Special Proceeding of the 21st Annual Conference of SSCA held at SV Agricultural College (ANGRAU), Tirupati, during January 29-31, 2019; pp 29-41

Small Area Estimation of Poverty Incidence under a Spatial Non-Stationary Generalized Linear Mixed Model*

Hukum Chandra and Bhanu Verma
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Received: 19 August 2019; Revised: 21 August 2019; Accepted: 22 August 2019

Abstract

The empirical plug-in predictor (EP) under an area level version of the generalized linear mixed model (GLMM) with logit link function is extensively used in small area estimation (SAE) for proportions. However, this model assumes that the fixed effect parameters are spatially invariant and does not account for the presence of spatial non-stationarity in the data. This paper describes a spatially non-stationary extension of the area level version of GLMM (NSGLMM) and SAE under this model. In particular, the empirical plug-in predictor for small area proportions (NSEP) under the area level version of NSGLMM with logit link function is delineated. The NSEP method is then applied to estimate the extent of household poverty in different districts of the rural part of the state of Uttar Pradesh using data from the 2011-12 Household Consumer Expenditure Survey collected by the National Sample Survey Office of India, and the 2011 population Census. A poverty map for State of Uttar Pradesh is also produced which provides an important information for analysis of spatial distribution of poverty in the state.

Keywords: Spatial non-stationarity, Binary data, MSE estimation, Poverty mapping.

1 Introduction

Sample surveys are generally planned to produce estimates for populations, sub-populations or larger domains. Sample sizes are fixed to provide reliable and representative estimates with a pre-determined level of precision for such planned domains. However, policy planners, researchers, government and public agencies often require estimates for unplanned domains. Such unplanned domains can be small geographic areas (e.g. municipalities, developmental blocks, tehsils, gram panchayats, etc.) or small demographic groups (e.g. age-sex-race groups within larger geographical areas) or a cross classification of both. The sample sizes for such unplanned domains in the available survey data may be very small or even zero. A domain is regarded as small if the domain-specific sample size is not large enough to ensure that a direct survey estimator has adequate precision. In such cases it becomes necessary to employ indirect small area estimators that make use of the sample data from related areas or domains through linking models, thus increasing the effective

Corresponding Author: Hukum Chandra
Email: hchandra12@gmail.com

sample size in the small areas. Such estimators can have significantly smaller coefficient of variation than direct estimators, provided the linking models are valid. The statistical methodology that tackles this problem of small sample sizes is often referred as small area estimation (SAE) theory in the survey literature, see Rao and Molina (2015). Based on the level of auxiliary information available, the models used in SAE are categorized as area level or unit level. 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. The Fay–Herriot model (Fay and Herriot, 1979) is a widely used area level model in SAE that assumes area-specific survey estimates are available, and that these follow an area level linear mixed model with area random effects.

When the variable of interest is binary, the use of standard SAE methods based on linear mixed models becomes problematic. In this context, generalized linear mixed model (GLMM) with logit link function (also referred as the logistic linear mixed model) is commonly used. When only area level data are available, an area level version of a GLMM can be used for SAE, see Johnson *et al*. (2010) and Chandra *et al*. (2011). The fixed effect parameters in GLMM is assumed to be spatially invariant. However there are situations where this assumption is inappropriate, and the parameters associated with the model covariates vary spatially. This paper describes a spatial nonstationary extension of the area level version of GLMM to incorporate spatial non-stationarity (referred as the NSGLMM), and then used this model in SAE via its corresponding empirical predictor (NSEP), see for example Chandra *et al*. (2017).

The structure of the paper is as follows. Section 2 describes the data from the 2011-12 Household Consumer Expenditure Survey of the National Sample Survey Office (NSSO) of India and the 2011 Indian Population Census that will be used to estimate the district level incidence of household poverty in the rural part of the Indian State of Uttar Pradesh. In Section 3 we set out the theoretical background of the area level version of the GLMM which is then used to define the plug-in empirical predictor (EP) for small areas. The extension of the area level GLMM to spatially nonstationary data (the NSGLMM) is defined in Section 4 and SAE using the plug-in empirical predictor (the NSEP) based on this model is presented. Section 5 demonstrates the application of the NSEP to poverty mapping in Uttar Pradesh. Finally, section 6 summarizes the main conclusions.

2 Data Description

This section introduces the basic sources of the data, i.e. the 2011-12 Household Consumer Expenditure Survey (HCES) of the National Sample Survey Office (NSSO) for rural areas of the State of Uttar Pradesh in India and the 2011 Population Census, used in the small area application reported in this paper. Data obtained from these sources are then used to estimate the proportion of poor households at district level in Uttar Pradesh. The State of Uttar Pradesh is the most populous State in the country and accounts for about 16.16 per cent of India's population. It covers 243,290 square km, equal to 6.88% of the total area of the country. Poverty estimates in India are produced for all the States separately for both rural and urban sectors. Our analysis is restricted to the rural areas of Uttar Pradesh because about 78% of the population of this State live in rural areas according to 2011

Population Census. The NSSO conducts nationwide HCE surveys at regular intervals as part of its "rounds", with the duration of each round normally being a year. These surveys are aimed at generating estimates of average household monthly per capita consumer expenditure (MPCE), the distribution of households and persons over the MPCE range, and the break-up of average MPCE by commodity group, separately for the rural and urban sectors of the country, for States and Union Territories, and for different socio-economic groups. These indicators are amongst the most important measures of the living conditions of the relevant domains of the population. The sampling design used in the NSSO survey is stratified multi-stage random sampling with districts as strata, villages as first stage units and households as second stage units. Although, these surveys provide reliable and representative state and national level estimates, they cannot be used directly to produce reliable estimates at the district level due to small sample sizes. In the 2011-12 HCES, a total of 5916 households from the 71 districts of Uttar Pradesh were surveyed. The district sample sizes ranged from 32 to 128 with average of 83. It is evident that these district level sample sizes are relatively small, with an average sampling fraction of 0.0002. As a consequence, it is difficult to generate reliable district level direct survey estimates with associated standard errors from this survey.

The target variable Y at the unit (household) level in the published survey data file is binary, corresponding to whether a household is poor or not. In our application however we focus on estimation where the available data are the corresponding counts of the number of poor households in sample in each district. In this context a household having MPCE below the state poverty line is defined as being poor. The poverty line used in this study (Rs. 768) is the same as that set by the Planning Commission, Govt. of India, for 2011-12. The parameter of interest is then the proportion of poor rural households within each district. The auxiliary variables (covariates) used in our analysis are taken from the Indian Population Census of 2011. These auxiliary variables are only available as counts at district level, and so SAE methods based on area level small area models, as described in next section, must be employed to derive the small area estimates. There are approximately 50 such covariates that are available for use in SAE analysis. We therefore carried out a preliminary data analysis in order to define appropriate covariates for SAE modelling, using Principal Component Analysis (PCA) to derive composite scores for selected groups of variables. In particular, we carried out PCA separately on three groups of variables, all measured at district level and identified as G1, G2 and G3 below. The first group (G1) consisted of literacy rates by gender and proportions of worker population by gender. The first principal component for this group explained 51% of the variability in the G1 group, while adding the second principal component (G12) increased explained variability to 85%. The second group (G2) consisted of the proportions of main worker by gender, proportions of main cultivator by gender and proportions of main agricultural labourer by gender. The first principal component (G21) for this second group explained 49% of the variability in the G2 group, while adding the second component (G22) increased explained variability to 67%. Finally, the third group (G3) consisted of proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. The first principal component (G31) for this third group explained 61% of the variability in the S3 group, while adding the second component (G22) increased explained variability to 78%. Using the methods detailed in the following sections, we fitted a generalised linear model using direct survey estimates of proportions of poor rural households as

the response variable and the six principal component scores G11, G12, G21, G22, G31 and G32 as potential covariates. The final selected model included the three covariates G11, G21 and G31, with residual deviance and AIC values of 327.18 and 636.89, respectively. This final model was then used to produce district wise estimates of rural poverty incidence, i.e. estimates of the head count ratio (HCR) at this level.

3 Small Area Estimation under the Area Level GLMM

We assume that a probability sampling method is used to draw a sample *s* of size *n* from a finite population U of size *N,* which consists of *m* non-overlapping domains U_i $(i = 1,....,m)$. We refer to these domains as small areas or just areas. Furthermore, we assume that there is a known number N_i of population units in small area *i*, with n_i of these sampled. The total number of units in the population is $N = \sum_{i=1}^{m} N_i$, with corresponding total sample size $n = \sum_{i=1}^{m} n_i$. We use *s* to denote the collection of units in sample, with s_i the subset drawn from small area *i* (i.e. $|s_i| = n_i$), and use expressions like $j \in i$ and $j \in s$ to refer to the units making up small area *i* and sample *s,* respectively. Similarly, r_i denotes the set of units in small area *i* that are not in sample, with $|r_i| = N_i - n_i$ and $U_i = s_i \cup r_i$. Let y_{ij} denotes the value of the variable of interest for unit *j* $(j = 1,..., N_i)$ in area *i*. The variable of interest, with values y_{ij}, is binary (e.g., $y_{ij} = 1$ if household *j* in area *i* is poor household and 0 otherwise), and the aim is to estimate the small area population count, $y_i = \sum_{j \in U_i} y_{ij}$, or equivalently the small area proportion, $P_i = N_i^{-1} \sum_{j \in U_i} y_{ij}$, in area i $(i = 1,..., m)$. The direct survey estimator (denoted by direct) for P_i is $p_{iw} = \sum_{j \in s_i} \tilde{w}_{ij} y_{ij}$, where $\tilde{w}_{ij} = w_{ij} \Big/ \sum_{j \in s_i} w_{ij}$ is the normalized survey weight for unit *j* in area *i* with $\sum_{j \in s_i} \tilde{w}_{ij} = 1$ and w_{ij} is the survey weight for unit *j* in area *i*. The estimated design-based variance of direct is approximated by $v(p_{iw}) \approx \sum_{j \in s_i} \tilde{w}_{ij} (\tilde{w}_{ij} - 1)(y_{ij} - p_{iw})^2$. This formula for the variance estimator of direct is obtained from Särndal *et al.* (1992; see pp. 43, 185 and 391), with the simplifications $w_{ij} = a_{ij}^{-1}$, $a_{ij,ij} = a_{ij}$ and $a_{ij,ik} = a_{ij} a_{ik}, j \neq k$, where a_{ij} is the first order inclusion probability of unit *j* in area *i* and $a_{ij,ik}$ is the second order inclusion probability of units *j* and *k* in area *i.* Under simple random sampling (SRS), $w_{ij} = N_i n_i^{-1}$ and direct is then $p_i = n_i^{-1} y_{si}$, with estimated variance $v(p_i) \approx n_i^{-1} p_i (1 - p_i)$, where $y_{si} = \sum_{j \in s_i} y_{ij}$ denotes the sample count in area *i*.

Suppose now that the available data consist of the sample aggregates y_{si} (i.e. the sample counts of poor households), together with the values of area specific contextual covariates. That is, for area *i* we observe the count y_{si} together with a *p*-vector of area-specific covariates $\mathbf{x}_i$ derived from secondary data sources (e.g. the census or administrative registers). If we ignore the sampling design, the sample count

y_{si} in area *i* can be assumed to follow a Binomial distribution with parameters n_i and π_i, i.e. $y_{si} \sim \text{Bin}(n_i, \pi_i)$, where π_i is the probability of occurrence of an event for a population unit in area *i* or the probability of prevalence in area *i*. Following Johnson *et al*. (2010) and Chandra *et al*. (2011), the model linking the probability π_i with the covariates $\mathbf{x}_i$ is the logistic linear mixed model of form

$$logit(\pi_i) = \ln\left\{\pi_i(1-\pi_i)^{-1}\right\} = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i, \quad (1)$$

with $\pi_i = \exp(\eta_i)\{1+\exp(\eta_i)\}^{-1} = \text{expit}(\mathbf{x}_i^T\boldsymbol{\beta} + u_i)$. Under this model, the mean of y_{si} given u_i is $\mu_{si} = E(y_{si}|u_i) = n_i \,\text{expit}(\mathbf{x}_i^T\boldsymbol{\beta} + u_i)$. Here $\boldsymbol{\beta}$ is the *p*-vector of regression coefficients, often referred to as the vector of fixed effects, and u_i is the area-specific random effect, with u_i : $N(0, \sigma_u^2)$. Note that the estimation of the fixed effect parameter and the area specific random effects in model (1) uses the data from all small areas. Without loss of generality we focus on the Binomial case and put $\boldsymbol{\pi} = (\pi_1, ..., \pi_m)^T$ and $\mathbf{X} = (\mathbf{x}_1^T, ..., \mathbf{x}_m^T)^T$. We assume that $\mathbf{u} = (u_1, ..., u_m)^T$ is a $m \times 1$ vector of area random effects which has a Gaussian distribution with zero mean vector and covariance matrix $\Sigma_u = \sigma_u^2 \mathbf{I}_m$. Here $\mathbf{I}_m$ is the identity matrix of order *m*. We adopt a Penalized Quasi-Likelihood method of estimation for the parameters $\boldsymbol{\beta}$ and **u** in the GLMM (2), combined with restricted maximum likelihood estimation of the variance parameter σ_u^2. See Manteiga *et al.* (2007). Under (1), a plug-in empirical predictor (EP) of the population count y_i in area *i* is

$$\hat{y}_i^{EP} = y_{si} + \hat{\mu}_{ri} = y_{si} + \left(N_i - n_i\right)\left\{\text{expit}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \hat{u}_i)\right\}. \quad (2)$$

An estimate of the corresponding proportion or rate in area *i* is obtained as $\hat{P}_i^{EP} = N_i^{-1}\hat{y}_i^{EP}$**.** For areas with zero sample sizes, the conventional approach to estimating area proportions is synthetic estimation, based on a suitable GLMM fitted to the counts from the sampled areas. For non-sampled area *i* with associated vector of covariates $\mathbf{x}_{i,out}$, the synthetic estimator of y_i is $\hat{y}_i^{SYN} = N_i\left\{\text{expit}(\mathbf{x}_{i,out}^T\hat{\boldsymbol{\beta}})\right\}$.

4 Small Area Estimation under the Area Level NSGLMM

The vector of fixed effect parameters $\boldsymbol{\beta}$ in (1) is spatially invariant. However there are situations where this assumption is inappropriate, and the parameters associated with the model covariates vary spatially. We introduce a spatial nonstationary extension of (1) that can be used in such situations. Let d_i denote the coordinates of an arbitrary spatial location (longitude and latitude) in area *i*. Typically, this will be its centroid. Let $\mathbf{d} = (d_1, ..., d_m)^T$ denote the corresponding *m*-vector of such spatial locations, and let $\pi_i(d_i)$ be the probability of occurrence of a characteristic of interest in area *i*, defined relative to the location d_i. A model for a nonstationary GLMM (NSGLMM) for $\pi_i(d_i)$ is then

$$\eta_i(d_i) = \mathbf{x}_i^T\boldsymbol{\beta}(d_i) + u_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\gamma}(d_i) + u_i, \quad (3)$$

where we assume that non-stationarity is characterised by an area-specific vector of fixed effects $\boldsymbol{\beta}(d_i) = \boldsymbol{\beta} + \boldsymbol{\gamma}(d_i)$; u_i is the area-specific random effect, assumed to follow a Gaussian distribution with zero mean and variance ϕ; and

$\boldsymbol{\gamma}(d) = \left(\boldsymbol{\gamma}_1(d),\ldots,\boldsymbol{\gamma}_p(d)\right)^T$ is a spatially correlated vector-valued random process with $E(\boldsymbol{\gamma}(d_i)) = \mathbf{0}$ and such that $\text{cov}\left(\gamma_k(d_i),\gamma_l(d_j)\right) = c_{kl}\left(1+L(d_i,d_j)\right)^{-1}$. Here $L(d_i,d_j)$ is the spatial distance between locations l_i and l_j and $\mathbf{c} = (c_j)$ is a p-vector of unknown positive constants that satisfies the conditions for the $pm \times pm$ matrix $\boldsymbol{\Sigma}_\gamma = \boldsymbol{\Omega} \otimes (\mathbf{c}\mathbf{c}^T)$ to be a covariance matrix, with $\boldsymbol{\Omega} = \left[\left(1+L(l_i,l_j)\right)^{-1}\right]$ defining the matrix of distances between the sample areas, and where $\otimes$ denotes Kronecker product. We can write the population level version of (4) as

$$\boldsymbol{\eta}(d) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}(d) + \mathbf{u} \tag{4}$$

where $\boldsymbol{\eta}(d) = \left(\eta_1(d),\ldots,\eta_m(d)\right)^T$, $\mathbf{Z} = \{diag(\mathbf{x}_1),.....,diag(\mathbf{x}_m)\}$ is a $m \times pm$ matrix and $\boldsymbol{\gamma}(d)$ is a $pm \times 1$ vector of spatial Gaussian random effects that capture the spatial non-stationarity in the data. We assume that $\boldsymbol{\gamma}(d)$ has a zero mean vector and a covariance matrix $\boldsymbol{\Sigma}_\gamma = \boldsymbol{\Omega} \otimes (\mathbf{c}\mathbf{c}^T)$. In general, the only constraint on the vector $\mathbf{c}$ is that $\boldsymbol{\Sigma}_\gamma = \boldsymbol{\Omega} \otimes (\mathbf{c}\mathbf{c}^T)$ is symmetric and non-negative definite. In practice ϕ and $\mathbf{c}$ are unknown and have to be estimated from the data. We restrict ourselves to the simple specification $\mathbf{c} = \sqrt{\lambda}\mathbf{1}_p$ so that $\text{cov}\left(\gamma_k(d_i),\gamma_l(d_j)\right) = \lambda\left(1+L(d_i,d_j)\right)^{-1}$, where $\lambda \geq 0$ and $\mathbf{1}_p$ denotes the unit vector of order p. In this case, we assume that the distance metric used to define $L(d_i,d_j)$ is such that the matrix $\boldsymbol{\Omega} \otimes (\mathbf{1}_p\mathbf{1}_p^T)$ is positive semidefinite, with the parameter λ then reflecting the 'intensity' of spatial clustering in the data, so $\lambda = 0$ corresponds to the situation where the model is spatially homogeneous. Given this specification, there are just 2 parameters (λ and ϕ) that need to be estimated. Replacing these unknown parameters by their estimated values $\hat{\phi}$ and $\hat{\mathbf{c}}$, and denoting subsequent plug-in estimators by a 'hat', we define the nonstationary empirical predictor (NSEP) of the population count in area i as

$$\hat{y}_i^{NSEP} = y_{si} + \left(N_i - n_i\right)\left\{\text{expit}\left[\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \mathbf{z}_i^T\hat{\boldsymbol{\gamma}}(d_i) + \hat{u}_i\right]\right\}. \tag{5}$$

Here $\mathbf{z}_i^T$ is the i-th row of $\mathbf{Z}$. A nonstationary empirical predictor of the proportion in area i is $\hat{p}_i^{EP} = N_i^{-1}\hat{y}_i^{EP}$. The synthetic prediction for a non-sample area is also straightforward. We set the estimated area effect to zero in this case, and evaluate (5) at the location d_i of the non-sampled area. The result is a nonstationary synthetic predictor (NSSYN) of the total count for the area of the form $\hat{y}_i^{NSSYN} = N_i\left\{\text{expit}\left[\hat{\eta}_i(d_i) = \mathbf{x}_{out,i}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{out,i}^T\hat{\boldsymbol{\gamma}}(d_i)\right]\right\}$.

In practice, the variance components that define the matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\gamma$ are unknown and must be estimated from the sample data. It is well known that the maximum penalised quasi-likelihood (MPQL) estimates of these variance components are biased, and that this bias increases with the relative contributions of the associated random effects to overall variability. Consequently, this approach is not recommended. Alternative estimates based on a hybrid of MPQL for fixed and random effects and maximum likelihood (ML) or restricted maximum likelihood (REML) for variance components can be defined. These can reduce, but not

eliminate, the aforementioned bias. Since prediction of small area quantities, rather than parameter estimation, is our focus, we use the hybrid approach. Under the hybrid approach, parameter estimates for the NSGLMM are obtained by a two-stage iterative process. At the first stage, MPQL estimates of $\boldsymbol{\beta}$, $\mathbf{u}$ and $\boldsymbol{\gamma}(d)$ are calculated based on specified values of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\gamma$, and at the second stage $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\gamma$ are estimated via ML or REML given these MPQL estimates. This hybrid approach, where MPQL estimation is combined with ML or REML estimation in generalized linear mixed models is due to McGilchrist (1994). Readers may refer to Chandra *et al.* (2017) for further details about the algorithm for estimation of the parameters.

Chandra *et al.* (2017) also suggetsed a diagnostic for spatial nonstationarity in the NSGLMM. In particular, they describe a bootstrap procedure to test the spatial nonstationarity hypothesis in the context of the simple one parameter NSGLMM (4) considered, i.e. the hypothesis $H_0 : \lambda = 0$ versus the one-sided alternative $H_1 : \lambda > 0$. Two models are fitted, first without spatial random effects, and second with these effects. The test then involves comparing the restricted log-likelihoods under each hypothesis.

Turning now to mean squared error (MSE) estimation of EP and NSEP predictors of small area proportions. Analytic estimation of the MSE of the small area empirical predictor EP (2) follows from the Johnson *et al.* (2010), Chandra *et al.* (2011). A corresponding analytic approach to estimating the MSE of the NSEP (5) under the NSGLMM (4) is developed in Chandra *et al.* (2017). This MSE is a second order approximation to the MSE of the NSEP.

5 Application to Small Area Poverty Estimation

In this section, we use NSEP under NSGLMM (3) to obtain estimates of the proportions of poor households at District level in the State of Uttar Pradesh. We use survey data from the Household Consumer Expenditure Survey 2011-12 of NSSO 68th round and the Population Census 2011. Section 2 illustrates the data and model specification for this application. Figure 1 shows contour maps of the estimated District-specific intercepts and slopes from a geographically weighted regression (GWR) fit (Fotheringham *et al.*, 2002) to the sample proportions for the different Districts. These maps support the case for spatial non-stationarity in the NSSO data. In particular, we see that the coefficients for G11, G21 and G31 in the GWR fit vary considerably, ranging from −1.3 to 2.9 for G11, from -1.2 to 2.8 for G21 and from -0.9 to 4.3 for G31. Moreover the contour map of the intercept coefficients also shows considerable spatial variation, ranging from -5.3 (South-East) to 0.2 (Centre-West).

The diagnostic procedure for testing for the presence of spatial non-stationarity, that is, the hypothesis $H_0 : \lambda = 0$ versus the one-sided alternative $H_1 : \lambda > 0$, was also applied to the NSSO data. The test statistic value that was generated was highly significant (p-value = 0.00), indicating strong evidence for non-stationarity, with the estimated value of the variance component λ characterising the intensity of this non-stationarity equal to 0.244. Hence, we applied the NSEP method to produce estimates of the proportions of poor households (Head Count Ratio, or HCR) by District across Uttar Pradesh. Table 1 shows the estimates of the regression coefficients for the global GLMM as well as descriptive statistics for the District-specific estimated parameter coefficients produced under the NSGLMM. The G11

and G21 variables had both negative and positive parameter values, although most values are negative. In contrast, all parameter values for the G31 variable are positive. As one would expect, the NSGLMM is an improvement over the GLMM for predicting HCR, with the NSGLMM log-likelihood (2375.5) significantly larger than the corresponding log-likelihood generated by the GLMM (-2970.2). The contour maps shown in Figure 2 support the non-stationarity hypothesis. Here we see that the fitted regression coefficents in the NSGLMM applied to the district level data change substantially over the study space. Although not reported here, the fitted distributions of the predicted random effects in the NSGLMM applied to the survey data were also produced. These include histograms as well as Normal *q-q* plots of District residuals as well as the spatial random effects. These indicate that the Gaussian assumption fits reasonably well for this application.

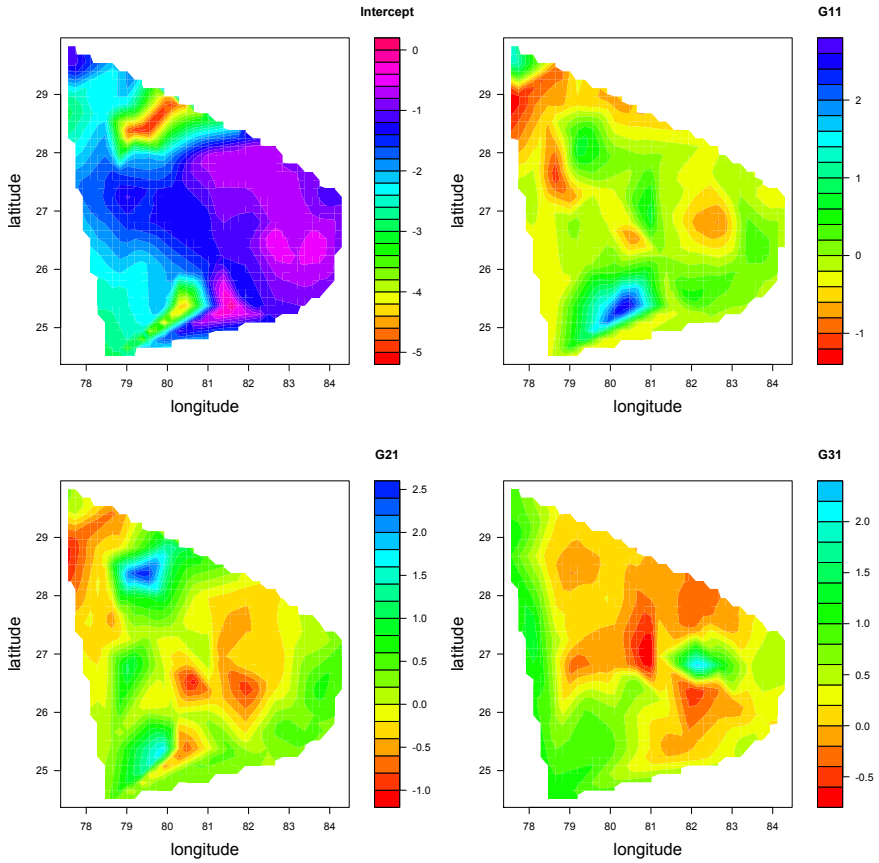

Figure 1. Maps showing the spatial variation in the District specific intercept and slope estimates that are generated when the GWR model is fitted to the NSSO data.

Table 1. Summary statistics for the NSGLMM parameter estimates, including the overall percentages of negative (% –) and positive (% +) values and parameter estimates for the global GLMM

	Intercept	G11	G21	G31
Descriptive Statistics		NSGLMM		
Minimum	-2.01	-0.33	-0.54	0.00
Q1	-1.67	-0.21	-0.16	0.10
Mean	-1.32	-0.12	-0.09	0.17
Median	-1.26	-0.13	-0.09	0.16
Q3	-1.01	-0.04	-0.02	0.19

Maximum	-0.70	0.14	0.21	0.53
% –	100	90.2	77.5	0
% +	0	9.8	22.5	100
		GLMM		
GLMM Model	-1.35	-0.21	-0.10	0.36

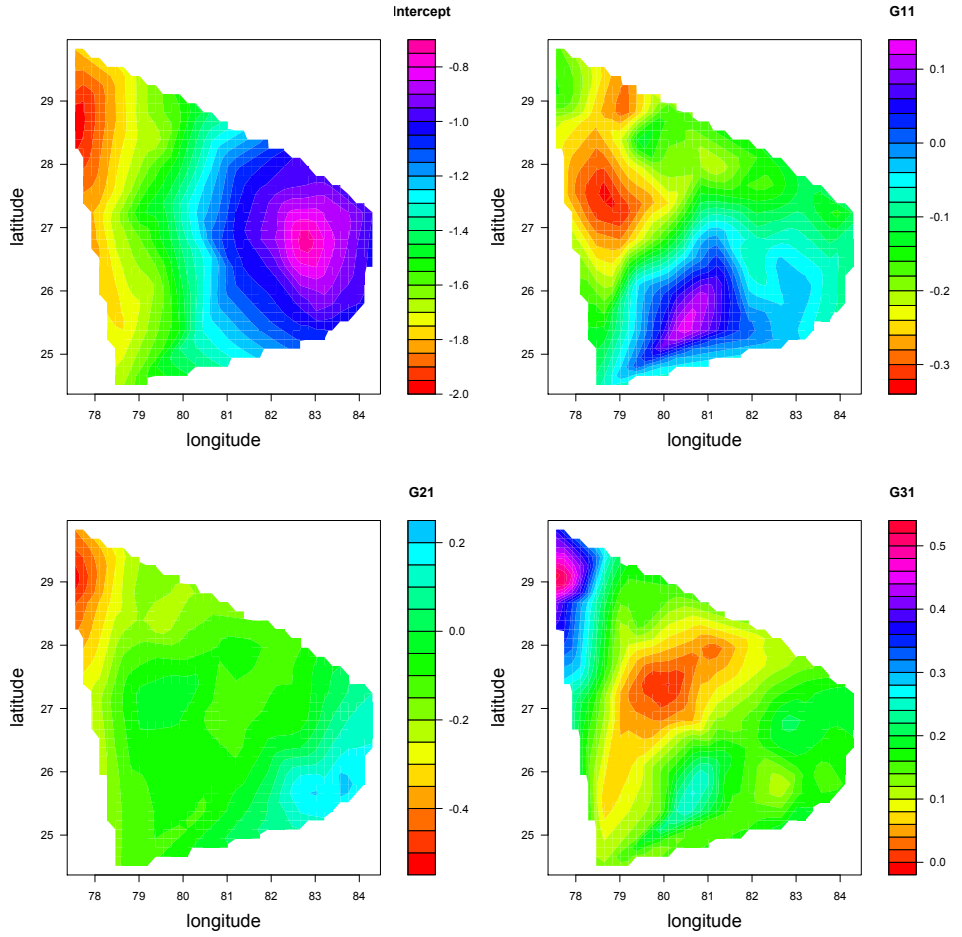

Figure 2. Maps showing the spatial variation of estimated regression coefficients $\beta(d_i) = \beta + \boldsymbol{\gamma}(d_i)$ when NSGLMM model is fitted to the NSSO data.

Brown *et al.* (2001) discuss diagnostics for SAE, rather than diagnostics for model fit. They note that small area estimates should be (a) consistent with unbiased direct survey estimates, i.e. they should provide an approximation to the direct survey estimates that is consistent with these values being "close" to the expected values of the direct estimates and (b) more precise than direct survey estimates, as evidenced by lower MSE estimates. We, therefore, consider three commonly used diagnostics developed by these authors for this purpose: the bias diagnostic, the goodness of fit (GOF) statistic and the percent coefficient of variation (CV) diagnostic.

The basic idea underpinning the bias diagnostic is that since direct estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If model-based small area estimates are close to these true values the regression of the direct estimates on these model-based estimates should be similar. We therefore plot direct estimates (y-axis) vs. model-based small area estimates (x-axis) and we looked for divergence of the fitted least squares regression line from the

line of equality. The scatter plot of the District-level direct estimates against the corresponding NSEP estimates is shown in Figure 3, with fitted least squares regression line (dashed line) and line of equality (solid line) superimposed.

Inspection of Figure 3 reveals that the District estimates generated by NSEP are less extreme than the direct survey estimates, demonstrating the typical SAE outcome of shrinkage towards the average. However, it is also clear that the NSEP estimates deviate somewhat from the line of equality at the extremes of their distribution. This is not unexpected, since the NSEP estimates are random variables and so the regression of the direct estimates on the NSEP estimates is biased for a test of common expected values. Such a test is provided by the GOF statistic, which is equivalent to a Wald test for whether the differences $D_i = \hat{y}_i^{direct} - \hat{y}_i^{NSSYN}$ have a zero mean, and is computed as $W = \sum_i \left\{ \frac{D_i^2}{\widehat{\text{var}}(\hat{y}_i^{direct}) + \widehat{mse}(\hat{y}_i^{NSSYN})} \right\}$, where $\hat{y}_i^{NSSYN}$ is the synthetic version NSSYN of the NSEP, defined following (5). See Brown *et al.* (2001). Under the assumption that $\hat{y}_i^{direct}$ and $\hat{y}_i^{NSSYN}$ are independently distributed, which is not unreasonable for large sample sizes, the value of W can be compared with an appropriate critical value from a chi square distribution with degrees of freedom D equal to the number of Districts. For our analysis, D = 71, with a critical value of 91.7 at a 5% level of significance. Here, we conclude that the NSEP estimates are consistent with the direct estimates.

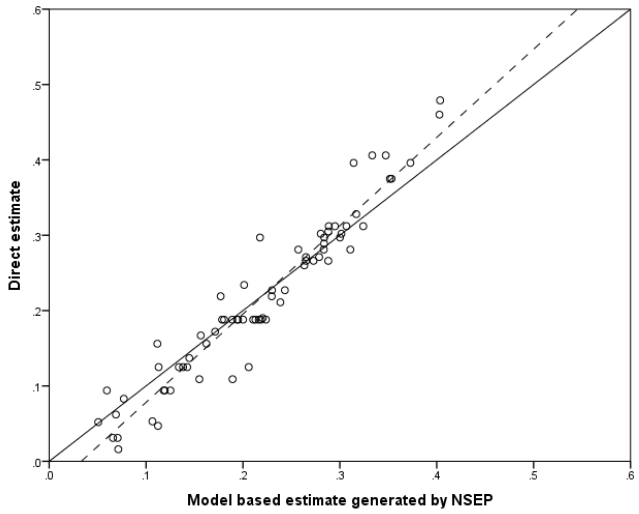

Figure 3. Bias diagnostic plot with $y = x$ line (solid) and regression line (dotted) for proportion of poor households in Uttar Pradesh: NSEP estimates versus direct estimates.

The final SAE diagnostic that we compute compares the extent to which the model-based small area estimates (EP and NSEP) improve in precision compared to the direct survey estimates (direct). We assess precision via the percentage CV of an estimate. Summary percentage CV values for the different SAE methods are shown in Table 2, while Figure 4 displays the District level values of percentage CV for the direct, EP and NSEP methods. These show that the direct survey estimators for the proportion of poor households within each District are unstable, with CVs varying from 8.67% to 232.44 %. Furthermore, the CVs of the direct estimators are greater

than 20% (40%) in 36 (16) out of the 71 Districts (Figure 4). Two small area estimators (EP and NSEP) are better than the direct estimator with respect to CV, with the NSEP outperforming the EP. In particular, the NSEP records smaller CVs than the EP in 53 of the 71 Districts of Uttar Pradesh.

Figure 5 is a map showing the proportions of poor households by District for Uttar Pradesh, estimated by the NSEP method. This clearly shows an east-west divide in the distribution of household poverty. For example, in the western part of Uttar Pradesh there are many districts with low poverty incidence (Saharanpur, Hathras, Meerut, Baghpat, Muzaffarnagar, Bulandshahar, etc). Similarly, in the eastern part and in the Bundelkhand region (north-east) we see many districts (Azamgarh, Sitapur, Chitrakoot, Bahraich, Siddharthnagar, Banda, Fatehpur, Basti and Kaushambi, etc) with high poverty incidence. This is an example of a "poverty map" showing reliable estimates of poverty incidence across a region of interest. This type of map is a useful aid for policy planners and administrators charged with taking effective financial and administrative decisions that can impact differentially across the region.

Table 2. Summary of area distributions of percentage coefficients of variation (CV, %) for different SAE methods applied to NSSO data.

Values	Direct	EP	NSEP
Minimum	8.67	6.72	6.54
Q1	13.94	12.70	11.42
Mean	32.61	21.58	20.67
Median	19.97	17.93	15.41
Q3	33.32	28.05	26.04
Maximum	232.40	57.38	70.02

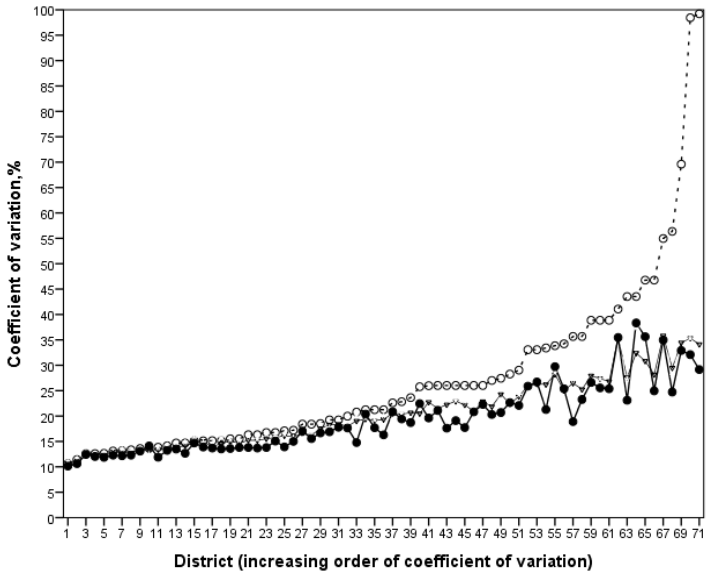

Figure 4. Percentage coefficients of variation (CV, %) by District for direct (dotted line, o) , EP (thin line, ∇) and NSEP(solid line, •) estimators applied to NSSO data.

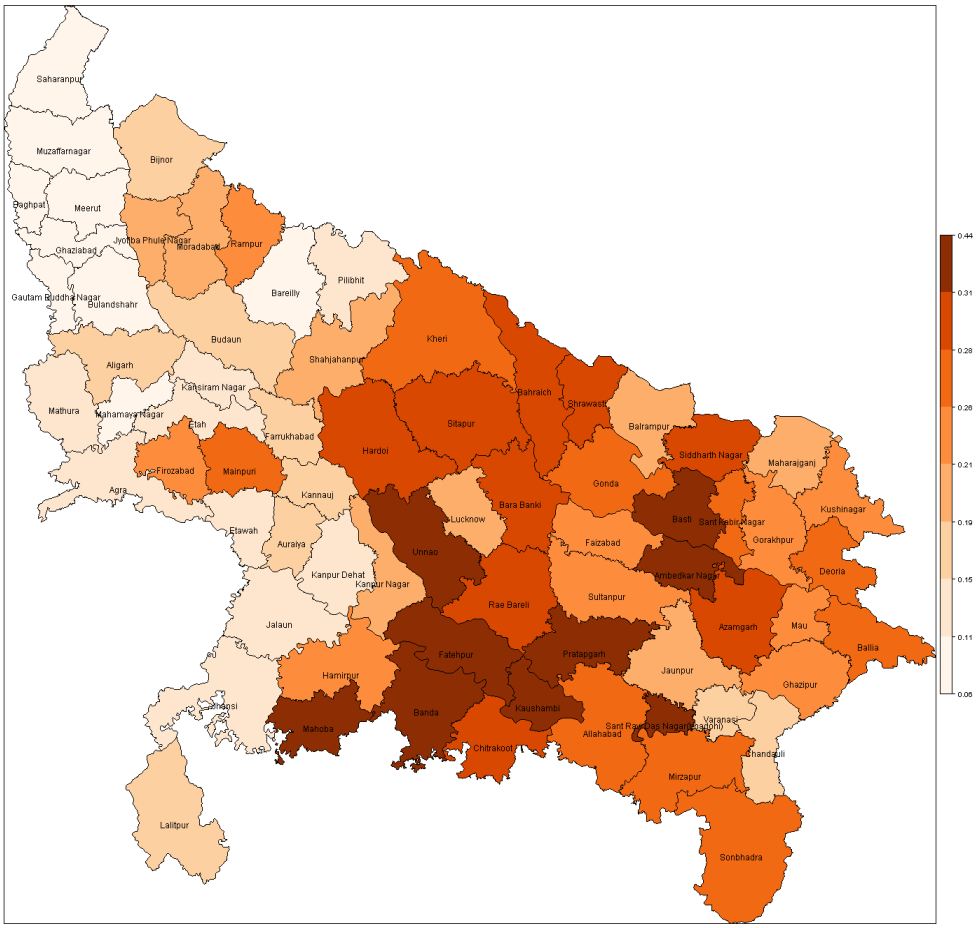

Figure 5. NSEP estimates showing the spatial distribution of proportions of poor households by District in Uttar Pradesh.

6 Conclusions

This paper describes plug-in non-stationary empirical predictor (NSEP) and plug-in empirical predictor (EP) for small area proportions. Our empirical results show that the NSEP is efficient than the EP in NSSO data with evidence of spatial non-stationarity. We therefore used the NSEP method to produce a poverty map showing how household poverty incidence varies by District across the State of Uttar Pradesh in India. We conclude by observing that the estimates and spatial distribution of poverty incidence generated from this research should be useful for meeting the data requirements for policy research and strategic planning by different international organizations and by Departments and Ministries in the Government of India.

References

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods - an application to the unemployment estimates from the UK LFS. In the Proceedings of the Statistics Canada Symposium. Achieving Data Quality in a Statistical Agency: A Methodological Perspective. *Statistics Canada*.

Chandra, H., Salvati, N. and Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics,* **20,** 30–56.

Chandra, H., Salvati, N. and Sud, U.C. (2011). Disaggregate-level Estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal of Applied Statistics*, **38(11)**, 2413-2432**.**

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.

Johnson, F.A., Chandra, H., Brown, J. and Padmadas, S. (2010). Estimating district-level births attended by skilled attendants in Ghana using demographic health survey and census data: an application of small area estimation technique. *Journal of Official Statistics,* **26 (2)**, 341–359.

Manteiga, G.W., Lombardìa, M.J., Molina, I., Morales, D. and Santamarìa, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, **51**, 2720-2733.

McGilchrist, C.E. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society B,* **56**, 61-69.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*. 2nd Edition. John Wiley and Sons.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.