Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 25-38

District-level Estimates of Extent of Food Insecurity for the State of Uttar Pradesh in India by Combining Survey and Census Data

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

Received: 18 April 2020; Revised: 26 April 2020; Accepted: 02 May 2020

Abstract

This paper describes small area estimation (SAE) method that incorporates the sampling information when estimating small area proportions. This method is applied to estimate the incidence of food insecurity in different districts of rural areas of the state of Uttar Pradesh in India by linking data from the 2011-12 Household Consumer Expenditure Survey collected by the National Sample Survey Office of India and the 2011 Population Census. A map showing district level inequalities in the distribution of food insecure households in Uttar Pradesh is also produced which provides an important information for analysis of spatial distribution of food insecurity in the state.

Key words: Food insecurity; SDG; Small area estimation; Precise, Representative.

1. Introduction

The food security is one of the highest priority of the Government of India to achieve the Sustainable Development Goal 2. In India, the Household Consumer Expenditure Survey (HCES) data collected by National Sample Survey Office (NSSO), Ministry of Statistics and Program Implementation, Government of India is used to generate the estimates of food insecurity indicators at state and national level for both rural and urban sectors separately. In spite of high importance, the estimates of food insecurity indicators are not available at local area or lower administrative unit (*e.g.* district) level in the country. Policy planners, researchers, government and public agencies are more and more interested in obtaining statistical summaries for smaller domains called small areas, created by cross classifying demographic and geographic variables such as small geographic areas (*e.g.* districts) or small demographic groups (*e.g.* age-sex groups, land category, social groups) or a cross classification of both. However, the sample sizes for such small areas in the existing large scale survey data (*e.g.* HCES in India) may be very small or even zero. The SAE methodology provides a viable and cost effective solution this problem of small sample sizes (Rao and Molina, 2015). The SAE methods produce reliable estimates for small areas with small sample sizes by borrowing strength from data of other areas, other time periods or both.

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

The SAE methods are generally based on model-based methods. The idea is to use statistical models to link the variable of interest with auxiliary information, *e.g*. Census and Administrative data, for the small areas to define model-based estimators for these areas. 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, Chandra (2013) and Chandra *et al* (2015). Standard SAE methods based on linear mixed models for continuous data can produce inefficient and sometime invalid estimates when the variable of interest is binary. If the variable of interest is binary and the target of inference is a small area proportion (*e.g.* for estimating food insecurity proportions), then the generalized linear mixed model with logit link function, also referred as the logistic linear mixed model (LLMM)) is generally used. An empirical plug-in predictor (EPP) under a LLMM is commonly used for the estimation of small area proportions, see for example, Chandra *et al*. (2012), Rao and Molina (2015) and references therein, although it is not the most efficient predictor under that model. An alternative to EPP is the empirical best predictor (EBP, Jiang, 2003). This predictor does not have a closed form and can only be computed via numerical approximation. This is generally not straightforward, and so national statistical agencies favour computation of an approximation like the EP.

In this context, when only area level data are available, an area level version of a LLMM is used for SAE, see for example, Johnson *et al*. (2010), Chandra *et al*. (2011), Chandra *et al*. (2017), Chandra *et al*. (2018), Anjoy *et al*. (2020). Unlike the Fay-Herriot model, this approach implicitly assumes simple random sampling with replacement within each area and ignores the survey weights. Unfortunately, this has the potential to seriously bias the estimates if the small area samples are seriously unbalanced with respect to key population characteristics, and consequently use of the survey weights appears to be inevitable for if one wishes to generate representative small area estimates. Chandra *et al*. (2019) deliberated the idea of Korn and Graubard (1998) and model the survey weighted estimates as binomial proportions, with an "effective sample size" chosen to match the binomial variance to the sampling variance of the estimates. Using the effective sample size rather than the actual sample size allows for the varying information in each area under complex sampling. This article considers Chandra *et al*. (2019) approach to model survey weighted small area proportions under a LLMM and attempts to produce the district level estimates of proportion of food insecurity (also refers as food insecurity prevalence or incidence of food insecurity) for rural areas of Uttar Pradesh. Throughout this article, proportion of food insecurity, food insecurity prevalence and incidence of food insecurity will be used interchangeably. The state of Uttar Pradesh is the most populous state in the country and accounts for about 16.16 percent of India's population. It covers 243,290 square km, equal to 6.88% of the total area of the country. The analysis is restricted to rural areas of Uttar Pradesh because about 78% of the population of the State live in rural areas according to 2011 Population Census.

Rest of the article is organized as follows. Next Section describes the data from the 2011-12 HCES of the NSSO and the 2011 Population Census that will be used to

estimate the district-wise proportion of household food insecurity for rural areas of Uttar Pradesh. Section 3 presents the SAE methodology. The empirical results and a map showing district-level inequalities in the distribution of food insecurity in rural Uttar Pradesh along with various diagnostic measures are reported in Section 4. Finally, Section 5 provides concluding remarks.

2. Data and Model Specification

This section introduces the basic sources of the data, *i.e.* the 2011-12 HCES of the NSSO for rural areas of Uttar Pradesh and the 2011 Population Census, used in SAE application reported in this paper. Data obtained from these sources are then used to estimate the proportion of food insecurity (or incidence of food insecurity) at district level in Uttar Pradesh. The NSSO conducts nationwide HCE surveys at regular intervals as part of its "rounds", with the duration of each round normally being a year. The surveys are conducted through interviews of a representative sample of households selected randomly through a suitable sampling design and covering almost the entire geographical area of the country. The sampling design used in the 2011-12 HCES 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. Although district is a very important domain of the planning process in India, there are no surveys aimed at producing estimates at this level. The lack of robust and reliable outcome measures at the district level puts constraints on the design of targeted interventions and policy development. In the 2011-12 HCES, a total of 5916 households from the 71 districts of rural areas 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 (see Table 1). Due to this sample size limitation, it is challenging to generate reliable district level direct estimates with associated standard errors from this survey (Rao and Molina, 2015 and Chandra *et al.*, 2011). This paper addresses this small sample size issue in the 2011-12 HCES data for producing district level estimates by adopting SAE approach and using auxiliary information from the 2011 Population Census to strengthen the limited sample data from the districts.

Table 1: Summary of sample size, number of food insecure households in sample (sample count) and sampling fraction in 2011 HCES data

Features	Minimum	Maximum	Average	Total
Sample size	32	128	83	5915
Sample count	10	111	53	3778
Sampling fraction	0.00015	0.00032	0.00023	0.01647

The target variable Y at the unit (household) level in the 2011-12 HCES survey data file is binary, corresponding to whether a household is food insecure (household consuming less than 2400 Kcal per day) or not. Average dietary energy intake per person per day in rural India is 2400 kilocalorie (Kcal), as defined by the Ministry of Health and Family Welfare, Government of India. The target is to estimate the proportion of rural households that are not getting satisfactory proportion of calories

consistently at small area level, also referred to as the incidence of food insecurity or proportion of household food insecurity.

As noted above, the auxiliary variables used in this analysis are taken from the 2011 Population Census of India. These auxiliary variables are only available as counts at district level, and so SAE methods based on area level small area models must be employed to derive the small area estimates. There are nearly 30 such auxiliary variables that are available for use in SAE analysis. We, therefore, carried out an exploratory data analysis to choose few auxiliary variables to determine appropriate covariates for SAE modelling. We also employed Principal Component Analysis (PCA) to derive composite scores for some selected groups of variables. In particular, we did PCA separately on two groups of variables, all measured at district level and identified as S1 and S2 below. The first group (S1) consisted of the proportions of main workers by gender, proportions of main cultivators by gender and proportions of main agricultural labourers by gender. The first principal component (S11) for this first group explained 44% of the variability in the S1 group, while adding the second component (S12) increased explained variability to 69%. The second group (S2) consisted of proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. The first principal component (S21) for this second group explained 52% of the variability in the S2 group, while adding the second component (S22) increased explained variability to 90%.

We fitted a generalised linear model using direct estimates of proportions of food insecure households as the response variable and the four principal component scores S11, S12, S21, S22 and few other selected auxiliary variables from the 2011 Population Census as potential covariates. The final selected model included five covariates namely proportional scheduled caste population (SC), literacy rate (Lit), proportion of working population (WP), index for main worker population (S11) and index for marginal worker population (S21), with Akaike Information Criterion (AIC) value of 636.34. For this model, null deviance is 430.88 on 70 degrees of freedom and including the five independent has decreased the deviance to 294.72 on 65 degrees of freedom, a significant reduction in deviance. The residual deviance has reduced by 136.16 with a loss of five degrees of freedom. We use Hosmer Lemeshow goodness of fit test to examine the fitted model (*i.e.* model fits depends on the difference between the model and the observed data). The *p*-value of Hosmer Lemeshow goodness is 0.9987. This indicates that model appears to fit well because we have no significant difference between the model and the observed data (*i.e*. the *p*-value is above 0.05). In this fitted model it can be noted that SC, Lit, WP, S11 influence proportion of food insecure households positively, while S21 has a slightly negative effect. Further, the coefficients of SC (–1.3741), Lit (–1.10334), WP (–5.0617), S11 (–0.385) and S21 (0.3123) are significant ($p < 0.001$). This final model was then used to produce district wise estimates of food insecurity.

3. Small Area Estimation Methodology

Let us assume that a finite population U of size N consists of D non-overlapping and mutually exclusive small areas (or areas), and a sample s of size n is drawn from this population using a probability sampling method. We use a subscript d

to index quantities belonging to small area *d*. Let U_d and s_d be the population and sample of sizes N_d and n_d in area *d*, respectively such that $U = \bigcup_{d=1}^{D} U_d$, $N = \sum_{d=1}^{D} N_d$, $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. We use subscript *s* and *r* respectively to denote quantities related to sample and non-sample parts of the population. Let y_{di} denotes the value of the variable of interest for unit *i* $(i = 1,...,N_d)$ in area *d*. The variable of interest, with values y_{di}, is binary (*e.g.*, $y_{di} = 1$ if household *i* in small area *d* is food insecure and 0 otherwise), and the aim is to estimate the small area population count, $y_d = \sum_{i \in U_d} y_{di}$, or equivalently the small area proportion, $P_d = N_d^{-1} y_d$, in area *d*. The standard direct estimator (denoted by Direct) for P_d is $\hat{p}_d^{Direct} = \left(\sum_{i \in s_d} w_{di} \right)^{-1} \sum_{i \in s_d} w_{di} y_{di}$, where w_{di} is the survey weight for unit *i* in area *d*. The estimate of variance of direct estimator is $v(\hat{p}_d^{Direct}) \approx \left(\sum_{i \in s_d} w_{di} \right)^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1)(y_{di} - \hat{p}_d^{Direct})^2$. Under simple random sampling (SRS), $\hat{p}_d^{Direct} = n_d^{-1} y_{sd}$, and $v(\hat{p}_d^{Direct}) \approx n_d^{-1} p_d (1 - p_d)$, where $y_{sd} = \sum_{i \in s_d} y_{di}$ denotes the sample count in area *d*. Similarly, $y_{rd} = \sum_{i \in s_r} y_{di}$ denotes the non-sample count in area *d*. If the sampling design is informative, this SRS-based version of Direct may be biased. If we ignore the sampling design, the sample count y_{sd} in area *d* can be assumed to follow a Binomial distribution with parameters n_d and π_d, i.e. $y_{sd} | u_d \sim \text{Bin}(n_d, \pi_d)$. Similarly, for the non-sample count, $y_{rd} | u_d \sim \text{Bin}(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent binomial variables with π_d being a common success probability. This leads to $E(y_{sd} | u_d) = n_d \pi_d$ and $E(y_{rd} | u_d) = (N_d - n_d)\, \pi_d$.

Let $\mathbf{x}_d$ be the *k*-vector of covariates for area *d* from available from secondary data sources. Following Johnson *et al.* (2010), Chandra *et al.* (2011) and Anjoy *et al.* (2020), the model linking the probability π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model (LLMM) of form

$$logit(\pi_d) = \ln\left\{ \pi_d (1 - \pi_d)^{-1} \right\} = \eta_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d, \tag{1}$$

with $\pi_d = \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d) \left\{ 1 + \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d) \right\}^{-1} = \text{expit}(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)$. Here $\boldsymbol{\beta}$ is the *k*-vector of regression coefficients and u_d is the area-specific random effect that capture the area dissimilarities. We assume that u_d is independent and normally distributed with mean zero and variance σ_u^2. The total population counts y_d can be written as $y_d = y_{sd} + y_{rd}$, where y_{sd}, the sample count is known whereas y_{rd}, the non-sample count, is unknown. Under (1), a plug-in empirical predictor (EPP) of y_d in area *d* is

$$\hat{y}_d^{EPP} = y_{sd} + (N_d - n_d) \left[\text{expit}(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{u}_d) \right]. \tag{2}$$

An estimate of the corresponding proportion in area d is $\hat{p}_d^{EPP} = N_d^{-1}\hat{y}_d^{EPP}$. It is obvious that in order to compute the small area estimates by equation (2), we require estimates of the unknown parameters $\boldsymbol{\beta}$ and $\mathbf{u} = (u_1,...,u_D)^T$. We use an iterative procedure that combines the Penalized Quasi-Likelihood estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with REML estimation of σ_u^2 to estimate unknown parameters.

The model (1) is based on unweighted sample counts, and hence it assumes that sampling within areas is non-informative given the values of the contextual variables and the random area effects. The EPP predictor based on (2) therefore ignores the complex survey design used in HCES data. But, the sampling design used in HCES is informative. The precision of an estimate from a complex sample can be higher than for a simple random sample, because of the better use of population data through a representative sample drawn using a suitable sampling design. Following Chandra *et al.* (2019), we model the survey weighted probability estimate for an area as a binomial proportion, with an "effective sample size" that equates the resulting binomial variance to the actual sampling variance of the survey weighted direct estimate for the area. Hence, in our analysis we replaced the "actual sample size" and the "actual sample count" with the "effective sample size" and the "effective sample count" respectively. The mean squared error (MSE) estimation is followed from Chandra *et al.* (2019).

4. Results and Discussions

In this Section we first examine if sampling design in HCES sample data is informative. The sampling design is called informative design if the distribution in the sample is different from the distribution in the population. Such sampling design is also referred as non-ignorable design. The sampling design used in survey data collected must be incorporated in making the valid analytic inference about the population. For this purpose, we compute the effective sample sizes and the effective sample counts for the HCES data. Readers are suggested to refer Chandra *et al.* (2019) for details about calculation of the effective sample sizes and the effective sample counts.

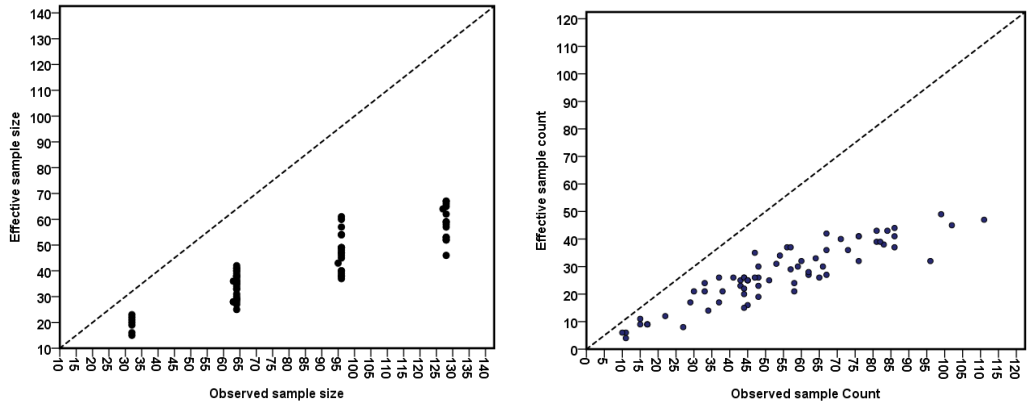

Figure 1: **Effective sample size versus observed sample size (left) and effective sample count versus observed sample count (right) in 2011 HCES data**

Figure 1 plots the effective sample sizes against the observed sample sizes (left side) and the effective sample counts against the observed sample counts (right side). It is evident from Figure 1 that the effective sample size is smaller than the observed sample sizes in almost all the districts. Similarly, the effective sample counts is lower than the observed sample counts. This indicates that the sampling design results in a loss in information, when compared with simple random sampling, in all the districts.

Figure 2 presents the district-wise survey weighted and unweighted direct estimates of proportion of household food insecurity. It can be seen from Figure 2 that the unweighted direct estimates underestimate the proportion of food insecurity, in majority of the districts. These examples are evident that the sampling design is informative and therefore must be accounted in SAE. Following the idea of Korn and Graubard (1998) and Chandra *et al.* (2019), we use the effective sample sizes in replace of observed sample sizes to incorporate the sampling design of HCES data.

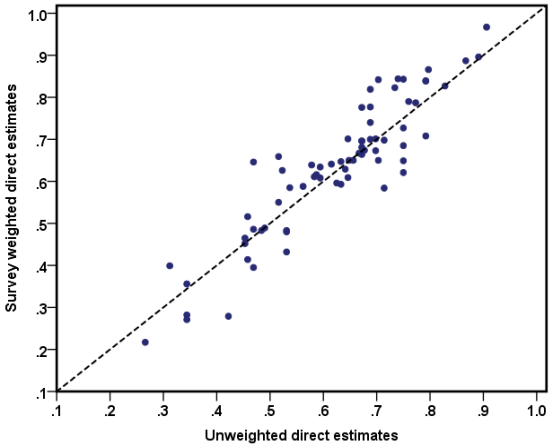

Figure 2: **District-wise survey weighted direct estimates versus unweighted direct estimates of proportion of food insecure households**

The estimates of proportion of food insecurity (or incidence of food insecurity) at district level for rural areas in the state of Uttar Pradesh is generated from the EPP method described in Section 3 using 5 significant covariates described in Section 2. Here we assume a binomial specification for the "effective" district level sample counts of food insecurity. Some important diagnostics measures are now discussed to examine the assumptions of the underlying models, and to validate the empirical performances of the EPP method. Generally, two types of diagnostics measures are advised in SAE applications. These are (i) the model diagnostics, and (ii) the diagnostics for the small area estimates. See Brown *et al.* (2001). The model diagnostics are applied to verify model assumptions. The other diagnostics are used to validate reliability of the model-based small area estimates of incidence of food insecurity generated by the EPP method. In LLMM (1) the random specific effects are assumed to have a normal distribution with mean zero and fixed variance. If the model assumptions are satisfied then the district level residuals are expected to be randomly distributed around zero. Histogram and normal probability (q-q) plot can be used to examine the normality assumption. Figure 3 shows the histogram (left plot), the normal probability (q-q) plot (centre plot) and the distribution of the district-level residuals (right plot). We also use the Shapiro-Wilk test (implemented using the *shapiro.test()* function in R) to examine the normality of the district random effects.

The Shapiro-Wilk test with p-value lower than 0.05 indicate that the data deviate from normality. Here, the value of Shapiro-Wilk test statistics is 0.988 with 71 degree of freedom and p-value 0.746. In Figure 3, the district level residuals appear to be randomly distributed around zero. Further, histogram and the q-q plot also provide evidence in support of the normality assumption. The Shapiro-Wilk p-value is larger than 0.05 and hence, the district random effects are likely to be normally distributed.

Following Chandra *et al.* (2011) and Brown *et al.* (2001), we use three commonly used measures for assessing the validity and the reliability of the model-based estimates generated by the EPP: the bias diagnostic, the percent coefficient of variation (CV) diagnostic and the 95 percent confidence interval diagnostic. The first diagnostics assesses the validity and last two assess the improved precision of the model based small area estimates. We also implemented a calibration diagnostic where the EPP estimates are aggregated to higher level and compared with direct estimates at this level. The bias diagnostic is based on following idea. The direct estimates are unbiased estimates of the population values of interest (i.e. true values), 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.

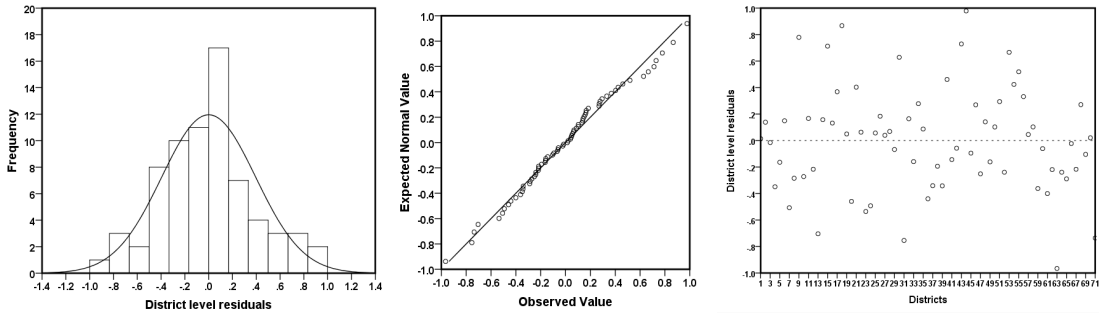

Figure 3: Histograms (left plot), normal q-q plots (centre plot) and distributions of the district-level residuals (right plot)

Figure 4 provides a bias diagnostic plot, defined by plotting direct estimates (*Y* axis) against corresponding small area estimates generated by the EPP (*X*-axis) and testing for divergence of the fitted least squares regression line (dashed line) from the line of equality, i.e. $Y = X$ line (solid line). The bias diagnostic plot in Figure 4 clearly indicate that the EPP estimates are less extreme when compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. The value of R^2 for the fitted regression line between the direct estimates and the EPP estimates is 95.6 per cent. The bias diagnostics indicates that the estimates generated by the EPP appear to be consistent with the direct estimates.

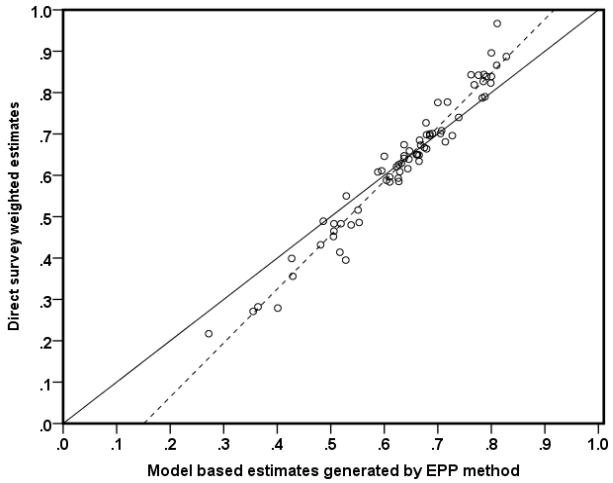

Figure 4: Bias diagnostic plot with $y = x$ line (solid) and regression line (dotted) for proportion of food insecurity for rural areas in Uttar Pradesh: EPP estimates versus direct survey estimates.

We now illustrate the second set of diagnostics to assess the extent to which the EPP estimates improve in precision compared to the direct estimates. The percent coefficient of variation (CV) is the estimated sampling standard error as a percentage of the estimate. Small area estimates with large CVs are considered unreliable. Table 2 provides a summary of CVs of the direct estimates and the EPP estimates. Figure 5 presents the District-wise values of CV for the direct and EPP methods. In one of the 71 districts, smaller CV (2.16%) of direct estimate is due to extreme value of proportion. Sample size and sample count for this district are 64 and 58 respectively while and direct estimate of proportion of food insecurity is 0.967. Note that the effective sample size and effective sample count for this districts are 25 and 24 respectively. In Table 2, we therefore presented the summary based on 70 districts (excluding one district extreme value of proportion). In further discussion we refer summary based on 70 districts only. The CVs of the direct estimates are larger than the EPP estimates.

Table 2: **Summary of area distributions of percentage coefficients of variation (CV, %) for the direct and EPP methods applied to HCES data**

Values	Summary of 71 Districts		Summary of 70 Districts	
	Direct	EPP	Direct	EPP
Minimum	2.16	5.12	5.53	5.12
Q1	8.97	7.90	9.06	7.99
Mean	14.41	10.60	14.59	10.65
Median	12.31	9.56	12.38	9.56
Q3	12.31	9.56	12.38	9.56
Maximum	45.52	24.29	45.52	24.29

Table 2 and Figure 5 show that direct estimates of incidence food insecurity are unstable with CVs that vary from 5.53 to 45.52 % with average of 14.59 %. In contrast, the CV values of EPP range from 5.12 to 24.29% with average of 10.65%. The relative performance of the EPP as compared to the direct survey estimates improve with decreasing district specific observed sample sizes. The estimates

computed from the EPP are more reliable and provide a better indication of food insecurity incidence. The district-wise plot of the 95 % confidence intervals (CIs) generated by direct and EPP methods are displayed in Figure 6, which shows that the 95% CIs for the direct estimates are wider than the 95% CIs for the EPP.

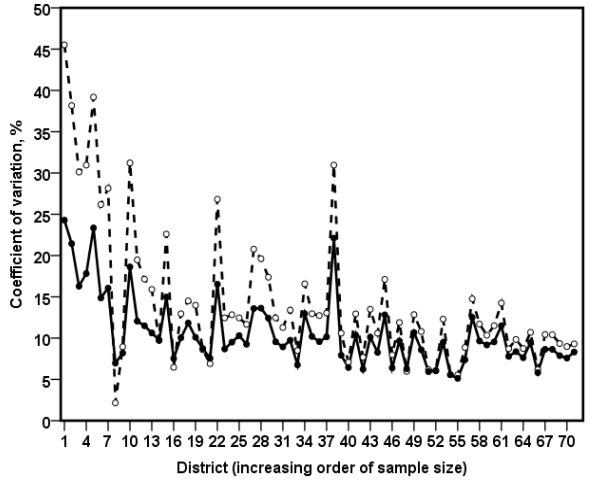

Figure 5: **District-wise percentage coefficient of variation (CV, %) for the direct (dotted line, o) and EPP (solid line, •) estimates for the food insecurity prevalence in Uttar Pradesh**

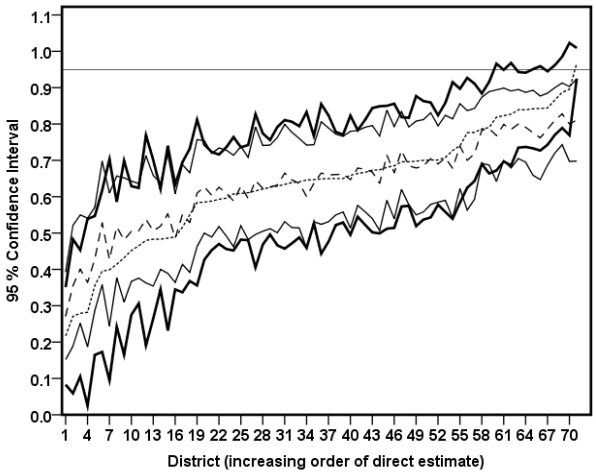

Figure 6: **District-wise 95 percentage nominal confidence interval (95% CI) for the direct (solid line) and EPP (thin line) methods. Direct (dotted point) and EPP estimates (dash point) for the food insecurity prevalence in Uttar Pradesh are shown in the 95% CI**

We inspect the aggregation property of the model-based district-level estimates generated by EPP at higher (*e.g.* State or Region) level. Let $\hat{P}_d$ and N_d denote the estimate of proportion of household food insecurity and population size for district *d.* The state-level estimate of the proportion of food insecure households is calculated as $\hat{P} = \sum_{d=1}^{D} N_d \hat{P}_d \Big/ \sum_{d=1}^{D} N_d$. The state of Uttar Pradesh is divided into Central, Eastern, Western and Southern regions, and calibration properties has been examined for these regions. State and regional level estimates of the proportion of food insecurity generated by the EPP is reported in Table 3. Comparing these with the corresponding direct estimates we see that the EPP estimates are very close to the direct estimates at

state level as well in each of the four regions. In Figure 7 we present a map showing the estimates of proportion of food insecurity in different districts in rural areas of Uttar Pradesh produced by the EPP method. This map provides the district-wise degree of inequality with respect to distribution of extent of food insecurity in rural areas of Uttar Pradesh. This map is supplemented by the results set out in Table 4, where we report the district-wise estimates along with CVs and 95 % confidence intervals generated by direct and EPP. The results indicate an east-west divide in the distribution of food insecurity. For example, in the western part of Uttar Pradesh there are many districts with low level of incidence of food insecurity. Similarly, in the eastern part and in the Bundelkhand region (north-east) we see districts with high incidence of food insecurity. This should prove useful for policy planners and administrators aiming to take effective financial and administrative decisions.

Table 3: Aggregated level estimates of incidence of food insecurity generated by direct and EPP method in different regions in Uttar Pradesh.

Estimator	State	Central	Eastern	Southern	Western
Direct	0.644	0.557	0.698	0.431	0.649
EPP	0.646	0.565	0.695	0.455	0.650

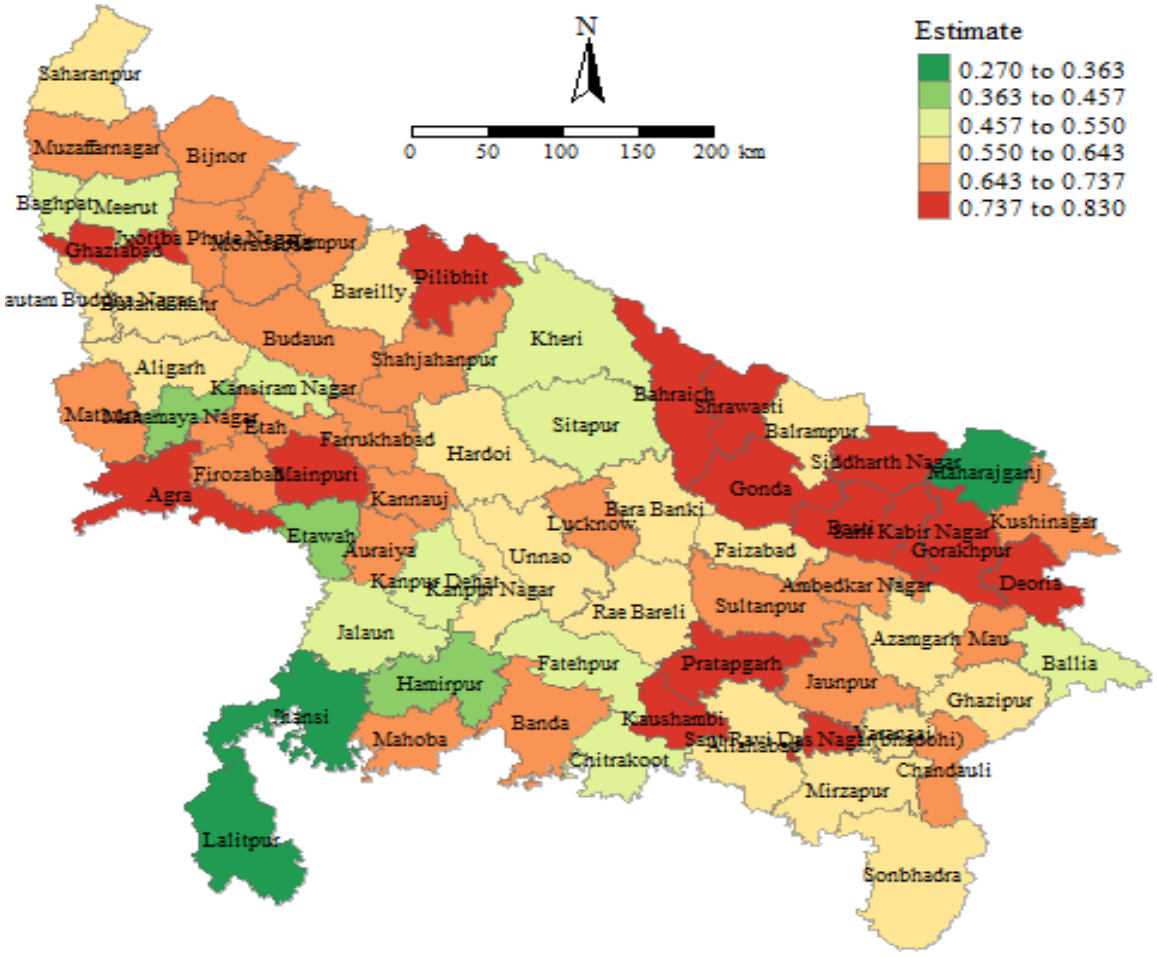

Figure 7: **EPP estimates showing the spatial distribution of incidence of food insecurity by District in Uttar Pradesh**

Table 4: Direct and EPP estimates along with 95 % confidence interval (95% CI) and percentage coefficient of variation (CV) of the incidence of food insecurity by District in rural areas of Uttar Pradesh

	Direct				EPP			
		95 % CI				95 % CI		
District	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Saharanpur	0.641	0.488	0.794	11.90	0.637	0.514	0.760	9.65
Muzaffarnagar	0.696	0.575	0.817	8.72	0.685	0.578	0.792	7.79
Bijnor	0.667	0.523	0.811	10.81	0.675	0.559	0.791	8.59
Moradabad	0.664	0.545	0.783	8.98	0.679	0.576	0.782	7.58
Rampur	0.681	0.512	0.850	12.44	0.714	0.590	0.838	8.69
Jyotiba Phule Nr	0.700	0.537	0.863	11.67	0.685	0.558	0.812	9.26
Meerut	0.452	0.275	0.629	19.62	0.505	0.367	0.643	13.63
Baghpat	0.480	0.191	0.769	30.12	0.538	0.363	0.713	16.29
Ghaziabad	0.843	0.727	0.959	6.90	0.762	0.647	0.877	7.55
Gautam B. Nr	0.486	0.232	0.740	26.17	0.553	0.389	0.717	14.87
Bulandshahr	0.611	0.480	0.742	10.70	0.595	0.482	0.708	9.46
Aligarh	0.516	0.337	0.695	17.37	0.551	0.414	0.688	12.42
Hathras	0.356	0.165	0.547	26.82	0.429	0.287	0.571	16.53
Mathura	0.685	0.514	0.856	12.45	0.666	0.539	0.793	9.56
Agra	0.844	0.743	0.945	5.97	0.786	0.687	0.885	6.28
Firozabad	0.698	0.519	0.877	12.83	0.679	0.550	0.808	9.51
Etah	0.777	0.643	0.911	8.59	0.718	0.593	0.843	8.73
Mainpuri	0.967	0.925	1.009	2.16	0.811	0.698	0.924	6.96
Budaun	0.701	0.543	0.859	11.29	0.705	0.579	0.831	8.94
Bareilly	0.585	0.427	0.743	13.50	0.627	0.500	0.754	10.12
Pilibhit	0.842	0.733	0.951	6.46	0.776	0.659	0.893	7.54
Shahjahanpur	0.673	0.502	0.844	12.72	0.668	0.540	0.796	9.56
Kheri	0.465	0.306	0.624	17.12	0.506	0.376	0.636	12.82
Sitapur	0.483	0.345	0.621	14.25	0.519	0.400	0.638	11.50
Hardoi	0.626	0.496	0.756	10.35	0.627	0.512	0.742	9.18
Unnao	0.608	0.452	0.764	12.85	0.588	0.462	0.714	10.68
Lucknow	0.639	0.472	0.806	13.04	0.646	0.515	0.777	10.16
Rae Bareli	0.647	0.527	0.767	9.30	0.637	0.531	0.743	8.32
Farrukhabad	0.649	0.443	0.855	15.89	0.665	0.524	0.806	10.63
Kannauj	0.776	0.625	0.927	9.71	0.700	0.563	0.837	9.76
Etawah	0.279	0.105	0.453	31.22	0.401	0.252	0.550	18.63
Auraiya	0.659	0.495	0.823	12.45	0.647	0.514	0.780	10.31
Kanpur Dehat	0.483	0.265	0.701	22.60	0.506	0.354	0.658	14.99
Kanpur Nagar	0.646	0.459	0.833	14.51	0.600	0.458	0.742	11.83
Jalaun	0.550	0.368	0.732	16.57	0.529	0.392	0.666	12.96
Jhansi	0.217	0.083	0.351	30.96	0.272	0.152	0.392	22.12
Lalitpur	0.271	0.059	0.483	39.19	0.355	0.189	0.521	23.35
Hamirpur	0.399	0.095	0.703	38.16	0.427	0.244	0.610	21.44
Mahoba	0.282	0.025	0.539	45.52	0.364	0.187	0.541	24.29
Banda	0.727	0.539	0.915	12.96	0.678	0.542	0.814	10.04
Chitrakoot	0.432	0.165	0.699	30.95	0.481	0.310	0.652	17.82
Fatehpur	0.489	0.345	0.633	14.76	0.486	0.364	0.608	12.52
Pratapgarh	0.887	0.789	0.985	5.53	0.828	0.743	0.913	5.12
Kaushambi	0.896	0.769	1.023	7.10	0.800	0.697	0.903	6.43

Allahabad	0.621	0.468	0.774	12.31	0.623	0.505	0.741	9.48
BaraBanki	0.674	0.499	0.849	12.95	0.637	0.507	0.767	10.20
Faizabad	0.584	0.356	0.812	19.49	0.610	0.463	0.757	12.05
Ambedkar Nr	0.701	0.578	0.824	8.74	0.690	0.585	0.795	7.63
Sultanpur	0.650	0.529	0.771	9.35	0.662	0.558	0.766	7.88
Bahraich	0.740	0.583	0.897	10.60	0.739	0.622	0.856	7.93
Shrawasti	0.819	0.672	0.966	8.96	0.768	0.642	0.894	8.19
Balrampur	0.616	0.405	0.827	17.14	0.644	0.496	0.792	11.49
Gonda	0.866	0.770	0.962	5.56	0.810	0.720	0.900	5.56
Siddharthnagar	0.839	0.735	0.943	6.17	0.800	0.705	0.895	5.93
Basti	0.839	0.737	0.941	6.05	0.791	0.695	0.887	6.05
Sant Kabir Nr	0.823	0.697	0.949	7.68	0.799	0.699	0.899	6.23
Mahrajganj	0.708	0.558	0.858	10.60	0.707	0.590	0.824	8.27
Gorakhpur	0.787	0.690	0.884	6.19	0.783	0.692	0.874	5.81
Kushinagar	0.696	0.573	0.819	8.87	0.727	0.620	0.834	7.37
Deoria	0.790	0.664	0.916	7.98	0.788	0.687	0.889	6.40
Azamgarh	0.593	0.470	0.716	10.41	0.626	0.518	0.734	8.65
Mau	0.634	0.457	0.811	13.98	0.665	0.531	0.799	10.10
Ballia	0.414	0.242	0.586	20.77	0.517	0.377	0.657	13.58
Jaunpur	0.650	0.522	0.778	9.84	0.661	0.550	0.772	8.37
Ghazipur	0.609	0.482	0.736	10.43	0.629	0.520	0.738	8.63
Chandauli	0.650	0.476	0.824	13.40	0.660	0.531	0.789	9.75
Varanasi	0.596	0.456	0.736	11.71	0.610	0.492	0.728	9.66
Bhadohi	0.827	0.686	0.968	8.50	0.785	0.679	0.891	6.76
Mirzapur	0.588	0.453	0.723	11.50	0.604	0.489	0.719	9.54
Sonbhadra	0.629	0.466	0.792	12.94	0.632	0.501	0.763	10.36
Kanshiram Nr	0.395	0.173	0.617	28.14	0.528	0.358	0.698	16.06

Nr- Nagar

5. Concluding Remarks

In this paper we outlined a plug-in empirical predictor (EPP) for small area proportions and employed for estimating the district-wise incidence of food insecurity in rural areas of the state of Uttar Pradesh using the 2011-12 HCES data collected by the NSSO of India. The auxiliary variables used in this analysis were taken from the 2011 Population Census. The effective sample sizes in place of the observed sample sizes were used to account for sampling design information of the 2011-12 HCES. The use of survey information through effective sample size leads to better representative and realistic estimates of incidence of food insecurity. The empirical results were also evaluated through several diagnostic measures and showed that the model-based SAE method defined by EPP provide significant gains in efficiency for generating district level estimates of proportion of food insecurity. Spatial map produced from the estimates generated by the EPP provides an evidence of inequality in distribution of incidence food insecurity across different districts in Uttar Pradesh. Availability of reliable district level estimates can definitely be useful for various Departments and Ministries in Government of India as well as International organizations for their policy research and strategic planning. These estimates will also be useful for budget allocation and to target welfare interventions by identifying the districts/regions with high food insecurity incidence. This application clearly demonstrates the advantage of using SAE technique to cope up the small sample size

problem in producing the cost effective and reliable disaggregate level estimates and confidence intervals from existing survey data by combining auxiliary information from different published sources with direct survey estimates.

References

Anjoy, P., Chandra, H. and Parsad, R. (2020). Estimation and spatial mapping of incidence of indebtedness in the state of Karnataka in India by combining survey and census data. *Statistics and Applications*, **18(1)**, 21-33.

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. (2013). Exploring spatial dependence in area level random effect model for disaggregate level crop yield estimation. *Journal of Applied Statistics*, **40**, 823-842.

Chandra, H., Chambers, R. and Salvati, N. (2019). Small area estimation of survey weighted counts under aggregated level spatial model. *Survey Methodology*, **45**(**1**), 31-59.

Chandra, H., Chambers, R. and Salvati, N. (2012). Small area estimation of proportions in business surveys. *Journal of Statistical Computation and Simulation*. **82**(**6**), 783-795.

Chandra, H., Salvati, N. and Chambers, R. (2018). Small area estimation under a spatially non-linear model. *Computational Statistics and Data Analysis*, **126**, 19–38.

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 Chambers, R. (2015). A spatially nonstationary Fay-Herriot model for small area estimation. *Journal of Survey Statistics and Methodology*, **3**, 109–135.

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.

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.

Korn, E. and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, **23**, 192-201.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning Inference*, **111**, 117–127.

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