Estimation of District Level Poor Households in the State of Uttar Pradesh in India by Combining NSSO Survey and Census Data

Hukum Chandra[1*], U.C. Sud[1] and Nicola Salvati[2]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*University of Pisa, Pisa, Italy*

SUMMARY

The National Sample Survey Organisation (NSSO) surveys are the main source of official statistics in India. A range of invaluable data at the macro level (e.g. state and national level) is generated through these surveys. However, the NSSO data cannot be used directly to produce reliable estimates at the micro level (e.g. district or further disaggregate level) due to small sample sizes. There is a rapidly growing demand of such micro level statistics in India as the country is moving from centralized to more decentralized planning system. In this article we employ small area estimation (SAE) techniques to derive model-based estimates of proportion of poor households at district level in the State of Uttar Pradesh in India by linking data from the Household Consumer Expenditure Survey 2006-07 of NSSO 63rd round and the Population Census. The poverty line used in this study is same as those of year 2004-05, given by Planning Commission, Govt of India. The poverty line is used to identify whether a given household is poor or not. A household having monthly per capita consumer expenditure below the state's poverty line is categorised as poor household. The results show that the model-based estimates are reliable. In contrast, the direct estimates are very unstable. These estimates are expected to provide invaluable information to policy-analysts and decision-makers.

Keywords: NSSO survey, Population census, Small area estimation, Poor households, Proportion.

1. INTRODUCTION

There is great emphasis on district level planning in India. The efforts to develop databases required for planning and decision-making at lower than the State level, were initiated quite some time back with the Planning Commission in the Government of India setting up a "Working Group on Districts Planning" in September, 1982. The Working Group in its report clearly highlighted the data requirement for planning and decision-making at the district level. However, it was found that though a lot of data are collected, processed and published for the country as a whole or for individual states, not much disaggregation of the data for sub-state level is done. The National Sample Survey Organisation (NSSO) surveys are main source

of official statistics in India. However, these surveys are planned to generate statistics at state and national level. There is no regular flow of estimates at further below level, e.g. at the district level. Therefore, NSSO surveys provide reliable state and national level estimates; they can not be used to derive reliable direct estimates at the district level owing to small sample sizes which lead to high levels of sampling variability (Pfeffermann 2002 and Rao 2003). Although in Indian context, district is a very important domain of planning process, we do not have surveys to produce estimates at this level. At the same time it is also true that conducting district specific surveys is going to be very trivial and costly as well as time consuming job. Using the state level survey data to derive the estimates at district or further smaller level

---

Corresponding author : Hukum Chandra
E-mail address : hchandra@iasri.res.in

may result in small sample sizes leading to very unstable estimates.

Due to the lack of statistics at lower level, proper planning, fund allocation and also monitoring of various plans is likely to suffer. An obvious solution to this problem is to use small area estimation (SAE) techniques. The SAE produces reliable estimates for such small areas with small sample sizes by borrowing strength from data of other areas. The SAE techniques 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. Small area models can be classified into two broad types:

(i) Area level random effect models, which are used when auxiliary information is available only at area level. They relate small area direct estimates to area-specific covariates (Fay and Herriot 1979) and

(ii) Nested error unit level regression models, proposed originally by Battese *et al*. (1988). These models relate the unit values of a study variable to unit-specific covariates.

We adopt the area level model since covariates are available only at the area level. In this paper we use small area estimation techniques to derive model-based estimates of proportion of poor households at small area levels in the State of Uttar Pradesh in India by linking data from the Household Consumer Expenditure (HCE) Survey 2006-07 of NSSO 63rd round and the Population Census 2001. Small areas are defined as the different districts of State of Uttar Pradesh in India. This article, in particular, illustrates how the HCE survey and Census data can be combined to derive reliable small area estimates for various policy relevant parameters. The rest of the paper is organised as follows. In Section 2 we describe the data used for the analysis and in Section 3 we present an overview of the methodology used for analysis. Section 4 discusses the diagnostic procedures for examining the model assumptions, validating the small area estimates and describes the results. Section 5 finally set out the main conclusions.

2. DATA

In this study covariates are available at districts level. Therefore, we adopt an area level area model to derive the small area level estimates (Fay and Herriot 1979). Two types of variables are required for this analysis.

(i) The variable of interest for which small area estimates are required is drawn from the Household Consumer Expenditure Survey 2006-07 of NSSO 63rd round data for rural areas of the State of Uttar Pradesh. The target variable used for the study was poor households. The poverty line has been used to identify whether given household is poor or not. A household having monthly per capita consumer expenditure below the state's poverty line (i.e. Rs. 365.84) is categorised as poor household. The poverty line used in this study is same as those of year 2004-05, given by Planning Commission, Govt of India. The parameter of interest is the proportion of poor household at the district level.

(ii) The auxiliary (covariates) variables are drawn from the Population Census 2001.

It is noteworthy that use of covariates from the Population Census 2001 to model poor household from the HCE Survey 2006-07 of NSSO may raise issues of comparability. However, relationship between variable of interest and covariates used in this study are assumed not to change significantly over the period. There were more than 100 covariates available from the Population Census for the purpose of modelling. Out of these, suitable covariates were selected for the analysis as follows: We first examined the correlation of all these covariates with the target variable and then selected the covariates with reasonably good correlation with the target variable. This was followed by step-wise regression analysis. Finally, six variables namely (i) sex ratio of SC population, (ii) sex ratio of ST population, (iii) percentage of other worker population, (iv) percentage of literate male, (v) main other workers female and (vi) marginal other population were identified for the further analysis which significantly explained the model. The R^2 for the chosen model was 48 per cent.

The sampling design used in the NSSO data is stratified multi-stage random sampling with districts as strata, villages as first stage units and households as the second stage units. A total of 2322 households were surveyed from the 70 districts of the Uttar Pradesh. The district-wise sample size varied from 19 to 48 with average of 33 (Table 1). Our aim is to estimate proportion of poor households at district level. It is evident that district level sample sizes are very small with very low values of average sampling fraction as 0.0001. Therefore, it is difficult to derive reliable estimates and their standard errors at district level. The SAE is an obvious choice for such cases. The SAE technique is expected to provide reliable estimates for the districts having small sample data (Rao 2003). The theory of SAE has been illustrated in next Section.

3. METHODOLOGY

This Section illustrates theoretical framework used to produce small area estimates of proportion and their measure of precision. The theory presented has successfully also been applied in many other small area estimation applications, see Johnson *et al.* (2010). To start with, we fix our notations. Throughout, we use a subscript d to index the quantities belonging to small area d ($d = 1, ..., D$), where D is the number of small areas (or areas) in the population. The subscript s and r are used for denoting the quantities related to the sample and non-sample parts of the population. So that n_d and N_d represent the sample and population (i.e., number of households in sample and population) sizes in district d, respectively. The direct estimator of proportion of poor household is given by $\hat{p}_d^{Direct} = \left(\sum_{i \in s_d} w_{di}\right)^{-1} \left(\sum_{i \in s_d} w_{di} y_{di}\right)$, where w_{di} is the survey weight associated with household i in area d. Assuming that joint inclusion $1/w_{di,d'j} = 0$ for $d \neq d'$ or $i \neq j$, the estimate of mean squared error of direct estimator $\hat{p}_d^{Direct}$ is

$$v(\hat{p}_d^{Direct}) = \left(\sum_{i \in s_d} w_{di}\right)^{-2} \left\{\sum_{i \in s_d} w_{di}(w_{di} - 1)(y_{di} - \hat{p}_d^{Direct})^2\right\},$$

Table 1. Distribution of districts-wise sample size

District	Sample size	District	Sample size	District	Sample size	District	Sample size
1	48	19	48	37	24	55	24
2	48	20	48	38	24	56	24
3	48	21	24	39	24	57	24
4	48	22	24	40	24	58	48
5	24	23	48	41	24	59	48
6	24	24	48	42	24	60	48
7	24	25	48	43	48	61	48
8	24	26	48	44	24	62	24
9	24	27	24	45	48	63	48
10	24	28	48	46	48	64	48
11	48	29	24	47	24	65	48
12	48	30	24	48	24	66	19
13	24	31	24	49	48	67	24
14	24	32	24	50	48	68	24
15	24	33	24	51	23	69	24
16	24	34	24	52	24	70	24
17	48	35	24	53	48		
18	24	36	24	54	24	**Total**	**2322**

see for example Särndal *et al.* (1992, page 391). The value of variable of interest y (the poor households) in the area d is defined by y_d and we denote by y_{sd} and y_{rd} the sample and non-sample counts of poor households in area d. Indeed, the variable of interest y_{sd} has a Binomial distribution with parameters n_d and π_d, denoted by $y_{sd} \sim Bin(n_d, \pi_d)$, where π_d is the probability of a poor household in area d, often termed as the probability of a 'success'. Similarly, $y_{rd} \sim Bin(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent Binomial variables with π_d being a common success probability. As mentioned in previous Section in model-based small area estimation the survey data is supplemented by the availability of auxiliary information from various sources, e.g. Census and Administrative records. Let $\mathbf{x}_d$ be the k-vector of the covariates for area d from the previous sources. The model linking the probabilities of success π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model given by

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

$$(d = 1, ..., D), \qquad (1)$$

where $\boldsymbol{\beta}$ is the k-vector of regression coefficient often known as fixed effect parameters and u_d is the area-specific random effect that accounts for between area dissimilarity beyond that explained by the auxiliary variables included in the in the fixed part of the model. We assume that u_d's are independent and normally distributed with mean zero and variance ϕ. Under model (1), we get

$$\pi_d = \exp(\eta_d)\{1 + \exp(\eta_d)\}^{-1}$$

$$= \exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d)\{1 + \exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d)\}^{-1}.$$

It is evident that model (1) relates the area level proportions to area level covariates. This type of model is often referred to as 'area-level' model in SAE terminology, see for example Rao (2003). Such a model was originally used by Fay and Herriot (1979) for the prediction of mean per-capita income (PCI) in small geographical areas (less than 500 persons) within counties in the United States. The Fay and Herriot (FH) method for SAE is based on area level linear mixed model and their approach is applicable to a continuous variable. In contrast, model (1) is a special case of a generalized linear mixed model (GLMM) with logit link

function (Breslow and Clayton 1993) and suitable for discrete, particularly binary variable. It is noteworthy that the FH model is not applicable in such cases. Saei and Chambers (2003) and Manteiga *et al.* (2007) described this model in the context of SAE. By definition, the means of y_{sd} and y_{rd} given u_d under model (1) are:

$$E(y_{sd}|u_d) = n_d \pi_d$$

$$= n_d \left[\exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d)(1 + \exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d))^{-1} \right] \quad (2)$$

$$E(y_{rd}|u_d) = (N_d - n_d) \pi_d$$

$$= (N_d - n_d) \left[\exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d)(1 + \exp(\mathbf{x}'_d \boldsymbol{\beta} + u_d))^{-1} \right]. \quad (3)$$

Let T_d denotes the total number of poor households in district d. We can write $T_d = y_{sd} + y_{rd}$, where the first term y_{sd}, the sample count is known whereas the second term y_{rd}, the non-sample count, is unknown. Therefore, an estimate $\hat{T}_d$ of the total number of poor households in area d is obtained by replacing y_{rd} by its predicted value under the model (1). That is,

$$\hat{T}_d = y_{sd} + \hat{y}_{rd} = y_{sd} + (N_d - n_d)$$

$$\left[\exp(\mathbf{x}'_d \hat{\boldsymbol{\beta}} + \hat{u}_d)(1 + \exp(\mathbf{x}'_d \hat{\boldsymbol{\beta}} + \hat{u}_d))^{-1} \right]. \quad (4)$$

An estimate of proportion of poor households p_d in a small area d is obtained as

$$\hat{p}_d = \frac{\hat{T}_d}{N_d} = \frac{1}{N_d} \left\{ y_{sd} + (N_d - n_d) \right.$$

$$\left. \left[\exp(\mathbf{x}'_d \hat{\boldsymbol{\beta}} + \hat{u}_d)(1 + \exp(\mathbf{x}'_d \hat{\boldsymbol{\beta}} + \hat{u}_d))^{-1} \right] \right\}. \quad (5)$$

It is obvious that in order to compute the estimates given by equation (4) or (5), we require estimates of the unknown parameters $\boldsymbol{\beta}$ and $\mathbf{u}$. A major difficulty in use of logistic linear mixed model (LLMM) for SAE is the estimation of unknown model parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ since the likelihood function for LLMM often involves high dimensional integrals (computed by integrating a product of discrete and normal densities, which has no analytical solution) which are difficult to evaluate numerically. We used an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\beta}$ and $\mathbf{u} = (u_1, ..., u_D)$ with restricted maximum likelihood (REML) estimation of ϕ to estimate these unknown parameters. Detailed

description of the approach can be followed from Saei and Chambers (2003) and Manteiga *et al.* (2007).

We now turn to estimation of mean squared error (MSE) for predictors given by equation (5). The MSE estimates are computed to assess the reliability of estimates and also to construct the confidence interval (CI) for the estimates. The mean squared error estimate of (5) under model (1) is (Saei and Chambers 2003 and Manteiga *et al.* 2007) given by

$$mse(\hat{p}_d) = m_1(\hat{\phi}) + m_2(\hat{\phi}) + 2m_3(\hat{\phi}). \quad (6)$$

The first two components m_1 and m_2 constitute the largest part of the overall MSE estimates in (6). These are the MSE of the best linear unbiased predictor (BLUP)-type estimator when ϕ is known (Rao 2003). The third component m_3 is the variability due to the estimate of ϕ. For simplicity, we used few notations to write the analytical expression of various components of the MSE (6). We denote by $\hat{\mathbf{V}}_{sd} = diag\{n_d \hat{p}_d(1-\hat{p}_d)\}$ and $\hat{\mathbf{V}}_{rd} = diag\{(N_d - n_d) \hat{p}_d(1-\hat{p}_d)\}$, the diagonal matrices defined by the corresponding variances of the sample and non-sample part respectively. Similarly,

$$\mathbf{A} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}$$

$$\mathbf{B} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}\mathbf{X}_r - \mathbf{A}\hat{\mathbf{T}}_s\hat{\mathbf{V}}_{sd}\mathbf{X}_s, \text{ and}$$

$$\hat{\mathbf{T}}_s = \left(\phi^{-1}\mathbf{I}_D + \hat{\mathbf{V}}_{sd}\right)^{-1},$$

where $\mathbf{X}_s$ and $\mathbf{X}_r$ are the sample and non-sample part of covariates and $\mathbf{I}_D$ is an identity matrix of order D. We further write

$$\hat{\mathbf{T}}_{(1)} = \left\{\mathbf{X}_s'\hat{\mathbf{V}}_{sd}\mathbf{X}_s - \mathbf{X}_s'\hat{\mathbf{V}}_{sd}\hat{\mathbf{T}}_s\hat{\mathbf{V}}_{sd}\mathbf{X}_s\right\}^{-1} \text{ and}$$

$$\hat{\mathbf{T}}_{(2)} = \hat{\mathbf{T}}_s + \hat{\mathbf{T}}_s\hat{\mathbf{V}}_{sd}\mathbf{X}_s\hat{\mathbf{T}}_{(1)}\mathbf{X}_s'\hat{\mathbf{V}}_{sd}'\hat{\mathbf{T}}_s.$$

With these notations, assuming model (1) holds, the various components of equation (6) are

$$m_1(\hat{\phi}) = \mathbf{A}\hat{\mathbf{T}}_s\mathbf{A}', \quad m_2(\hat{\phi}) = \mathbf{B}\hat{\mathbf{T}}_{(1)}\mathbf{B}', \text{ and}$$

$$m_3(\hat{\phi}) = trace\left(\hat{\mathbf{V}}_i\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_j'v(\hat{\phi})\right) \text{ with}$$

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}'.$$

Here $v(\hat{\phi})$ is the asymptotic covariance matrix of the estimates of variance components $\hat{\phi}$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\phi}$. Note that this also depends

upon whether we are using maximum likelihood (ML) or restricted maximum likelihood (REML) estimates for $\hat{\phi}$. We used REML estimates for $\hat{\phi}$, then $v(\hat{\phi}) = 2\left(\hat{\phi}^{-2}(D-2t_1) + \hat{\phi}^{-4}t_{11}\right)^{-1}$ with $t_1 = \hat{\phi}^{-1}trace(\hat{\mathbf{T}}_{(2)})$ and $t_{11} = trace(\hat{\mathbf{T}}_{(2)}\hat{\mathbf{T}}_{(2)})$. Let us write $\Delta = \mathbf{A}\hat{\mathbf{T}}_s$ and $\hat{\mathbf{V}}_i = \partial(\Delta_i)/\partial\phi|_{\phi=\hat{\phi}} = \partial(\mathbf{A}_i\hat{\mathbf{T}}_s)/\partial\phi|_{\phi=\hat{\phi}}$, where $\mathbf{A}_i$ is the ith row of the matrix $\mathbf{A}$. The numerical results reported in Sections 4 are obtained using R version 2.9.2.

4. RESULTS AND DISCUSSIONS

4.1 Diagnostic Procedures

Generally two types of diagnostics procedures are tested in small area estimation, the model diagnostics and the diagnostics for the small area estimates, see for example Ambler *et al.* (2001). The first diagnostics are used to verify the assumptions of underlying model and the second diagnostics are applied to validate the reliability of the model-based small area estimates. The random area effects $u_d (d = 1, ..., D)$ in model (1) are assumed to have a normal distribution with mean zero and variance ϕ. If the model assumptions are satisfied then the district level residuals are expected to be randomly distributed and not significantly different from the regression line $y = 0$, where under model (1), the area level residuals are defined as $r_d = \hat{\eta}_d - \mathbf{x}_d'\hat{\boldsymbol{\beta}}$. The distribution of the district level residuals (left side plots) and q-q plots (right side plots) are shown in Fig. 1. The Fig. 1 clearly reveals that the randomly distributed district level residuals and the line of fit does not significantly differ from the line $y = 0$ as expected in all the plots. The q-q plots also confirm the normality assumption. Therefore the model diagnostics are fully satisfied for the data.

To validate the reliability of the model-based small area estimates we used the bias diagnostics, coefficient of variation (CV) and computed the 95 per cent confidence intervals. The bias diagnostics are used to investigate if the model-based estimates are less extreme as compared to the direct survey estimates, when they are available, see Johnson *et al.* (2010). The bias scatter plot of the model-based estimates against the direct estimates is set out in Fig. 2. The plot show that the model-based estimates are less extreme as
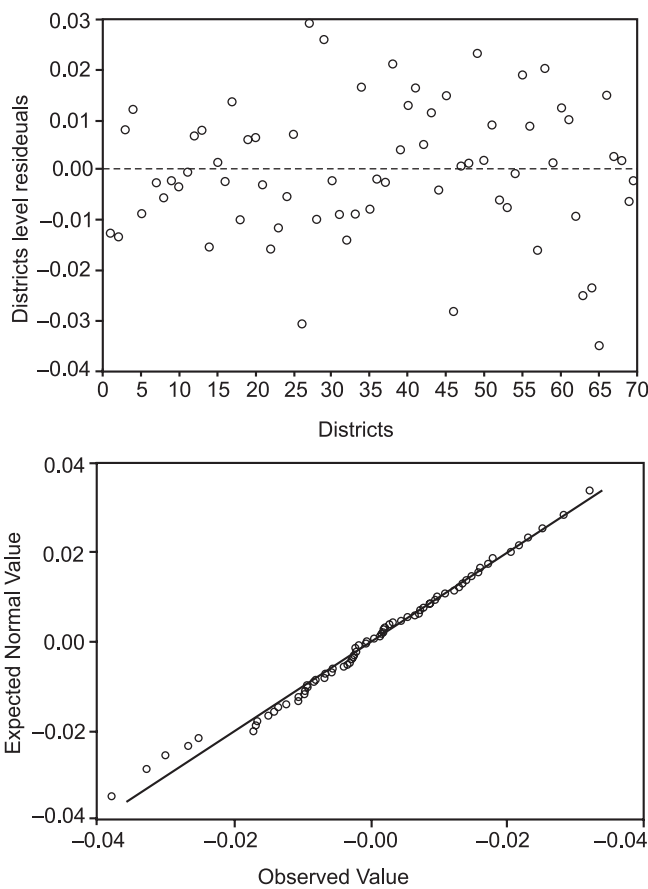
Fig. 1. Distribution of the district level residuals (upper plot) and normal q-q plot of the district level residuals (lower plot)
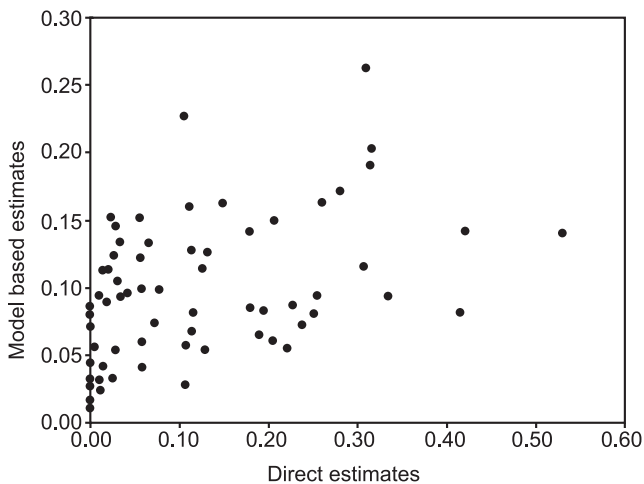

Fig. 2. Bias diagnostics plot

compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average.

We computed the CV to assess the improved precision of the model-based estimates compared to the direct estimates. The CVs show the sampling variability as a percentage of the estimate. Estimates with large CVs are considered unreliable (i.e. smaller is better). There are no internationally accepted tables available that allow us to judge how large is 'too large' (Ambler *et al.* 2001 and Johnson *et al.* 2010). Fig. 3 presents the district-wise distribution of the percentage CV of model based estimates and direct estimates. The estimated CVs show that model-based estimates have a higher degree of reliability as compared to the direct estimates.
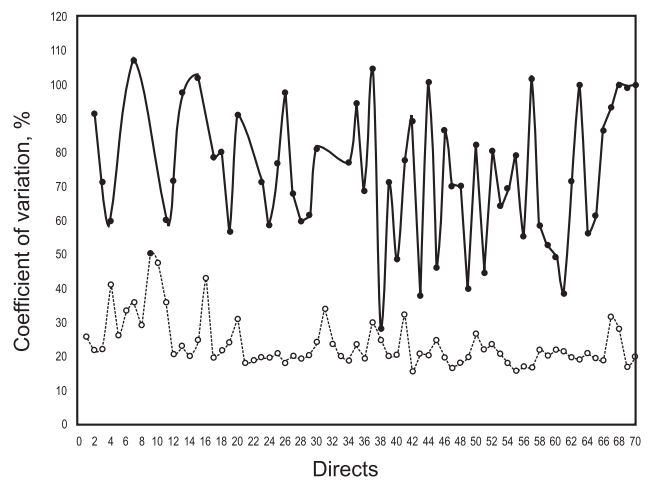

Fig. 3. District-wise coefficient of variation for direct (solid line) and model-based estimate (dash line)

In Table 2 we present the districts-wise 95 per cent confidence intervals of the model-based and the direct estimates. The standard errors of the direct estimates are too large and therefore the estimates are unreliable. Note that for many districts we can even not produce the confidence intervals due to unavailability of standard errors.

4.2 DISCUSSIONS

The small area estimates diagnostic measures clearly depict that the model-based estimates are reliable and more stable than the corresponding direct estimates (Fig. 3). Table 2 presents the direct estimates and model-based estimates along with 95 per cent confidence intervals for the State of Uttar Pradesh. These results show the degree of inequality with respect to distribution of poor households in different districts.

Table 2. District-wise model-based and direct estimates of proportion of poor households

District	Direct			Model-based			District	Direct			Model-based		
	Estimate	Lower	Upper	Estimate	Lower	Upper		Estimate	Lower	Upper	Estimate	Lower	Upper
1	0.00	0.000	0.000	0.03	0.016	0.049	36	0.07	−0.024	0.154	0.13	0.081	0.185
2	0.13	−0.106	0.362	0.05	0.031	0.079	37	0.03	−0.030	0.084	0.05	0.022	0.087
3	0.01	−0.006	0.034	0.04	0.023	0.060	38	0.53	0.232	0.826	0.14	0.071	0.211
4	0.33	−0.066	0.734	0.09	0.016	0.171	39	0.03	−0.012	0.068	0.15	0.087	0.205
5	0.00	0.000	0.000	0.04	0.021	0.067	40	0.42	0.011	0.829	0.14	0.083	0.200
6	0.00	0.000	0.000	0.03	0.009	0.044	41	0.19	−0.108	0.496	0.08	0.029	0.138
7	0.03	−0.032	0.088	0.05	0.015	0.095	42	0.08	−0.060	0.214	0.10	0.068	0.130
8	0.00	0.000	0.000	0.03	0.011	0.044	43	0.31	0.073	0.545	0.26	0.153	0.372
9	0.00	0.000	0.000	0.01	0.000	0.022	44	0.06	−0.058	0.172	0.06	0.036	0.085
10	0.00	0.000	0.000	0.02	0.001	0.033	45	0.23	0.017	0.437	0.09	0.044	0.130
11	0.13	−0.026	0.288	0.13	0.035	0.217	46	0.02	−0.015	0.055	0.11	0.069	0.159
12	0.19	−0.082	0.460	0.07	0.038	0.093	47	0.15	−0.060	0.356	0.16	0.108	0.218
13	0.06	−0.055	0.171	0.04	0.022	0.061	48	0.11	−0.045	0.267	0.16	0.102	0.218
14	0.00	0.000	0.000	0.08	0.047	0.111	49	0.31	0.062	0.566	0.19	0.115	0.267
15	0.02	−0.025	0.073	0.03	0.017	0.050	50	0.11	−0.070	0.284	0.06	0.027	0.088
16	0.00	0.000	0.000	0.01	0.001	0.019	51	0.41	0.042	0.786	0.08	0.045	0.119
17	0.02	−0.010	0.046	0.09	0.054	0.126	52	0.24	−0.145	0.619	0.07	0.038	0.107
18	0.03	−0.020	0.086	0.13	0.075	0.193	53	0.12	−0.033	0.263	0.08	0.047	0.116
19	0.11	−0.015	0.241	0.07	0.035	0.101	54	0.11	−0.044	0.270	0.13	0.082	0.174
20	0.01	−0.009	0.031	0.02	0.009	0.040	55	0.06	−0.032	0.142	0.15	0.104	0.200
21	0.00	0.004	0.004	0.06	0.036	0.077	56	0.26	−0.028	0.548	0.16	0.107	0.219
22	0.00	0.000	0.000	0.08	0.050	0.111	57	0.03	−0.027	0.079	0.12	0.082	0.166
23	0.01	−0.006	0.034	0.11	0.068	0.158	58	0.03	−0.006	0.074	0.09	0.053	0.135
24	0.04	−0.007	0.091	0.10	0.058	0.135	59	0.18	−0.011	0.367	0.14	0.084	0.200
25	0.18	−0.096	0.454	0.09	0.050	0.122	60	0.13	0.002	0.248	0.11	0.064	0.165
26	0.06	−0.054	0.168	0.10	0.063	0.136	61	0.32	0.073	0.557	0.20	0.116	0.291
27	0.22	−0.079	0.519	0.06	0.033	0.078	62	0.28	−0.121	0.681	0.17	0.104	0.240
28	0.21	−0.041	0.453	0.15	0.092	0.208	63	0.03	−0.030	0.090	0.11	0.065	0.146
29	0.31	−0.072	0.684	0.12	0.069	0.163	64	0.11	−0.013	0.223	0.23	0.132	0.322
30	0.25	−0.159	0.667	0.09	0.048	0.140	65	0.02	−0.005	0.051	0.15	0.092	0.212
31	0.00	0.000	0.000	0.05	0.014	0.077	66	0.2	−0.149	0.557	0.06	0.038	0.084
32	0.00	0.000	0.000	0.07	0.037	0.105	67	0.11	−0.092	0.304	0.03	0.010	0.046
33	0.00	0.000	0.000	0.09	0.051	0.121	68	0.01	−0.010	0.030	0.03	0.014	0.050
34	0.25	−0.136	0.636	0.08	0.050	0.112	69	0.07	−0.071	0.215	0.07	0.049	0.099
35	0.06	−0.050	0.162	0.12	0.064	0.181	70	0.01	−0.010	0.030	0.09	0.056	0.132

A critical review of Table 2 shows that in many districts the lower bound (Lower) of 95% confidence interval (CI) is negative which results in practically impossible and inadmissible values of CI for direct estimates. In contrast, the model estimate with precise CI and reasonable CV percent are reliable. This problem was mostly observed when there was no variability in the sample data of district. For example where all *y* values in sample were 0 estimated direct proportions was 0. In such circumstance, SAE plays an

important role in generating micro level statistics. The results clearly show the advantage of using SAE technique to cope up the small sample size problem in producing the estimates or reliable confidence intervals. These estimates can definitely be useful for resource allocation and policy decision-making relating the living condition of people in rural areas.

5. CONCLUSIONS

The method of estimation of proportions for small areas is well developed (Saei and Chambers 2003 and Manteiga *et al.* 2007), however, there is limited application in the area of agricultural or social sciences. Further, there is very less known application to the Indian data, particularly, NSSO data. In this article we demonstrate the application of SAE techniques to estimate the district level statistics of poor households using survey and census data. The diagnostic procedures clearly confirm that the model-based district level estimates have reasonably good precision. As the quantum of work involved in the conduct of Census is quite appreciable, Censuses are generally carried out after a fixed period of time. Thus, the Census data is available only after a certain time period. The NSSO survey, on the other hand, contributes to providing estimates on a regular basis at the State and National level. They do not provide sub-state level statistics. However, it is known that regional and national estimates usually mask variations (heterogeneity) at the sub-state or district level and render little information for micro level planning and allocation of resources.

A lot of emphasis now a days is being given to micro level planning all over the world. In the Indian context, district is an important domain for planning process in the country and therefore availability of district level statistics is vital to monitoring of policy and planning. This study produces reliable statistics at micro level using existing surveys and other already available secondary data and can be seen as an initiative example for further applications. Such micro level statistics can be generated without conducting micro level specific survey. This has the merit that unlike Census, estimates can be produced on a regular basis from existing surveys.

REFERENCES

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. and Wang, S. (2001). Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method. *Proceedings of the International Association of Survey Statisticians,* Meeting of the ISI, Seoul, August 2001.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.,* **83**, 28-36.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9-25.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **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. *J. Off. Stat.*, **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. *Comput. Statist. Data Anal.*, **51**, 2720-2733.

Pfeffermann, D. (2002). Small area estimation: new developments and directions. *Internat. Statist. Rev.*, **70**, 125-143.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology, John Wiley and Sons Inc.

Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *Southampton Statistical Sciences Research Institute, University of Southampton, UK, W.P. No. M03/15.*

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