loc}_{out}$ denote these auxiliary data. The spatially nonstationary predictor (denoted by NSSYN) of the vector of population values $\mathbf{Y}_{out}$ for the non-sampled areas is the plug-in estimator of the MMSE predictor of $\mathbf{Y}_{out}$ given **y**, **z**, **loc**, $\mathbf{z}_{out}$ and $\mathbf{loc}_{out}$,

$$\hat{\mathbf{Y}}_{out} = \mathbf{X}_{out}\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}_{out}\hat{\boldsymbol{\gamma}}_{out}$$

where $\mathbf{X}_{out}$ is the $m_{out} \times p$ matrix of covariates for the non-sampled areas,

$$\boldsymbol{\Psi}_{out} = diag\left\{\mathbf{x}_{j,out}^T; j = 1,\ldots,m_{out}\right\},\ \hat{\boldsymbol{\gamma}}_{out} = \hat{\boldsymbol{\Sigma}}_{out/in}\boldsymbol{\Psi}^T\hat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \text{ with } \hat{\boldsymbol{\Sigma}}_{out/in} = \boldsymbol{\Omega}_{out/in} \otimes \hat{\mathbf{C}}.$$

Here $\boldsymbol{\Omega}_{out/in}$ is the known $m_{out} \times m$ matrix of spatial contiguities between the non-sampled areas and the sampled areas. In particular, for non-sampled area *j*, we have

$$\hat{Y}_{j,out} = \mathbf{x}_{j,out}^T\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}_{j,out}\hat{\boldsymbol{\gamma}}_{out} \tag{20}$$

where $\boldsymbol{\Psi}_{j,out}$ denotes the row of $\boldsymbol{\Psi}_{out}$ corresponding to this non-sampled area. We refer to the predictor (20) as the NSSYN predictor. In contrast to the SYN predictor (7), the NSSYN predictor uses the location data for the non-sampled areas to 'borrow strength' from neighbouring sampled areas, and so has the potential to improve conventional synthetic prediction for non-sampled areas. In particular, we expect that if in fact the population data exhibit spatial nonstationarity, then the NSSYN predictor (20) will exhibit less bias than the standard SYN predictor (7).

2.1 Parameter estimation

In what follows we restrict our development to the simple single parameter specification $\mathbf{C} = \eta\mathbf{I}_p$ for the matrix $\mathbf{C}$, where $\mathbf{I}_p$ denotes the identity matrix of order *p*. That is, the components of the random vector $\boldsymbol{\gamma}$ are uncorrelated at any particular location, with the parameter $\eta \geq 0$ reflecting the strength or 'intensity' of spatial clustering in the data, and $\eta = 0$ corresponding to the situation where the model is spatially homogeneous (no spatial correlation in Y_i). In this case there are just 2 parameters (η and σ_u^2) that need to be estimated. This can be done by maximising a restricted maximum likelihood under a Gaussian assumption. Put $\boldsymbol{\varphi} = \left\{\eta, \sigma_u^2\right\} = \left(\varphi_1, \varphi_2\right)$. Under (10), the restricted log-likelihood function is then

$$l\left(\boldsymbol{\varphi}\right) = const - \frac{1}{2}\log\left|\mathbf{P}^T\mathbf{V}\mathbf{P}\right| - \frac{1}{2}\left(\mathbf{P}^T\mathbf{y}\right)^T(\mathbf{P}^T\mathbf{V}\mathbf{P})^{-1}\left(\mathbf{P}^T\mathbf{y}\right) \tag{21}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$ and, for $s = 1,2$

$$\frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_s} = -\frac{1}{2} tr\left(\mathbf{P}\mathbf{V}_{(s)}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_{(s)}\mathbf{P}\mathbf{y} \tag{22}$$

with

$$\mathbf{V}_{(1)} = \frac{\partial \mathbf{V}}{\partial \varphi_1} = \frac{\partial \mathbf{V}}{\partial \eta} = \boldsymbol{\Psi}\left(\boldsymbol{\Omega}\otimes\mathbf{I}_p\right)\boldsymbol{\Psi}^T \tag{23}$$

$$\mathbf{V}_{(2)} = \frac{\partial \mathbf{V}}{\partial \varphi_2} = \frac{\partial \mathbf{V}}{\partial \sigma_u^2} = \mathbf{I}_m\,. \tag{24}$$

The restricted ML estimate of $\boldsymbol{\varphi}$ can be obtained by setting the system of equations (22) to zero and solving for $\boldsymbol{\varphi}$. This can be done using a Fisher scoring algorithm as follows:

1. Compute the distance matrix $\boldsymbol{\Omega}$ between the centroids of the areas and define a starting value for $\boldsymbol{\varphi}$.
2. Use the current values of $\boldsymbol{\varphi}$ to calculate $\boldsymbol{\Sigma}$ and $\mathbf{V}$.
3. Update $\hat{\boldsymbol{\beta}}$ and **P**.
4. Calculate the value $\hat{\boldsymbol{\varphi}}$ such that the components of (22) are zero.
5. Return to step 3 and repeat the procedure until the estimates converge, i.e. when difference between the estimated model parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varphi}}$ obtained from two successive iterations is less than a very small value.

R code (R Development Core Team, 2010) that implements this algorithm is available from the authors.

3. Mean Squared Error Estimation

3.1 Analytic mean squared error estimation

Analytic estimation of the mean squared error (MSE) of the EBLUP (6) is usually carried out using the estimator of Prasad and Rao (1990). A corresponding analytic approach to estimating the MSE of the NSEBLUP (19) is developed below. After some algebra, we can show that the prediction error of the NSBLUP is

$$\tilde{Y}_i - Y_i = \mathbf{b}_i(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{h}_i\left\{\boldsymbol{\Sigma}_H\mathbf{H}^T\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)-\boldsymbol{\zeta}\right\}, \tag{25}$$

where $\mathbf{b}_i = \mathbf{x}_i^T - \left(\boldsymbol{\Psi}_i\boldsymbol{\Sigma}\boldsymbol{\Psi}^T + \sigma_u^2\boldsymbol{\delta}_i^T\right)\mathbf{V}^{-1}\mathbf{X} = \mathbf{x}_i^T - \mathbf{h}_i\boldsymbol{\Sigma}_H\mathbf{H}^T\mathbf{V}^{-1}\mathbf{X}$, $\mathbf{H} = \left[\boldsymbol{\Psi}\ \ \mathbf{I}_m\right]$, $\boldsymbol{\zeta} = (\boldsymbol{\gamma}^T, \mathbf{u}^T)^T$ and

$\boldsymbol{\Sigma}_H = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2\mathbf{I}_m \end{bmatrix}$. Put $\boldsymbol{\varphi} = \left\{\eta, \sigma_u^2\right\}$. The MSE of the NSBLUP is then

$$E\left(\tilde{Y}_i - Y_i\right)^2 = M_1(\boldsymbol{\varphi}) + M_2(\boldsymbol{\varphi}), \tag{26}$$

where

$$M_1(\boldsymbol{\varphi}) = \mathbf{h}_i\boldsymbol{\Sigma}_H\left\{\mathbf{I}_{pm+m} - \mathbf{H}^T\mathbf{V}^{-1}\mathbf{H}\boldsymbol{\Sigma}_H\right\}\mathbf{h}_i^T$$

and

$$M_2(\boldsymbol{\varphi}) = \mathbf{b}_i\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{b}_i^T\,.$$

Similarly, we can express the prediction error of the NSEBLUP (19) as

$$\hat{Y}_i - Y_i = \hat{\mathbf{b}}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{h}_i\left\{\hat{\boldsymbol{\Sigma}}_H\mathbf{H}^T\hat{\mathbf{V}}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)-\boldsymbol{\zeta}\right\}, \tag{27}$$

with $\hat{\mathbf{b}}_i = \mathbf{x}_i^T - \mathbf{h}_i\hat{\boldsymbol{\Sigma}}_H\mathbf{H}^T\hat{\mathbf{V}}^{-1}\mathbf{X}$. Following Opsomer *et al.* (2008), and again using $\boldsymbol{\varphi} = \left(\varphi_1, \varphi_2\right)$ to denote the variance components, define **S** to be the matrix with rows ($s = 1, 2$)

$$\mathbf{S}_s = \mathbf{h}_i \left(\frac{\partial \boldsymbol{\Sigma}_H}{\partial \varphi_s} \mathbf{H}^T \mathbf{V}^{-1} + \boldsymbol{\Sigma}_H \mathbf{H}^T \frac{\partial \mathbf{V}^{-1}}{\partial \varphi_s} \right)$$

where $\frac{\partial \boldsymbol{\Sigma}_H}{\partial \varphi_1} = diag\left(\boldsymbol{\Omega} \otimes \mathbf{I}_p, \mathbf{0}_m\right)$, $\frac{\partial \boldsymbol{\Sigma}_H}{\partial \varphi_2} = diag\left(\mathbf{0}_{pm}, \mathbf{I}_m\right)$ and $\frac{\partial \mathbf{V}^{-1}}{\partial \varphi_s} = -\mathbf{V}^{-1}\mathbf{L}_s\mathbf{V}^{-1}$ with $\mathbf{L}_s = \frac{\partial \mathbf{V}}{\partial \varphi_s}$. Here $\mathbf{L}_1 = \boldsymbol{\Psi}\left(\boldsymbol{\Omega} \otimes \mathbf{I}_p\right)\boldsymbol{\Psi}^T$ and $\mathbf{L}_2 = \mathbf{I}$. Note also that the 2×2 Fisher information matrix $\boldsymbol{\Phi}$ with respect to $\boldsymbol{\varphi}$ contains elements $\phi_{rs} = 0.5 \times tr(\mathbf{PL}_r\mathbf{PL}_s)$. Replacing the unknown variance components in $\mathbf{S}$ and $\boldsymbol{\Phi}$ by their restricted maximum likelihood estimates then leads to the following estimator of the Prediction MSE (PMSE) of the NSEBLUP

$$\begin{aligned} \widehat{PMSE}\left(\hat{Y}_i\right) &= M_1(\hat{\boldsymbol{\varphi}}) + M_2(\hat{\boldsymbol{\varphi}}) + 2M_3(\hat{\boldsymbol{\varphi}}) \\ &= \mathbf{h}_i\hat{\boldsymbol{\Sigma}}_H\left\{\mathbf{I}_{pm+m} - \mathbf{H}^T\hat{\mathbf{V}}^{-1}\mathbf{H}\hat{\boldsymbol{\Sigma}}_H\right\}\mathbf{h}_i^T + \hat{\mathbf{b}}_i\left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\hat{\mathbf{b}}_i^T \\ &\quad + 2\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\hat{\mathbf{S}}^T\hat{\boldsymbol{\Phi}}^{-1}\hat{\mathbf{S}}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right). \end{aligned} \tag{28}$$

The asymptotic behaviour of (28) can be developed along the same lines as set out in Opsomer *et al.* (2008). In particular, we then have

$$E\left[\widehat{PMSE}\left(\hat{Y}_i\right)\right] = PMSE\left(\hat{Y}_i\right) + o\left(m^{-1}L_m^2\right)$$

provided that regularity conditions described in Chandra *et al.* (2015) are met. The mean cross product error (MCPE) matrix defined by the vector $\hat{\mathbf{Y}}_{out} = \mathbf{X}_{out}\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}_{out}\hat{\boldsymbol{\gamma}}_{out}$ of values of the NSSYN predictor (20) is estimated in a similar fashion. Noting that

$$\begin{aligned} E(\mathbf{DD}^T) &= \boldsymbol{\Psi}_{out}\left\{\boldsymbol{\Sigma}_{out/out} + \boldsymbol{\Sigma}_{out/in}\boldsymbol{\Psi}^T\mathbf{V}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_{out/in}^T - 2\boldsymbol{\Sigma}_{out/in}\boldsymbol{\Psi}^T\mathbf{V}^{-1}\boldsymbol{\Psi}Cov\left[\boldsymbol{\gamma},\boldsymbol{\gamma}_{out}\right]\right\}\boldsymbol{\Psi}_{out}^T \\ &= \boldsymbol{\Psi}_{out}\left\{\boldsymbol{\Sigma}_{out/out} - \boldsymbol{\Sigma}_{out/in}\boldsymbol{\Psi}^T\mathbf{V}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_{out/in}^T\right\}\boldsymbol{\Psi}_{out}^T \end{aligned}$$

where $\mathbf{D} = \boldsymbol{\Psi}_{out}\boldsymbol{\Sigma}_{out/in}\boldsymbol{\Psi}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) - \boldsymbol{\Psi}_{out}\boldsymbol{\gamma}_{out}$, this estimator is given by

$$\begin{aligned} \widehat{MCPE}\left(\hat{\mathbf{Y}}_{out}\right) &= \boldsymbol{\Psi}_{out}\left\{\hat{\boldsymbol{\Sigma}}_{out/out} - \hat{\boldsymbol{\Sigma}}_{out/in}\boldsymbol{\Psi}^T\hat{\mathbf{V}}^{-1}\boldsymbol{\Psi}\hat{\boldsymbol{\Sigma}}_{out/in}^T\right\}\boldsymbol{\Psi}_{out}^T \\ &\quad + \mathbf{X}_{out}\left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}_{out}^T + \hat{\sigma}_u^2. \end{aligned} \tag{29}$$

Here $\hat{\boldsymbol{\Sigma}}_{out/out} = \hat{\eta}\boldsymbol{\Omega}_{out/out} \otimes \left(\mathbf{1}_p\mathbf{1}_p^T\right)$ and $\boldsymbol{\Omega}_{out/out}$ is the $m_{out} \times m_{out}$ matrix of distances between the non-sampled areas.

3.2 Bootstrap procedure for mean squared error estimation

This Section describes an alternative procedure for estimating the MSE of the NSEBLUP based on the parametric bootstrap procedure of Gonzalez-Manteiga *et al.* (2008). Note that the MSE estimator defined by this procedure is consistent provided the model parameter estimators are consistent. The steps of this parametric bootstrap procedure are as follows.

1) Given $\mathbf{y}$, maximise the restricted log-likelihood (21) using the method described in Section 2.1. Let $\hat{\boldsymbol{\varphi}} = \left(\hat{\eta}, \hat{\sigma}_u^2\right)$ and $\hat{\boldsymbol{\beta}}$ denote the resulting estimates.

2) Given the estimates obtained in step 1, generate a vector $\mathbf{t}_1^*$ of length *pm* corresponding to a realisation from the $N\left(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{I}_p\right)$ distribution. Construct the bootstrap vector $\boldsymbol{\gamma}^*\left(loc\right) = \hat{\eta}^{1/2}\mathbf{t}_1^*$.

3) Generate a vector $\mathbf{t}_2^*$ whose elements are *m* independent realisations of a *N*(0,1) variable, independently of the generation of $\mathbf{t}_1^*$. Construct the bootstrap vector $\mathbf{u}^* = \hat{\sigma}_u\mathbf{t}_2^*$.

4) Calculate the bootstrap realisation $\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}\hat{\boldsymbol{\gamma}}^* + \hat{\mathbf{u}}^* = \left(Y_1^*,\ldots,Y_m^*\right)^T$ of the population quantities of interest.
5) Generate a vector $\mathbf{t}_3^*$ whose elements are *m* independent realisations of a *N*(0,1) variable, independently of the generation of $\mathbf{t}_1^*$ and $\mathbf{t}_2^*$, and construct the vector of random prediction errors $\mathbf{e}^*$ where $e_i^* = \sigma_{ei} t_{3i}^*$.
6) Construct the bootstrap vector $\mathbf{y}^* = \mathbf{Y}^* + \mathbf{e}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Psi}\boldsymbol{\gamma}^* + \mathbf{u}^* + \mathbf{e}^*$ of direct estimates.
7) Using these bootstrap values $\mathbf{y}^*$, as well as the values of $\mathbf{X}$ and **loc**, calculate the bootstrap estimators $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\varphi}}^*$ using the method described in Section 2.1.
8) Combining the formula (19) with these values of $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\varphi}}^*$, calculate the bootstrap value $\hat{Y}_i^*$ of the NSEBLUP.
9) Repeat steps 2 - 8 B times. In the *b*-th bootstrap replication, let $\hat{Y}_i^{*(b)}$ be the bootstrap NSEBLUP in area *i*. The bootstrap estimator for the MSE of the actual NSEBLUP $\hat{Y}_i$, see (19), for area *i* is then

$$MSE_{boot}\left(\hat{Y}_i\right) = B^{-1}\sum_{b=1}^{B}\left\{\hat{Y}_i^{*(b)} - \hat{Y}_i\right\}^2 . \tag{30}$$

3.3. A diagnostic for spatial nonstationarity

Following Opsomer *et al.* (2008), we describe a bootstrap procedure to test the hypothesis $H_{0\eta}: \eta = 0$ versus the one-sided alternative $H_{1\eta}: \eta > 0$. This involves first calculating the value $\ell_\eta = 2\left(\ell_1 - \ell_0\right)$, where $\boldsymbol{\ell}_0$ denotes the restricted log-likelihood under the null $H_{0\eta}$ and $\boldsymbol{\ell}_1$ denotes the corresponding value under the alternative $H_{1\eta}$. The level of significance of ℓ_η is then calculated via a parametric bootstrap. That is, if we put $\hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ equal to the estimates of σ_u^2 and $\boldsymbol{\beta}$ obtained under the null, then we generate bootstrap realisations of $\mathbf{y}$ as $\mathbf{y}^{*(b)} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{u}^{*(b)} + \mathbf{e}^{*(b)}$, where $\mathbf{u}^{*(b)}$ and $\mathbf{e}^{*(b)}$ are generated as in Section 3.2. For each bootstrap replication, the null and the alternative models are then fitted and $\ell_\eta^{*(b)}$ is calculated. The significance of the calculated value of ℓ_η is evaluated by comparing it with the bootstrap distribution of $\ell_\eta^{*(b)}$. A word of caution is appropriate at this point. A significant result from the above bootstrap-based test does not mean that the model (10) with $\mathbf{C} = \eta\mathbf{I}_p$ provides a good representation of the data, i.e. the set of direct estimates $\mathbf{y}$. It only means that this particular spatially non-homogeneous model provides a significant improvement in fit compared with the usual Fay-Herriot approach that ignores spatial heterogeneity.

4. Application

In Section we illustrate an application of NSEBLUP method of SAE to real agriculture survey data collected by the National Sample Survey Office (NSSO), India under the scheme titled ‘Improvement of Crop Statistics (ICS)’. In particular, aim is to estimate average yield for paddy (green) crop at small area (or district) levels in the State of Uttar Pradesh in India by linking data generated under ICS scheme by NSSO (data collected with much reduced sample size, however, the quality of data is very high) and the Census information. In this application we adopt an area level small area model (2) as well as nonstationary version of area level model (10) for SAE. We applied the EBLUP and the NSEBLUP estimator for SAE described in Table 1.

Table 1. Definition of various estimators.

Acronym	Description	MSE Estimator
EBLUP	Predictor (6) under model (2)	Prasad & Rao (1990)
NSEBLUP	Predictor (19) under model (10)	Expression (28)
SYN	Predictor (7) under model (2)	Prasad & Rao (1990)
NSSYN	Predictor (20) under model (10)	Expression (29)

Here we are working under aggregated or area level version of small area models. As a consequence, this approach of SAE requires area-specific information on direct survey estimates and covariates. In particular, two types of variables are required for this analysis.

(i) The variable of interest for which small area estimates are required is yield for paddy (green) crop. We use data pertaining to supervised CCE on paddy (green) crop under ICS scheme for kharif season for the State of Uttar Pradesh in India collected during the year 2009-10. We are interested in estimating the average yield for paddy (green) crop at the district level. In the State of Uttar Pradesh CCE is carried out in the plots of form equilateral triangle of side 10 meter each and with total area of 43.30 meter^2. Therefore, yield rate for paddy (green) crop is recorded as gram per 43.12 meter^2.

(ii) The covariates (auxiliary variables) known for the population are drawn from the Population Census 2001. Note that use of covariates from the 2001 Population Census to model yield data of paddy crop from the 2009-10 ICS scheme data may raise issues of comparability. However, the covariates used in this study are not expected to change significantly over a short period of time.

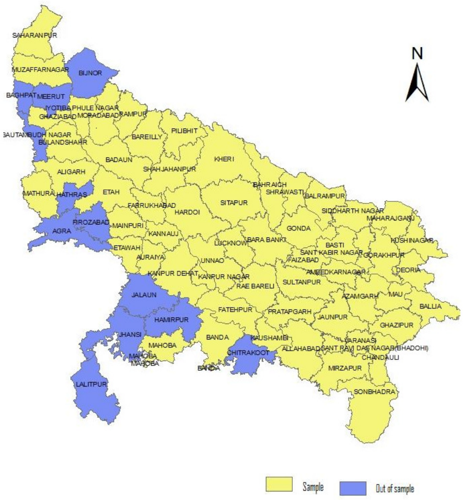


Figure 1. *Map of districts in the State of Uttar Pradesh in India.*

In the State of Uttar Pradesh there are 70 districts however supervision, on a sub-sample, of CCEs work under ICS scheme is carried out in 58 districts only and there is no sample data for the remaining 12 districts. We refer these 12 districts as the out of sample districts. These 70 (58 in sample and 12 out of sample) districts are the small areas for which we are interested in producing the estimates. Figure 1 shows the map of these 70 districts in the State of Uttar Pradesh. In this map the districts in sample are shown in yellow color while out of sample districts are shown in blue color. The area specific sample sizes for 58 sample districts range from minimum of 4 to maximum of 28 CCE with average of 11 (see Figure 2). A total of 655 CCE were supervised for recording yield data in the State of Uttar Pradesh for paddy crop for the year 2009-10. We see that in few districts the sample size is small so the traditional sample survey estimation approaches lead to unstable estimate. In addition, in 12 districts due to non availability of sample under ICS, we cannot estimate paddy yield**.** Indeed, there is no design based solution to provide estimates for these 12 out of sample districts. The SAE is an obvious choice for such cases.

There were 121 covariates available from the Population Census to consider for modeling and choosing appropriate covariates for this analysis. For this purpose, we used data of 58 sampled districts and did some exploratory data analysis, for example, first we segregated group of covariates with significant correlation with target variable and subsequently we implemented step wise regression analysis for identification of such covariates. Finally we identified two significant variables, average household size (HH_SIZE) and female population of marginal household (MARG_HH_F) with 26 per cent R^2 for the SAE. Although, the value of R^2 for this data is very good but this is best possible model we obtained from the available information. Note that for SAE of 12 out of sampled districts we used the same two covariates since we assume that the underlying model for sample areas also holds for out of sample districts.

For fitting the geographically weighted linear mixed model, i.e. spatial nonstationarity version of Fay Herriot model (1) we also require coordinates of different small areas. The developed SAE method is suitable for the data exhibiting the *spatial nonstationarity*. That is, if spatial nonstationarity is present in the data then the developed method of SAE accounts for nonstationarity while generating small area estimates.

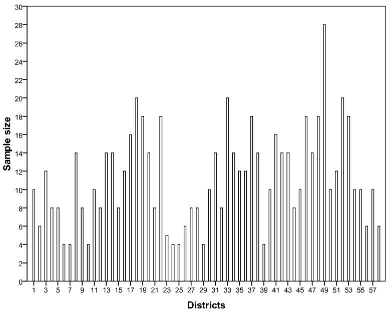


Figure 2. *Distribution of district-specific sample sizes in sample districts.*

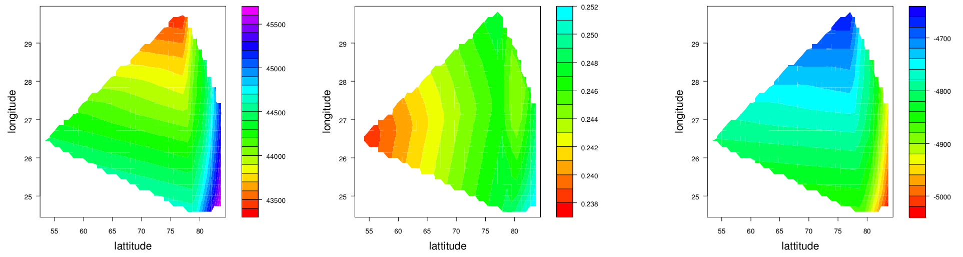
We did some exploratory data analysis to examine whether the ICS data used in this small area analysis (particularly, from sampled districts) have spatial nonstationarity. For this purpose, we computed the district-specific regression coefficients by fitting the spatial nonstationarity version of Fay Herriot model (10). We also obtained the global regression coefficients by fitting the Fay Herriot model (2). In the fitted model we have two covariates, average household size (HH_SIZE) and female population of marginal household (MARG_HH_F), therefore we have three regression coefficients (i.e., intercepts + two slope parameters with respect to HH_SIZE and MARG_HH_F). Table 2 reports the district wise estimates of regression coefficients for ICS data by fitting the spatial nonstationarity version of Fay Herriot model (10). Estimates of regression coefficients for ICS data by fitting the Fay Herriot model (2) are reported in Table 3. Surface plot of regression coefficients for ICS data by fitting the spatial nonstationarity version of Fay Herriot model (10) are shown in Figure 3. The diagnostic procedure to test the spatial nonstationarity described in Section 3.3, that is, the hypothesis $H_{0\eta}:\eta=0$ versus the one-sided alternative $H_{1\eta}:\eta>0$ is applied to NSSO data. Nonstationarity test is significant (p value 0.01), that is, rejected the null hypothesis of spatial stationarity of the model parameters of nonstationarity Fay Herriot model (10). Hence, there is evidence of nonstationarity in the data. In addition, the values of regression coefficients shown in Table 2 and in Figure 3 clearly indicate that the ICS data is not stationary. There is a marginal nonstationarity in the model parameters. Looking at the minimum, maximum, average and median values of model parameters of geographically weighted linear mixed model in Table 2 and the values of model parameters (i.e. global parameters) of linear mixed model in Table 3, we observe that there is evidence spatial nonstationarity in the model parameters. The surface plots of model parameters in Figure 3 and the conclusion from test result also confirm this. It is therefore interesting to apply the developed SAE method with ICS data. As there is evidence of spatial nonstationarity in the data, we expect a slightly better performance of small area estimates with the newly developed method of SAE.

Table 2. District wise estimates of regression coefficients for ICS data by fitting the spatial nonstationarity version of Fay Herriot model (10).

Districts	**Intercept**	**HH_F**	**HH_SIZE**	**Districts**	**Intercept**	**HH_F**	**HH_SIZE**
1	43336.5	0.25	-4653.6	32	44836.4	0.25	-4891.6
2	43620.5	0.25	-4697.1	33	44973.7	0.25	-4914.3
3	44046.5	0.25	-4760.6	34	44542.2	0.25	-4838.9
4	44022.4	0.25	-4757.2	35	44803.7	0.25	-4882.5
5	44076.7	0.25	-4765.2	36	44441.1	0.24	-4803.8
6	43951.9	0.25	-4746.9	37	44856.9	0.25	-4892.7
7	43687.2	0.25	-4707.8	38	44553.1	0.24	-4839
8	43863.3	0.25	-4734.4	39	44658.1	0.25	-4855.7
9	44243.8	0.24	-4790.8	40	44710.7	0.25	-4865.5
10	43934.3	0.25	-4746.7	41	44664.9	0.25	-4857.8
11	44111.6	0.25	-4772	42	44991.5	0.25	-4911.5
12	44192.6	0.25	-4784.4	43	44959.6	0.25	-4908.1
13	44148.4	0.24	-4776.3	44	45062.2	0.25	-4925
14	44165.9	0.24	-4778.3	45	45171.1	0.25	-4941.6
15	44208.8	0.24	-4784.3	46	45164.5	0.25	-4941.6
16	44253.6	0.24	-4791.4	47	45507	0.25	-4995.2
17	44349	0.24	-4805.2	48	45323.3	0.25	-4968.2
18	44811.6	0.25	-4888.8	49	45220.7	0.25	-4952.6
19	44302.4	0.24	-4799.3	50	45364.8	0.25	-4976.1
20	44451	0.24	-4824.3	51	45616.1	0.25	-5016.5
21	44489.4	0.24	-4830.2	52	45127.9	0.25	-4938.4
22	44670.7	0.25	-4862.2	53	45435.6	0.25	-4988.4
23	44244.6	0.24	-4791.2	54	45418.9	0.25	-4986.1
24	44302.1	0.24	-4800.1	55	45305.1	0.25	-4968.3
25	44245.9	0.25	-4793.9	56	45453.1	0.25	-4991.9
26	44680.3	0.25	-4870.5	57	45233.6	0.25	-4957.3
27	44439	0.25	-4823.3	58	45771.4	0.25	-5043.9
28	44085.3	0.25	-4767.8	**Min**	**43336.5**	**0.24**	**-5043.9**
29	44671.9	0.25	-4865.5	**Max**	**45771.4**	**0.25**	**-4653.6**
30	44644.3	0.25	-4858.8	**Mean**	**44627.4**	**0.25**	**-4855.1**
31	44970.8	0.25	-4913.5	**Median**	**44651.2**	**0.25**	**-4856.7**

Table 3. Estimates of regression coefficients for ICS data by fitting the Fay Herriot model (2).

Intercept	44415.191
HH_F	0.24
HH_SIZE	-4800



(a) Intercept (b) HH_SIZE (c) HH_F

Figure 3. *Surface plot of regression coefficients for ICS data by fitting the spatial nonstationarity version of Fay Herriot model (10).*

The coefficient of variation (CV) is used to assess the comparative precision of different small area estimates. The CVs show the sampling variability as a percentage of the estimate. Estimates with large CVs are considered unreliable (i.e. smaller is better). In general, there are no internationally accepted tables available that allow us to judge what is "*too large*". Different organization used different cut off for CV to release their estimate for the public use. For example, Office for National Statistics, United Kingdom has cut off CV value of 20% for acceptable estimates. We computed the percentage CV of direct estimates and two different model based estimates (i.e. EBLUP and NSEBLUP). Besides, comparison of model-based estimates versus direct estimates, we also want to compare the precision of two model-based estimates (i.e. EBLUP and NSEBLUP). Table 4 shows the district-wise distribution of the percentage CVs for the direct estimates and two different model-based estimates defined in Table 1. Figure 4 presents the percentage CV of direct estimates and two different model based estimates for sample districts.

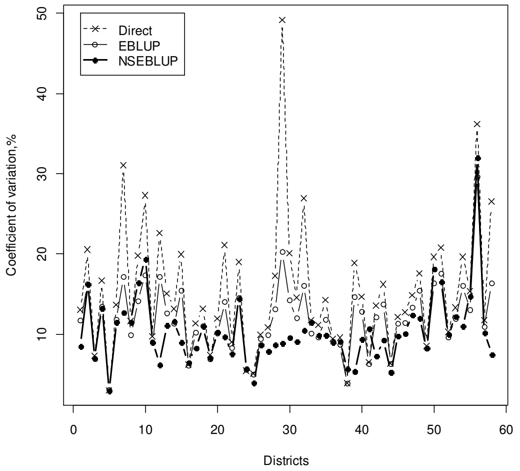


Figure 4. *District-wise plot of percent coefficient of variation for the Direct (dash line,×), EBLUP (thin line, o) and NSEBLUP (solid line, •) estimators.*

Two things stand out from Table 4 and Figure 4. First, the estimated CVs of two model-based estimators (i.e., EBLUP and NSEBLUP) are smaller than the traditional design-based direct estimator. Second, between two model based estimators, the NSEBLUP is better than the EBLUP. That is, the estimated CVs for the model-based estimates have a higher degree of reliability when compared to the direct estimates. In general, relative performance of model based estimates are better as sample size decreases. This result reveals that if spatial nonstationarity is incorporated in SAE it leads to significant gains in efficiency of small area estimates. It is interesting to note that out of 70 districts there are 12 districts with no sample data. For these 12 districts we cannot produce the direct estimates, however, model based estimates generated for these districts have reasonably good CV values and that to within the acceptable limit. In Table 4 If we look at the percent CV it is apparent that the standard errors of the direct estimates are large and therefore the estimates are unreliable. Further, for 12 out of sample districts we observed a significant gain in application of NSEBLUP as compared to the EBLUP method. That is, the NSEBLUP method leads to a drastic gain in precision when estimates were produced for out of sample districts.

Table 4. District wise estimates and percent coefficient of variation for the direct, EBLUP and NSEBLUP estimators from NSSO data.

Districts	**Direct**		**EBLUP**		**NSEBLUP**	
	Yield	**CV,%**	**Yield**	**CV,%**	**Yield**	**CV,%**
Saharanpur	19575	13.04	17737	11.64	17269	8.51
Muzaffarnagar	23483	20.53	17168	16.20	18194	16.26
Bijnor	19442	7.28	18918	6.95	19323	7.00
Moradabad	17700	16.67	16770	13.42	16830	13.18
Rampur	17250	3.01	17173	2.99	17110	2.97
Jyotiba Phule Nagar	10850	13.68	11635	11.78	12009	11.53
Ghaziabad	16800	31.03	16726	17.10	17420	12.70
Bulandshahar	17418	11.64	18126	9.89	17314	11.44
Aligarh	12419	19.77	14302	14.11	12365	16.43
Mathura	10483	27.27	12712	17.35	12146	19.38
Etah	12125	9.73	12514	8.95	12249	9.01
Mainpuri	14019	22.58	13707	17.14	13641	6.20
Budaun	12721	15.05	13315	12.64	13248	11.07
Bareilly	13511	13.18	14150	11.24	14114	11.60
Pilibhit	14938	19.94	14684	15.42	15277	8.96
Shahjahanpur	18863	6.23	18403	6.07	17307	6.47
Kheri	14975	11.37	15081	10.19	15392	8.28
Sitapur	15986	13.11	16428	10.99	16344	10.96
Hardoi	19286	7.39	19315	6.89	19468	7.05
Unnao	12843	11.92	14024	10.09	14144	10.24
Lucknow	17331	21.08	18251	14.00	17573	9.66
Rae Bareli	19506	9.03	19284	8.24	18858	7.61
Farrukhabad	8880	18.95	10470	14.52	10505	14.42
Kannauj *	34050	5.45	30396	5.51	33034	5.64
Etawah	15463	5.07	15431	4.97	15496	3.97
Auraiya	23717	9.96	20987	9.37	19081	8.69
Kanpur Dehat	21200	10.82	19526	9.89	17331	7.88
Kanpur Nagar	15375	17.27	16326	13.06	16434	8.70
Banda	8888	49.15	13406	20.20	15905	8.82
Fatehpur	14612	20.11	15895	14.23	16793	9.59
Pratapgarh	16304	14.52	16439	11.94	16749	9.06
Kaushambi	15450	26.93	16633	16.04	17038	10.50
Allahabad	19465	11.72	20227	10.10	19415	11.50
Barabanki	18668	11.12	18757	9.61	18044	9.87
Faizabad	16379	14.26	16559	11.77	16745	9.86
Ambedkar Nagar	17692	9.44	16650	9.10	16900	9.00
Sultanpur	16609	9.57	16796	8.66	16653	9.07
Bahraich	14714	3.89	14736	3.84	15197	5.63
Shrawasti	15075	18.90	15169	14.60	15947	5.36
Balrampur	11975	14.63	12343	12.76	14200	9.42
Gonda	16981	6.47	16704	6.29	15442	10.67

Siddharthnagar	12829	13.55	12922	12.11	13970	7.26
Basti	14268	16.21	14163	13.65	14327	9.25
Sant Kabir Nagar	13319	6.35	13272	6.21	13169	5.27
Mahrajganj	21690	12.15	18603	11.33	15745	9.78
Gorakhpur	12164	12.73	12441	11.42	12793	10.04
Kushinagar	19343	14.88	16669	13.33	16006	12.39
Deoria	8364	17.58	8873	15.41	9858	12.01
Azamgarh	11957	8.52	12034	8.16	11924	8.26
Mau	9820	19.64	10498	16.29	9578	18.14
Ballia	7029	20.78	7775	17.48	7988	16.53
Jaunpur	16990	10.27	16408	9.60	16745	9.93
Ghazipur	10858	13.29	11286	11.89	10933	12.15
Chandauli	12000	19.63	12231	16.06	12196	10.94
Varanasi	17665	15.38	17055	13.04	17594	14.69
Sant Ravidas Nagar	6693	36.21	7136	29.63	6522	32.04
Mirzapur	15625	11.71	15043	10.83	15162	10.19
Sonbhadra	15283	26.49	16337	16.29	17605	7.41
Meerut			14897	21.60	16167	19.97
Baghpat			11947	27.57	13030	11.03
Gautam Buddha Nr			16677	19.83	17756	6.82
Hatharas			15162	21.27	14326	16.86
Agra			14731	21.84	12789	30.74
Firozabad			14223	22.67	12954	22.99
Jalaun			15028	21.58	15927	10.82
Jhansi			17582	18.54	19186	22.03
Lalitpur			16959	19.43	18698	9.96
Hamirpur			16476	20.01	17280	7.19
Mahoba			16196	20.38	17123	6.01
Chitrakoot			14723	22.14	15202	5.79

5. Concluding Remarks

This paper illustrates that the SAE technique can be satisfactorily applied to produce reliable district level estimates of crop yield using CCE supervised under ICS scheme. Although the ICS supervised CCEs number only 30,000 in the entire country i.e. the sample size is very low, the collected data is of very high quality. The estimates generated using this data are expected to be relatively free from various sources of non-sampling errors. Further SAE provides estimates for those districts where there is no sample information under ICS and so direct estimates cannot be computed. It is, therefore, recommended that wherever it is not possible to conduct adequate number of CCEs due to constraints of cost or infrastructure or both, SAE technique can be gainfully used to generate reliable estimates of crop yield based on a smaller sample. In addition, when there is spatial nonstationarity in the data developed method should be used to improve these disaggregate level estimates. We noticed that the ICS data have evidence of spatial nonstationarity. As a consequence the developed NSEBLUP method when applied to ICS data enhanced the efficiency of small area estimates.

REFERENCES

Anselin, L. (1992) *Spatial Econometrics: Method and Models*. Kluwer Academic Publishers, Boston.

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity, *Geographical Analysis,* 28, 281-298.

Chandra, H., Salvati, N. and Chambers, R. (2015) A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation, *Journal of Survey Statistics and Methodology*. 3(2), 109-135.

Cressie, N. (1991) Small-Area Prediction of Undercount Using the General Linear Model, in the Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada, 93-105.

Datta, G., Day, B., and Maiti, T. (1998) Multivariate Bayesian small area estimation: An application to survey and satellite data, *Sankhya. Series A*, 60, 344-362.

Fay, R. E. and Herriot, R. A. (1979) Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data*, Journal of the American Statistical Association,* 74, 269-277.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2002) *Geographically Weighted Regression*. John Wiley & Sons, West Sussex.

Giusti, C., Marchetti, S., Pratesi, M. and Salvati, N. (2012), "Semiparametric Fay-Herriot Model using Penalized Splines", *Journal of the Indian Society for Agricultural Statistics,* 66(1)*,* pp.1-14.

Gonzalez-Manteiga, W., Lombardia, M., Molina, I., Morales, D. And Santamaria, L. (2008), "Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay–Herriot Model", *Computational Statistics and Data Analysis,* 52, pp. 5242–5252.

Henderson, C. R. (1975) Best Linear Unbiased Estimation and Prediction under a Selection Model, *Biometrics,* 31, 423-447.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008) Nonparametric Small Area Estimation Using Penalized Spline Regression, *Journal of the Royal Statistical Society, Series B*, 70, 265-286.

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.

Pratesi, M. and Salvati, N. (2008)Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects, *Statistical Methods and Applications*, 17, 114-131.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

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

Singh, B.B., Shukla, G.K. and Kundu, D. (2005) Spatio-Temporal Models in Small Area Estimation", *Survey Methodology*, 31, *2*, 183-195.