Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 73(2) 2019 143–152

Hierarchical Bayes Aggregated Level Spatial Model for Crop Yield Estimation

Priyanka Anjoy and Hukum Chandra
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 20 September 2018; Revised 28 March 2019; Accepted 01 May 2019

SUMMARY

The demand for acceptable disaggregated level statistics from sample surveys has grown substantially over the past decades due to their extensive and varied use in public and private sectors. Basically, it is the main endeavor of 'Small area estimation (SAE)' approach to produce sound prediction of a target statistic for small domains to answer the problem of small sample sizes. The traditional survey estimation approaches are not suitable enough for generating disaggregate or small domain level estimates because of sample size problem. The SAE techniques therefore provide a feasible way to produce the reliable estimates at disaggregate level from the existing survey data.This paper explores a spatial dependent aggregated level Hierarchical Bayes (HB) model for SAE to estimate the yield for paddy (green) crop at district level in the state of Uttar Pradesh in India. The approach uses survey data from the Improvement of crop statistics (ICS) scheme collected by National Sample Survey Office (NSSO) and linked with Population Census. A considerable gain has been obtained while exploiting spatial information in aggregated level small area model.

Keywords: Aggregated level; Small area estimation; Spatial Information.

1. INTRODUCTION

In survey sampling we attempt to produce a concrete statistical inference about the target population and hence endeavors of statistical theory and methodology supporting efficient survey design prioritize this objective only. Relevantly, our target parameters may be the characteristics pertaining to whole population for which reliable design-consistent and design-unbiased direct estimates are available in the literature or we may either be interested to develop valid estimates for subpopulations (domains). Relying on design-based approach in generating official statistics for such subpopulations or domains may end up with result having poor precision or reliability, because domain-specific sample size is not generally large enough to guarantee reliable estimates for all the target subpopulations, such domains are also referred as 'Small areas'. Further, model-assisted approaches are design-based while assisted by models that result in more accurate design-unbiased estimates but still suffer from instability in case of small sample sizes. Therefore, model-dependent or model-based approaches are widely preferred and extensively used in producing acceptable small domain statistics (Rao, 2003). Eventually, model-based estimators suffer from design bias problem but their overall accuracy measures remain small. Following this advantage, the demand for large scale acceptable small area statistics from sample surveys has to be served through suitable "Small area estimation (SAE)" approaches.

Small areas may be geographical regions (e.g., districts, municipalities, blocks, tehsils, gram panchayets, etc.), particular demographic groups (e.g., age-sex-race groups within geographical areas) or cross classification of both. Sample sizes for such small domains are small enough (even zero) to avail acceptable direct estimates with adequate precision, hence it becomes necessary to employ oversampling to increase sample size or to "borrow strength" from related areas or time (or both) through specific linking

Corresponding author: Priyanka Anjoy
E-mail address: anjoypriyanka90@gmail.com

models. Oversampling is not a feasible approach, because that may leave other domains with small sample sizes as total sample sizes is fixed by the budget beforehand, hence incurring extra cost cannot be considered. Thus, we have to take recourse to small area estimation methodologies in order to generate reliable small area statistics which can efficiently be utilized for policy making, localized development at small domain levels. However, to draw needful and robust inference from small domains it is necessary that specified model for the situation under consideration is a good fit to the available sample data and again success of this kind of models also depends on good amount of covariates. Informative auxiliary variates may probably explain unknown structures in the data, along with this incorporation of area-specific random effects is crucial to model unstructured heterogeneity across areas which may not possibly be captured by fixed covariates. Based on the level of auxiliary information available, there are basically two kinds of small area models, e.g., area level models utilizes aggregated auxiliary and target information and unit level models are based on unit-specific variable information. In absence of unit-wise data, area level models are broadly accepted.

Fay and Herriot (1979) first described the small area estimation method under area level small area model to estimate the mean per capita income in predefined small areas within countries. Their model uses direct survey estimates and area level covariates to obtain small area estimates. The application of basic FH model and its various extensions are widely available in various real life studies and literatures to solve the small domain estimation problems (Pratesi and Salvati, 2009; Molina *et al.*, 2009; Chandra *et al.*, 2011; Chandra, 2013; Porter *et al.*, 2014; Pratesi and Salvati, 2016; Chandra *et al.*, 2017). Potentiality of such small area models to provide efficient small domain estimates has magnificent importance in micro or disaggregates level planning processes.

SAE models are based on linear mixed modeling framework and incorporate random area-specific effects which account for unstructured heterogeneity across areas beyond that is explained by auxiliary variables included in the fixed effect part of the model (Rao and Molina, 2015). However, an implicit independence assumption is also imposed on the random effect while modeling such component which implies different small areas are simply uncorrelated. But, such postulation may not hold good in practice as there appears no good reason why neighbouring areas should not be correlated (Chandra, 2013). Area boundaries defining the small domains are typically arbitrarily set, hence neighbouring effect can never be ignored. Particularly, as an example in agricultural, environmental data neighbouring areas exhibit strong spatial dependency and therefore independence assumption of random area effects seems questionable. In this regard, the article is set to explore spatial association between small areas via spatial model. Chandra (2013) and Chandra *et al.* (2017) has earlier attempted to investigate such spatial association via Simultaneous Autoregressive (SAR) and Spatial Non Stationary process under frequentist framework of SAE. In this paper we explore such spatial association via Bayesian prospect. One of the advantages of using Bayes framework is the direct quantification of uncertainty. In addition, this approach leads to more reasonable interval estimates.

In India, the yield rate estimates are developed on the basis of scientifically designed crop-cutting experiments (CCEs) conducted under the scheme of General Crop Estimation Surveys (GCES). More than 9,50,000 CCEs are conducted annually to cover 68 crops (52 food and 16 non-food) at national level. This sample size is sufficient enough to obtain precise estimates of crop yield (i.e. production per hectare of land) at the district level. But due to huge spectrum of work, along with some infrastructural and resource constraints, the data quality of GCES is questionable. To improve the quality of data collected under the GCES, a scheme titled Improvement of crop statistics (ICS) has been introduced by the Directorate of Economics and Statistics, Ministry of Agriculture and Farmers' Welfare, Government of India. Under this scheme, quality check on the field operation of GCES is carried out by supervising around 30,000 CCEs by NSSO and State Government supervisory officers. The findings of the ICS results reveal that the CCEs are generally not carried out properly resulting in data that lack acceptable quality (Chandra, 2013). Therefore the need is to reduce the sample size under GCES. However, with reduced sample sizes, the estimates produced at district level may not be reliable. Hence, to deal the problem of small sample sizes, SAE approach detailed in this article can be a

proficient alternative. Basically, our endeavor here is to explore SAR process under Hierarchical Bayes (HB) SAE framework to estimate yield for paddy crop at district level in the state of Uttar Pradesh in India. As an illustration, yield for paddy (green) crop at district level in the state of Uttar Pradesh in India is being taken up from ICS scheme by National Sample Survey Office (NSSO) and linked with Population Census.

2. DESCRIPTIONS OF DATA

In India each state are consist of districts and districts are important domains for planning process and policy formulation. In estimating quantities like yield, state level estimates are not able to represent regional scenario. Hence, SAE methodology sets an important step in deriving out micro level estimates through borrowing strength from related sources. In this article, we consider an aggregated level or area-level random effect model for SAE. Area level small area models require area-specific information on direct survey estimates and covariates. Here, the variable of interest is yield for paddy (green) crop. Data pertaining to supervised Crop cutting experiments (CCE) on paddy (green) crop under the ICS scheme for the kharif season for the state of Uttar Pradesh in India is collected during the year 2009–2010. The aim is to estimate the 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 m each and with total area of 43.30 m^2. Therefore, yield rate for paddy (green) crop is recorded as gram per 43.12 m^2. Auxiliary variables required for the study are obtained from Population census 2001. However, it is expected that covariates used in the study are not going to change significantly over a short period of time. In the state of Uttar Pradesh there are 70 districts; however, supervision on a sub-sample of CCEs work under the ICS scheme is carried out in 58 districts only and there is no sample data for the remaining 12 districts. So, these 12 districts are the out-of-sample districts. The area-specific sample sizes for 58 sample districts range from 4 to 28 CCEs with an average of 11 (Fig. 1). In few districts, the sample size is small so the traditional sample survey estimation approaches will not lead to reasonable estimate. Additionally in 12 non-sample districts, direct survey estimation approach which is based on only domain-specific sample data fails for such districts. SAE technique enables us to obtain precise small area estimates not only for districts with negligible sample sizes but also for non-sample districts, where the direct estimation approach typically incapable.

The auxiliary variables used in the study at area level are drawn from Indian Population Census 2001. Initially, there were 121 covariates available for study. Therefore, preliminary data analysis was carried out to select appropriate covariates for SAE modeling. First examining the correlation of each of the covariates with the target variable (direct survey estimates), variables with reasonable good correlation has been retained for further analysis. Finally, based on step-wise regression method two significant variables, average household size (HH_SIZE) and female population of marginal household (MARG_HH_F) with Akaike information criterion (AIC) value 1138.9 has been included for SAE. Note that for SAE of 12 out-of-sampled districts same two covariates may be used, since the underlying model for sample areas also holds for out-of-sample districts. The basic theory of SAE has been described in the next section considering Bayesian perspective. Hierarchical Bayes approach has been implemented for illustration purpose in estimating average paddy (green) crop yield at district level in the state of Uttar Pradesh.

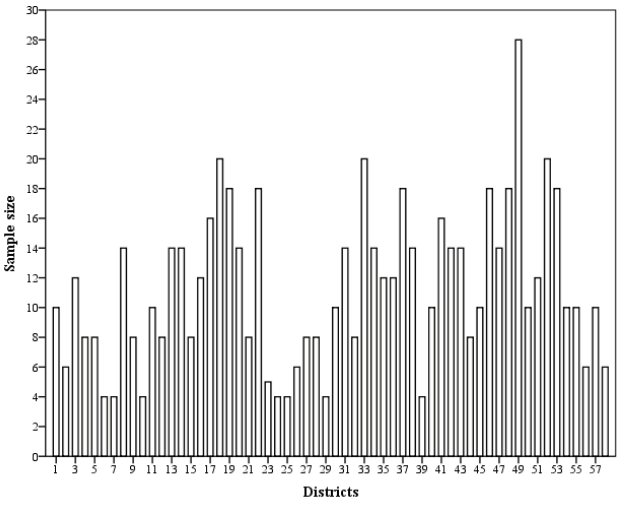

Fig. 1. Distribution of district-specific sample sizes in sample districts

3. HIERARCHICAL BAYES METHODOLOGY FOR SAE

Let us introduce some notation, U denotes the finite population of interest of size N partitioned into m disjoint small areas, a sample s of size n is drawn from this population with a given survey design.

The set of population units in area i is denoted as U_i with known size N_i, such that $\bigcup_{i=1}^{m} U_i = U$ and $\sum_{i=1}^{m} N_i = N$. Following standard practice, s and r denotes the units in the sampled and non-sampled parts of the population. With N_i and n_i respectively being the population and sample size from small area i ($i=1,...,m$), the units making up the sample in area i are denoted by s_i, so that $\bigcup_{i=1}^{m} s_i = s$ and $\sum_{i=1}^{m} n_i = n$. Basic area level FH model combines direct aggregate (district) level survey estimates with the available auxiliary information obtained from varied sources, e.g., census or administrative records. Thus the model has two components,

(1) Sampling model for the direct survey estimates
(2) Linking model to incorporate area-specific auxiliary information and random area effect through linear mixed modeling framework

For estimating small area population mean, assume that y_i denotes the direct survey estimate for unobservable population level quantities Y_i, hence the sampling model for y_i is expressed as follows,

$$\text{Sampling model: } y_i = Y_i + e_i, \text{i}=1,...,,m \qquad (1)$$

where, e_i's are independent sampling error assumed to have zero mean with known sampling variance σ_{ei}^2. Now, the linking model of Y_i can be written as,

$$\text{Linking model: } Y_i = \mathrm{x'_i}\,\beta + v_i,\ i = 1,...,m \qquad (2)$$

where, $\mathbf{x}_i'$ represent area level auxiliary information, $\boldsymbol{\beta}$ is the regression coefficient or fixed effect parameter and v_i being the area-specific random effect, independent and identically distributed as $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$. Random area-specific effects are included in the linking model to account for between areas variation beyond that is explained by auxiliary variables in the fixed part of the model. Two random errors v_i and e_i are independent of each other within and across areas (districts). So, the area level FH model can be written in combined form as below,

$$\mathrm{y}_i = \mathrm{x'}_i\beta + \mathrm{v_i} + e_i,\ i = 1,..., m \qquad (3)$$

Aggregating m-area-level model given by equation (3) leads to the population-level version of the random effect model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \qquad (4)$$

where $\mathbf{y} = (y_1,...,y_m)'$ is the $m \times 1$ vector of direct survey estimates, $\mathbf{X} = (\mathbf{x}_1',..., \mathbf{x}_m')'$ be $m \times p$ matrix of auxiliary variates, $\mathbf{Z} = (\mathbf{z}_1,...,\mathbf{z}_m)'$ is a matrix of known covariates of dimension $m \times m$ characterizing differences among small areas, $\mathbf{v} = (v_1,...,v_m)'$ is a $m \times 1$ vector of domain random effects and $\mathbf{e} = (e_1,...,\mathbf{e}_m)'$ is the m-component vector of sampling errors with $\mathbf{e} \sim N(0, \boldsymbol{\Omega})$ where $\boldsymbol{\Omega} = \mathrm{diag}\{\sigma_{ei}^2; 1 \le i \le m\}$ is the matrix of design variances.

In order to estimate small area population mean, several authors have considered EBLUP (Empirical Best Linear Unbiased Predictor) of Y_i considering frequentist perspective (Chandra, 2013; Pratesi and Salvati, 2008). But this approach requires an analytical expression of measure of precision which is based on very some approximation. In contrast, the strategic advantage in considering Bayesian approach is that, here estimations are described by assuming particular probability distributions, which render the opportunities to analyze the uncertainties involved in the decision process. Accordingly potential Bayesian analogue of FH model can be suitably considered rather its frequentist version. Particularly, Hierarchical Bayes (HB) method is implemented in this article employing Gibbs sampling approach. In the HB method, together with prior distribution of the parameters, prior of the hyper-parameters (model parameters) are also specified then inferences are made from the posterior distributions. A parameter is estimated by posterior mean and posterior variance is taken as the measure of the error or uncertainty of the estimates. HB approach can effectively deal with complex small area models using Monte Carlo Markov Chain (MCMC), which overcomes the computational difficulties of high-dimensional integrations of posterior densities (You and Rao, 2002; Anjoy *et al.*, 2018).The HB alternative of FH model (4) is expressed as below.

3.1 FH Model

$$\text{Sampling model: } \mathbf{y} \mid \mathbf{Y} \sim N(\mathbf{Y}, \Omega)$$

$$\text{Linking model: } \mathbf{Y} \mid \beta, \sigma_v^2 \sim N(\mathbf{X}\beta, \sigma_v^2 \mathbf{Z}\mathbf{Z}')$$

Choice of prior distributions plays a crucial role in Bayesian analysis, because inferences drawn from posterior densities depend on wide range of prior

distributions. Choice of improper or non-informative prior may be problematic due to small amount of data (Lambert and Sutton, 2005). Hence, selection of prior distributions should be preceded by sensitivity analysis under various parameterization processes. Generally, prior for the hyper-parameters (β, σ_v^2) are set as, individual β has improper uniform prior on R^p and $\sigma_v^2 \sim IG(a_0, b_0)$, where (a_0, b_0) are known positive quantity, usually set to be very small to reflect vague knowledge about σ_v^2. Here, distribution of β has been taken to be $N(0, 10^{-6})$ and distribution of σ_v^2 has been taken to be $IG(0.1, 0.1)$.

The FH model (4) is based on one of the implicit assumption that random area effects are distributed independently of each other. However, as evident from many real life examples, such as in agricultural, environmental or epidemiological data it is quite common to see that neighbouring small areas influences each other up to a great extent (Mercer *et al.,* 2014; Mercer *et al.,* 2015). Again such influences tend to decay as the distances between two areas increases. Keeping all this in regard, the obvious spatial association between neighbouring areas cannot be ignored in small area modeling while such correlation is specifically high enough. Furthermore, incorporation of such spatial information may significantly improve the model accuracy. Therefore, linear regression model with spatial dependence in error structure, particularly SAR error process is considered motivated from Chandra, 2013. But, our attempt here is to implement HB alternative of SAR model.

Let, define the domain random effect $\mathbf{v} = (v_1, ..., v_m)'$ satisfy,

$$\mathbf{v} = \rho \mathbf{W} \mathbf{v} + \mathbf{u} \tag{5}$$

where ρ is the spatial autoregressive coefficient measuring the strength of spatial relationship. $\mathbf{W}$ is a proximity matrix of order m which describes how random effects from neighboring areas are related. The elements of $\mathbf{W}$ take non-zero values only for those pairs of areas that are adjacent. Generally, for ease of interpretation, this matrix is defined in row-standardized form; in which case ρ is called the spatial autocorrelation parameter. Formally, the element w_{jk} $(j, k = 1,...,m)$ of a contiguity matrix takes the value 1 if area j shares an edge with area k and 0 otherwise. In a row-standardized form this becomes,

$$w_{jk} = \begin{cases} t_j^{-1} \text{ if } j \text{ and } k \text{ are contiguous,} \\ 0 \quad \text{otherwise} \end{cases} \tag{6}$$

where t_j is the total number of areas that share an edge with area j (including area j itself). Contiguity matrix $\mathbf{W}$ provides a simplest way to define spatial interaction between adjoining small areas. However, exploring this matrix in a better way while defining the elements w_{jk} as function of the length of shared border between neighboring areas or as a function of the distance between the area is crucial in producing reliable small area estimates. Let, l_i be the coordinates of an arbitrary spatial location (longitude and latitude) in i^{th} small area; generally this will be its centroid. Then the spatial distances between sample locations (l_i, l_j) can be given as $d_{j,k} = \| l_j - l_k \|$. Now consider various specification of weight matrix $\mathbf{W}$ defined as a function of distance measure.

1. Weights are assigned in proportion to distance between the areas as

$$\mathbf{W} = \{w_{jk}\} = d_{j,k} \tag{7}$$

2. Weights are defined as inverse of the distance between the areas as

$$\mathbf{W} = \{w_{jk}\} = \left(d_{j,k}\right)^{-1} \tag{8}$$

Now the HB version of SAR model is expressed as below,

3.2 SAR model

Sampling model: $\mathbf{y} \mid \mathbf{Y} \sim N(\mathbf{Y}, \mathbf{\Omega})$

Linking model: $\mathbf{Y} \mid \boldsymbol{\beta}, \rho, \sigma_u^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_u^2 \mathbf{Z}\left[(\mathbf{I}_m - \rho \mathbf{W})(\mathbf{I}_m - \rho \mathbf{W}')\right]^{-1} \mathbf{Z}')$

For SAR model, distribution of individual β has been taken to be $N(0, 10^{-6})$ and distribution of σ_u^2 has been taken to be $IG(0.1, 0.1)$. The value of ρ coefficient is treated as constant (obtained from the Spatial-EBLUP estimation based on frequentist approach). HB small domain estimates are computed for all the models using Metropolis-Hastings algorithm, drawing random samples from full conditional distributions of posterior quantities.

Here, four HB models will be explored including FH model and SAR model with three different structures for proximity matrix $\mathbf{W}$ as defined in

equation (6), (7) and (8). SAR models considering three different **W** structures denoted hereby as HB.S1, HB.S2 and HB.S3. The HB version of three SAR models, e.g., HB.S1, HB.S2 and HB.S3 is same except for different contiguity matrix.

4. EMPIRICAL ILLUSTRATIONS

This article attempts to study whether it is possible to achieve a substantial gain in spatially correlated random effect process (i.e., SAR model) over the non-spatial counterpart (i.e., FH model) as well as traditional direct estimation approach. HB version of EBLUP, HB.S1, HB.S2 and HB.S3 model is implemented therefore to see their relative proficiency in estimating yield of paddy (green) crop in each small area. To implement the Gibbs sampler, three independent chains were used each of length 20000. The first 10000 iterations were deleted as "burn-in" periods. Potential scale reduction factor $\hat{R}$ determines the convergence success. Stationarity is attained when $\hat{R} = 1$ (Rao, 2003). It is worth noting that choice of prior distribution plays a crucial role in Bayesian analysis because inferences drawn from posterior distribution depend on a wide range of prior distributions (Anjoy *et al.*, 2018). In this study the prior for σ_u^2 has been taken as $IG(0.1, 0.1)$. However other choices of prior, example $\sigma_u^2 \sim \text{uniform}(0,100)$ or $\sigma_u^2 \sim \text{uniform}(0,1000)$ can also be investigated.

Table 1 represents the descriptive statistics of CV% for direct estimates as well a small area model based estimates. Estimates with smaller CV% are preferred or more reliable than others. Comparing all the HB models, it is to be noted that the precision level of HB.S3 is much better than all other model based alternatives. HB.S3 has manifested substantially improved modeling power due to its special structure of proximity matrix where the weights are defined by decreasing function of distance. In direct estimation approach CV% was ranging from 3.01 to 49.15, whereas, in HB.S3 the range of CV% is 2.22–11.58. In terms of specifically mean, median, Q3 and maximum CV% all the model based method has outperformed over the direct survey based approach with HB.S3 being the best.

The reliability of model based estimates can be validated from bias diagnostics result also apart from CV%. The bias diagnostics are used to investigate whether the model based estimates are less extreme or not when compared to the direct survey estimates. If model based estimates are close to the true values, the regression of the direct estimates on the model based estimates should be similar. Plotting the direct estimates on the *Y*-axis and model-based estimates on the *X*-axis, divergence of regression line can be observed from $Y = X$ line and test for intercept = 0 and slope = 1. The bias scatter plots of the direct estimates against the model based estimates generated by different small area predictor are shown in Fig. 2. The results for the bias test are given in Table 2. The plots in Fig. 2 show that the model-based estimates are less extreme when compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. Reasonably good fit of HB.S3 is notable here. So, finally, HB.S3 can be opted as most proficient model based alternative to provide district level yield estimates.

Table 1. Summary of percentage coefficient of variation (%CV) generated by the different methods for sample districts

Values	Direct	EBLUP	HB.S1	HB.S2	HB.S3
Minimum	3.01	3.05	3.01	2.99	2.22
Q1	10.04	9.75	9.52	9.03	3.47
Mean	15.14	13.04	12.21	11.62	4.37
Median	13.42	12.39	11.92	11.27	4.09
Q3	19.46	16.68	15.35	13.80	4.71
Maximum	49.15	28.54	22.83	27.58	11.58

Table 2. Bias-diagnostics test results for different small area predictor

Predictor	Parameters	Estimate	Standard error	Probability
EBLUP	Intercept	2746.97	315.02	<0.0001
	Model based estimate	0.81	0.02	<0.0001
HB.S1	Intercept	2580.10	425.48	<0.0001
	Model based estimate	0.82	0.03	<0.0001
HB.S2	Intercept	2920.24	422.56	<0.0001
	Model based estimate	0.81	0.03	<0.0001
HB.S3	Intercept	10546.29	1002.66	<0.0001
	Model based estimate	0.33	0.06	<0.0001

District-wise estimates of yield implementing direct as well as HB.S3 model has been furnished in Table 3 along with 95% credible interval and CV% for both sample and out-of-sample districts. Direct

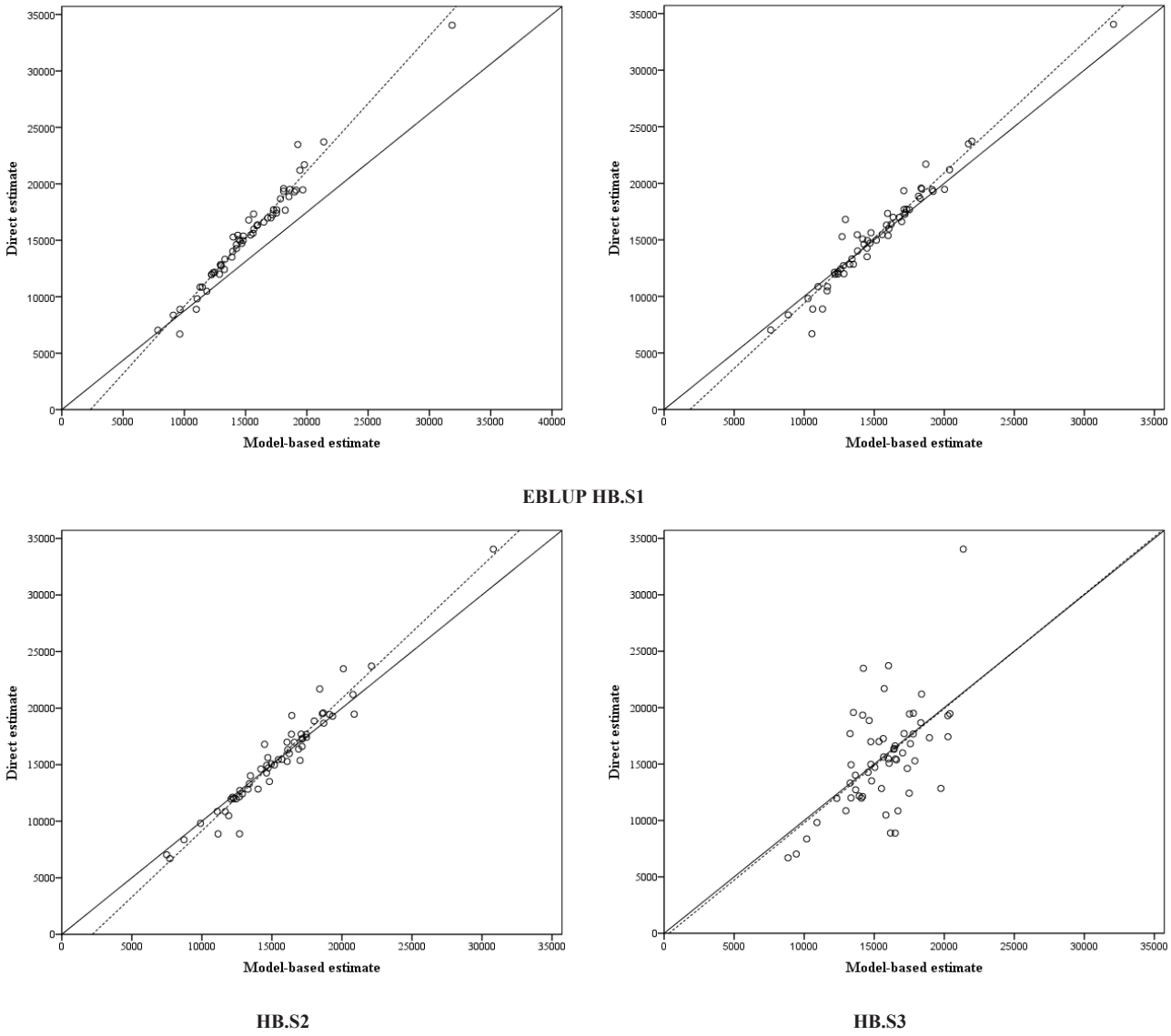

Fig. 2. Bias diagnostic plots for sample districts. Direct estimates versus model based estimates, y = x line (solid) and linear regression fit line (dashed)

estimation approach cannot provide estimate for out-of-sample districts. HB.S3 model therefore may be implemented for providing yield estimate for 12 out-of-sample districts based on auxiliary variable and estimate of hyper-parameter β. In Table 3, more than 20% CV for 11 districts has made the direct estimates in such districts highly unstable. A significant reduction in %CV has been achieved thereafter using the HB.S3 method over traditional direct estimation method, thus has resulted stable and precise yield estimates generated by HB.S3 method. Fig. 3 displays the graphical representation for district level values of % CV in increasing order respectively implementing direct and HB.S3 estimation methods. Average yield estimate aggregated over 58-sample districts by direct estimation approach is 15498 gram/43.12 m^2 and by HB.S3 estimation approach is 15635 gram/43.12 m^2.

5. CONCLUSIONS

The potentiality of SAE methodologies to generate reliable small domain inference is now quite established fact from varied theoretical researches, what needed is its real life implementation and applications. To strengthen the micro level planning, disaggregate level estimates are often required and small area models serve this purpose both adequately and efficiently. In this context, the current study also

Table 3. District-wise estimates of paddy (green) crop yield (gram per 43.12 m2) along with 95% confidence interval and percentage CV using direct and HB.S3 estimation approach

Districts	Direct estimation				HB estimation (HB.S3 model)			
	Estimates	Lower	Upper	%CV	Estimates	Lower	Upper	%CV
Saharanpur	19575	14574	24576	13.04	13513	12097	14929	5.34
Muzaffarnagar	23483	14035	32932	20.53	14223	13453	14994	2.76
Bijnor	19442	16669	22214	7.28	17522	16324	18719	3.49
Moradabad	17700	11916	23484	16.67	17135	16068	18202	3.18
Rampur	17250	16234	18266	3.01	15647	14968	16327	2.22
Jyotiba Phule Nagar	10850	7940	13760	13.68	16697	15413	17981	3.92
Ghaziabad	16800	6581	27019	31.03	17583	16138	19028	4.19
Bulandshahar	17418	13443	21393	11.64	20264	18571	21956	4.26
Aligarh	12419	7605	17232	19.77	17498	16331	18665	3.40
Mathura	10483	4880	16085	27.27	15837	14367	17307	4.73
Etah	12125	9813	14437	9.73	14166	12981	15352	4.27
Mainpuri	14019	7814	20224	22.58	13655	12707	14604	3.54
Budaun	12721	8968	16475	15.05	13676	12498	14855	4.40
Bareilly	13511	10021	17000	13.18	14814	13545	16084	4.37
Pilibhit	14938	9098	20777	19.94	13348	11928	14769	5.43
Shahjahanpur	18863	16560	21165	6.23	14640	13616	15663	3.57
Kheri	14975	11638	18312	11.37	14757	13455	16059	4.50
Sitapur	15986	11880	20093	13.11	17030	15627	18432	4.20
Hardoi	19286	16494	22078	7.39	20254	18970	21537	3.23
Unnao	12843	9841	15844	11.92	19758	18550	20966	3.12
Lucknow	17331	10170	24492	21.08	18938	17608	20269	3.58
Rae Bareli	19506	16053	22958	9.03	17798	16501	19095	3.72
Farrukhabad	8880	5582	12178	18.95	16503	15438	17567	3.29
Kannauj	34050	30416	37684	5.45	21360	19672	23048	4.03
Etawah	15463	13925	17000	5.07	16014	14858	17170	3.68
Auraiya	23717	19085	28348	9.96	16021	14588	17453	4.56
Kanpur Dehat	21200	16705	25695	10.82	18383	17116	19649	3.51
Kanpur Nagar	15375	10172	20578	17.27	16589	15249	17930	4.12
Banda	8888	326	17449	49.15	16166	14752	17579	4.46
Fatehpur	14612	8853	20371	20.11	17366	16200	18533	3.43
Pratapgarh	16304	11665	20942	14.52	16413	15384	17441	3.20
Kaushambi	15450	7295	23605	26.93	16511	15425	17598	3.36
Allahabad	19465	14994	23936	11.72	20400	18022	22778	5.95
Barabanki	18668	14600	22736	11.12	18328	16996	19661	3.71
Faizabad	16379	11802	20957	14.26	16419	14874	17963	4.80
Ambedkar Nagar	17692	14417	20966	9.44	13275	12687	13862	2.26
Sultanpur	16609	13493	19725	9.57	16490	14895	18085	4.93
Bahraich	14714	13593	15835	3.89	15044	14189	15900	2.90
Shrawasti	15075	9490	20660	18.90	16065	14885	17245	3.75
Balrampur	11975	8541	15409	14.63	14084	13158	15010	3.35
Gonda	16981	14828	19134	6.47	14752	13903	15601	2.94
Siddharth Nagar	12829	9422	16235	13.55	15522	13959	17085	5.14
Basti	14268	9736	18800	16.21	14557	13233	15882	4.64

Districts	Direct estimation				HB estimation (HB.S3 model)			
	Estimates	Lower	Upper	%CV	Estimates	Lower	Upper	%CV
Sant Kabir Nagar	13319	11660	14978	6.35	13284	12114	14455	4.50
Mahrajganj	21690	16526	26854	12.15	15716	14573	16859	3.71
Gorakhpur	12164	9129	15199	12.73	13924	12979	14869	3.46
Kushinagar	19343	13702	24984	14.88	14175	12583	15768	5.73
Deoria	8364	5482	11246	17.58	10177	8950	11403	6.15
Azamgarh	11957	9961	13953	8.52	12326	11219	13433	4.58
Mau	9820	6039	13601	19.64	10916	9510	12321	6.57
Ballia	7029	4167	9892	20.78	9435	7732	11138	9.21
Jaunpur	16990	13571	20409	10.27	15336	14106	16566	4.09
Ghazipur	10858	8029	13687	13.29	12976	11937	14016	4.09
Chandauli	12000	7382	16618	19.63	13342	12293	14391	4.01
Varanasi	17665	12341	22989	15.38	17789	15800	19777	5.70
Sant Ravidas Nagar	6693	1943	11443	36.21	8845	6837	10852	11.58
Mirzapur	15625	12039	19211	11.71	15671	14218	17125	4.73
Sonbhadra	15283	7347	23220	26.49	17906	15229	20583	7.63
Meerut*					15112	13115	17109	6.74
Baghpat*					12337	10158	14516	9.01
Gautam Buddha Nagar*					16757	14769	18745	6.05
Hathras*					15333	13259	17406	6.9
Agra*					14957	12895	17019	7.03
Firozabad*					14439	12362	16516	7.33
Jalaun*					15197	13172	17222	6.79
Jhansi*					17695	15628	19762	5.96
Lalitpur*					17040	15000	19080	6.1
Hamirpur*					16612	14611	18613	6.15
Mahoba*					16307	14258	18355	6.41
Chitrakoot*					14869	12837	16901	6.97

* Indicates out-of-sample districts

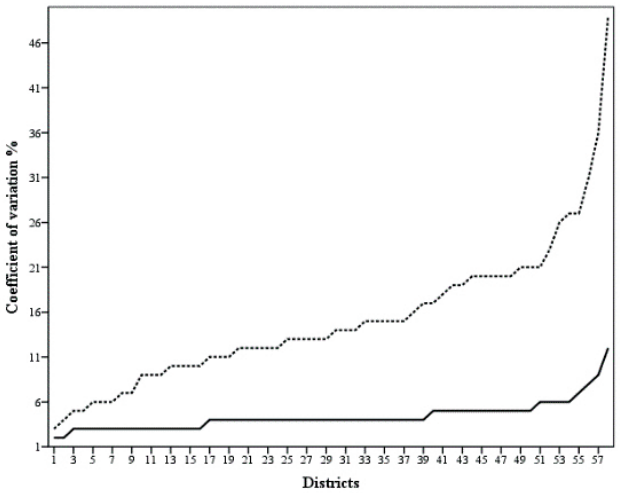

Fig. 3. District-wise percentage CV for direct (dotted line) and HB (solid line) estimation method (HB.S3)

reflects a suitable example of why small area model based methods should be preferred. Along with this, the relative proficiency of using existing spatial information in aggregated level small area model is also established than the non-spatial alternative. However, the study can also be extended to account for spatial non-stationarity and the same can be implemented in solving various small domain inference problems. As a profound application, the suitability of this study can be found in schemes like Pradhan Mantri Fasal Bima Yojana (PMFBY) to generate the micro level estimates of crop yield from existing survey data.

ACKNOWLEDGEMENT

We would like to express our sincere thanks to the anonymous reviewer for the valuable suggestions that helped us a lot in improving this manuscript.

REFERENCES

Anjoy, P., Chandra, H. and Basak, P. (2018). Estimation of Disaggregate-Level Poverty Incidence in Odisha under Area-Level Hierarchical Bayes Small Area Model. *Social Indicators Research*, DOI :10.1007/s11205-018-2050-9.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2002). *Geographically Weighted Regression*. New York: John Wiley and Sons.

Chandra, H., Salvati, N. and Sud, U.C. (2011). Disaggregate-level estimates of indebtedness inthe state of Uttar Pradesh in India – An application of small area estimation technique. *J. Applied statistics,* **38**, 2413-2432.

Chandra., H. (2013). Exploring spatial dependence in area-level random effect model for disaggregate-level crop yield estimation. *J. Applied statistics,* **40**, 823-842.

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

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.,* **74**, 269-277.

Lambert, P.C. and Sutton, A.J. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distribution in MCMC using WinBUGS. *Statistics in Medicine,* **24**, 2401-2428.

Mercer, L., Wakefield, J., Chen, C. and Lumely, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics,* **8**, 69-85.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A.M., Masanja, H. and Clark, S. (2015). Space –time smoothing of complex survey data: small area estimation for child mortality. *The Annals of Applied Statistics,* **9**,1889-1905.

Molina, I., Salvati, N. and Pratesi, M. (2009). Bootstrap for estimating the MSE of the spatial EBLUP. *Computational Statistics,* **24**, 441-458.

Porter, A.T., Holan, S.H., Wikle, C.K. and Cressie, N. (2014). Spatial Fay-Herriot models for small area estimation with functional covariates. *Spatial Statistics,* **10**, 27-42.

Pratesi, M. and Salvati, N. (2008). *Small area estimation:* The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Application,* **17**, 114-131.

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

Pratesi, M. and Salvati, N. (2016). Introduction on measuring poverty at local level using small area estimation methods. *Pratesi/ Analysis of poverty data by small area estimation*, 1-18. New York: John Wiley and Sons.

Rao, J.N.K. 2003. *Small area estimation*. New York: John Wiley and Sons.

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

You, Y. and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian J. Statist.,* **30**, 3-15.