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

Hukum Chandra, U.C. Sud and Nicola Salvati

ucsud@iasri.res.in

Indian Agricultural Statistics Research Institute

New Delhi-110012

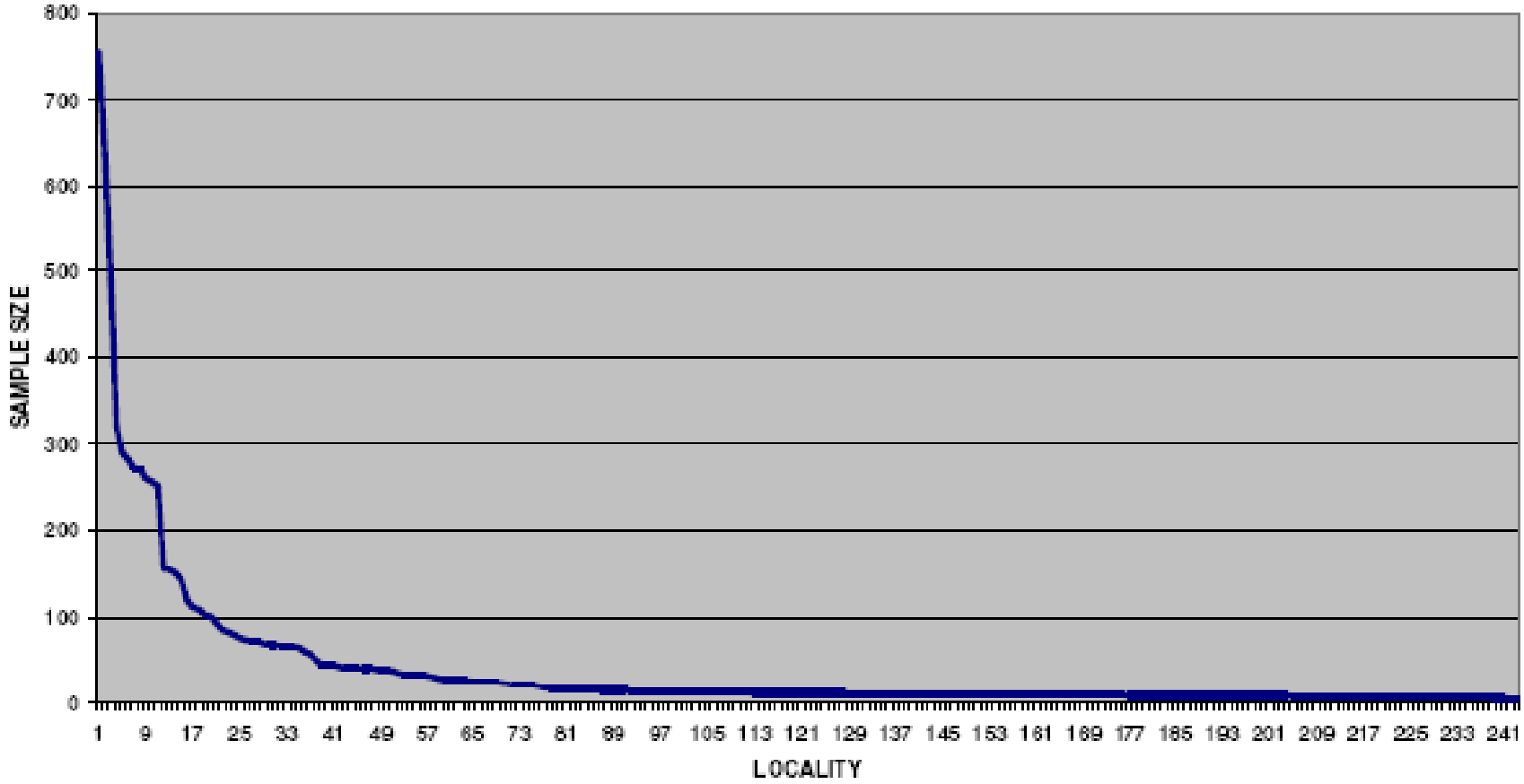
Introduction-Small Area Estimation (SAE) Problem

- Sample surveys are generally large scale
- Estimates are readily available at macro level
- There is emphasis on micro level planning
- Reliable and timely estimates are needed at the micro level
- Micro level estimates generated out of large scale surveys may have very high standard errors (due to small sample size)-unreliable
- Micro level specific sample surveys are not possible

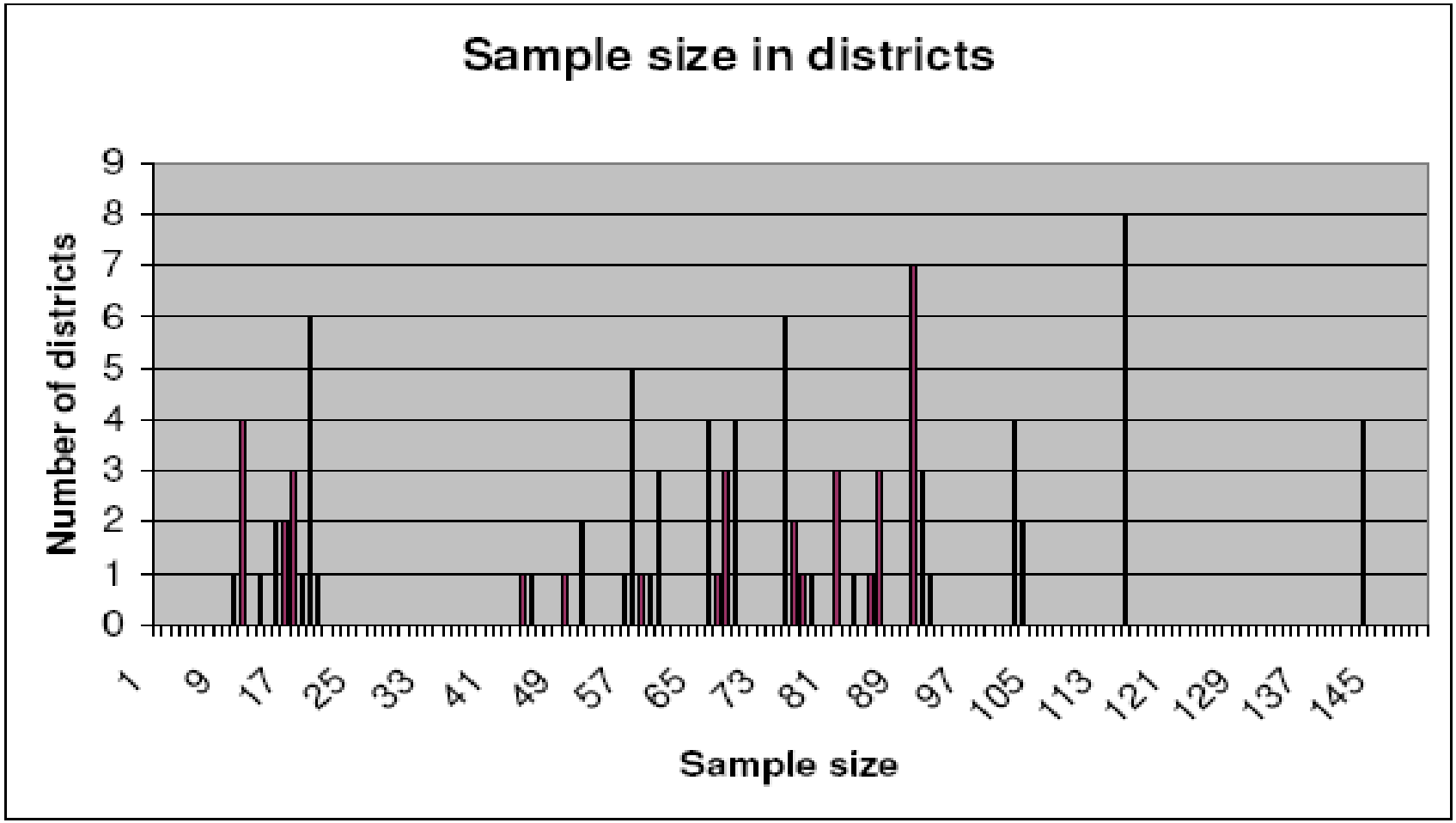
  i) High cost

  Ii) Massive infra-structure, time factor

SAMPLE SIZES IN ISRAEL LABOR FORCE SURVEY BY CITIES
FIRST QUARTER, 2001



Illiteracy survey in Cambodia, 1999



CV of the direct estimator (amount of loan outstanding) District level estimates (59th round)

25.56451
23.55819
39.52303
30.88995
17.97997
18.26402
19.75504
29.19395
26.34203
28.66078

25.56451
23.55819
39.52303
30.88995
17.97997
18.26402
19.75504
29.19395
26.34203
28.66078

Solution to the problem of SAE

- Mechanism is needed whereby reliable micro level estimates can be generated out of the large scale surveys
- SAE techniques can be used to generate reliable micro-level estimates
- The SAE techniques have the potential to produce 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

Two-types of small area models are available:

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 [4]) and

II. Nested error unit level regression models, proposed originally by Battese, Harter and Fuller [2]. 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.

National Sample Surveys Office

- A district, in India, is a very important domain of planning process, we do not have surveys to produce estimates at this level.
- The National Sample Survey Office (NSSO) surveys are one of the main sources of official statistics in India.
- The NSSO primarily carries out socio-economic surveys
- The NSSO was initiated in 1950 to conduct sampling enquiries with a view to providing the Government and other organizations with socio-economic data, which can be used for planning for national development and for various research purposes.

Contd.

- The NSSO surveys are planned to generate statistics at state and national level
- Cannot be used to derive reliable direct estimates at the district level owing to small sample sizes- estimates may have high level of sampling variability (see [7] and [8]).

Problem attempted in the paper

- SAE technique is used to derive model-based estimates of proportion of poor households at small area levels in the State of Uttar Pradesh, India
- Data linked 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.

Data Used

- Two types of variables are required for analysis:
- The target variable used for the study was poor households
- The poverty line has been used to identify whether a given household is poor or not.
- A household having monthly per capita consumer expenditure below the state poverty line (i.e., Rs. 365.84) is categorised as poor household.
- The poverty line used in this study is as that of year 2004-05, given by planning commission, Govt of India.
- The auxiliary (covariates) variables are drawn from the Population Census 2001.
- More than 100 covariates were available from the Population Census
- Suitable covariates were selected for the analysis using step-wise regression analysis.

Covariates used

Six variables were identified

(i) Sex ratio of Schedule Caste population,

(ii) Sex ratio of Schedule Tribe population,

(iii) Percentage of Other worker Population,

(iv) Percentage of Literate Male,

(v) Main Other workers female and

(vi) Marginal Other population

The R^2 for the chosen model was 48 per cent

Sampling Design of Survey

- The sampling design used in the NSSO data is stratified two-stage random sampling with districts as strata, villages as first stage units and households as the second stage units.
- Villages are selected by PPSWR (population size is the auxiliary variable)
- Households are selected by SRSWOR

Sample Size

- 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).
- It is evident that district level sample sizes are very small with very low values of average sampling fraction as 0.0001.

Table 1. Distribution of districts-wise sample size

S.No.	Sample size	S. No.	Sample size
1	48	36	24
2	48	37	24
3	48	38	24
4	48	39	24
5	24	40	24
6	24	41	24
7	24	42	24
8	24	43	48
9	24	44	24
10	24	45	48
11	48	46	48
12	48	47	24
13	24	48	24
14	24	49	48
15	24	50	48

Contd.

16	24	51	23
17	48	52	24
18	24	53	48
19	48	54	24
20	48	55	24
21	24	56	24
22	24	57	24
23	48	58	48
24	48	59	48
25	48	60	48
26	48	61	48
27	24	62	24
28	48	63	48
29	24	64	48
30	24	65	48
31	24	66	19
32	24	67	24
33	24	68	24
34	24	69	24
35	24	70	24
		Total	2322

Methodology

Direct Estimator of proportion of poor household

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

Variance of the direct estimator

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

Contd.

- y_d = the value of variable of interest *y* (the poor households) in the area *d*.
- y_{sd} = the sample count of poor households in area d
- y_{rd} = the non-sample count of poor households in area d

$$y_{sd} \sim Bin(n_d, \pi_d)$$

$$y_{rd} \sim Bin(N_d - n_d, \pi_d)$$

- Further, y_{sd} and y_{rd} are assumed to be independent Binomial variables with π_d being a common success probability .
- $\mathbf{x}_d$ = *k*-vector of the covariates for area *d* obtained from a secondary sources.

Contd.

- The model linking the probabilities of success π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model given by

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

- u_d's are independent and normally distributed with mean zero and variance φ

Model relating the area level proportions to area level covariates

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

Contd.

Area level model

- Used by Fay and Herriot [4] for the prediction of mean per-capita income (PCI) in small geographical areas (less than 500 persons) within counties in the United States
- FH model is not applicable

Contd.

- **Appropriate model is given by Saei and Chambers [9] and Manteiga et al. [6]. The means of y_{sd} and y_{rd} given u_d under above model**

$$E(y_{sd}|u_d) = n_d\pi_d = n_d\left[\exp(\mathbf{x}'_d\boldsymbol{\beta}+u_d)\left(1+\exp(\mathbf{x}'_d\boldsymbol{\beta}+u_d)\right)^{-1}\right]$$

$$E(y_{rd}|u_d) = (N_d-n_d)\pi_d = (N_d-n_d)\left[\exp(\mathbf{x}'_d\boldsymbol{\beta}+u_d)\left(1+\exp(\mathbf{x}'_d\boldsymbol{\beta}+u_d)\right)^{-1}\right]$$

- **Let**
- T_d **= the total number of poor households in district *d*.**

$$T_d = y_{sd} + y_{rd}$$

Contd.

$$\hat{T}_d = y_{sd} + \hat{y}_{rd} = y_{sd} + (N_d - n_d)\left[\exp(\mathbf{x}'_d\hat{\boldsymbol{\beta}} + \hat{u}_d)\left(1+\exp(\mathbf{x}'_d\hat{\boldsymbol{\beta}} + \hat{u}_d)\right)^{-1}\right]$$

- An estimate of proportion of poor households $\hat{p}_d$ in a small area *d* is obtained as

$$\hat{p}_d = \frac{\hat{T}_d}{N_d} = \frac{1}{N_d}\left\{y_{sd} + (N_d - n_d)\left[\exp(\mathbf{x}'_d\hat{\boldsymbol{\beta}} + \hat{u}_d)\left(1+\exp(\mathbf{x}'_d\hat{\boldsymbol{\beta}} + \hat{u}_d)\right)^{-1}\right]\right\}$$

A major difficulty in use of logistic linear mixed model (LLMM) for SAE is the estimation of unknown model parameters $\boldsymbol{\beta}$ $\mathbf{u}$

Contd.

- 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.
- An iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of β and $\mathbf{u} = (u_1, \ldots, u_D)$
- with restricted maximum likelihood (REML) estimation of ϕ to estimate these unknown parameters.
- Detailed description of the approach can be followed from [6, 9].

Mean Squared Error of Predictors

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

- The first two components m_1 and m_2 constitute the largest part of the overall MSE
- The third component m_3 is the variability due to the estimation of ϕ

Contd.

$$\hat{\mathbf{V}}_{sd} = diag\left\{n_d \hat{p}_d (1-\hat{p}_d)\right\}$$

$$\hat{\mathbf{V}}_{rd} = diag\left\{(N_d - n_d)\hat{p}_d (1-\hat{p}_d)\right\}$$

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

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

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

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

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

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

$$m_3(\hat{\phi}) = trace\left(\hat{\nabla}_i\hat{\boldsymbol{\Sigma}}\hat{\nabla}'_j v(\hat{\phi})\right) \quad , \quad \hat{\boldsymbol{\Sigma}} = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}'_{sd}$$

Contd.

- Here $v(\hat{\phi})$ is the asymptotic covariance matrix of the estimates of variance components $\hat{\varphi}$
- REML estimates were used for $\hat{\phi}$,

$$v(\hat{\phi}) = 2\left(\hat{\phi}^{-2}(D - 2t_1) + \hat{\phi}^{-4}t_{11}\right)^{-1}$$

$$t_1 = \hat{\phi}^{-1}\, trace(\hat{\mathbf{T}}_{(2)})$$

$$t_{11} = trace(\hat{\mathbf{T}}_{(2)}\hat{\mathbf{T}}_{(2)})$$

$$\Delta = \mathbf{A}\hat{\boldsymbol{T}}_s$$

$$\hat{\nabla}_i = \partial(\Delta_i)/\partial\boldsymbol{\phi}\big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} = \partial(\boldsymbol{A}_i\hat{\boldsymbol{T}}_s)/\partial\boldsymbol{\phi}\big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$$

Results and Discussions

- 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 [1].
- 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 in LLMM are

$$u_d \, (d = 1, \ldots, D)$$

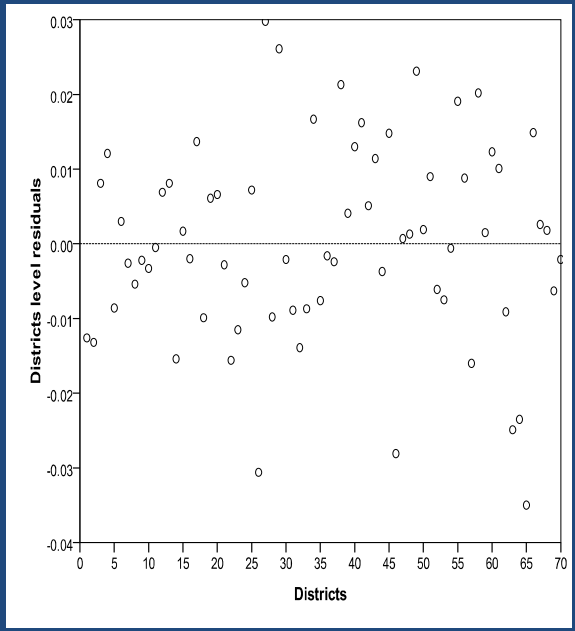
- These are assumed to have a normal distribution with mean zero and variance φ . If the model assumptions are satisfied then the district level residuals are expected to be randomly distributed and not significantly different from the regression line *y=0*, where under the LLMM , the area level residuals are defined as

$$r_d = \hat{\eta}_d - \mathbf{x}'_d \hat{\boldsymbol{\beta}}$$

The Figure 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.

Figure 1. Distribution of the district level residuals (left hand side) and normal q-q plot of the district level residuals (right hand



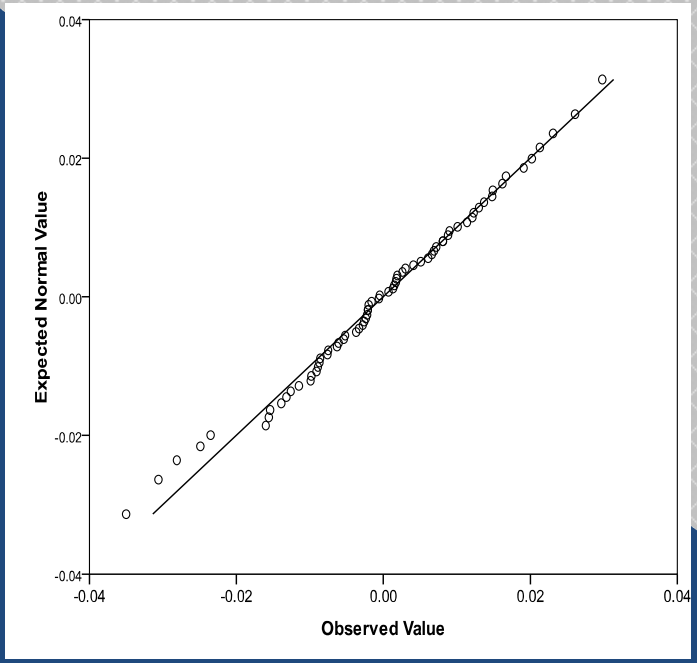
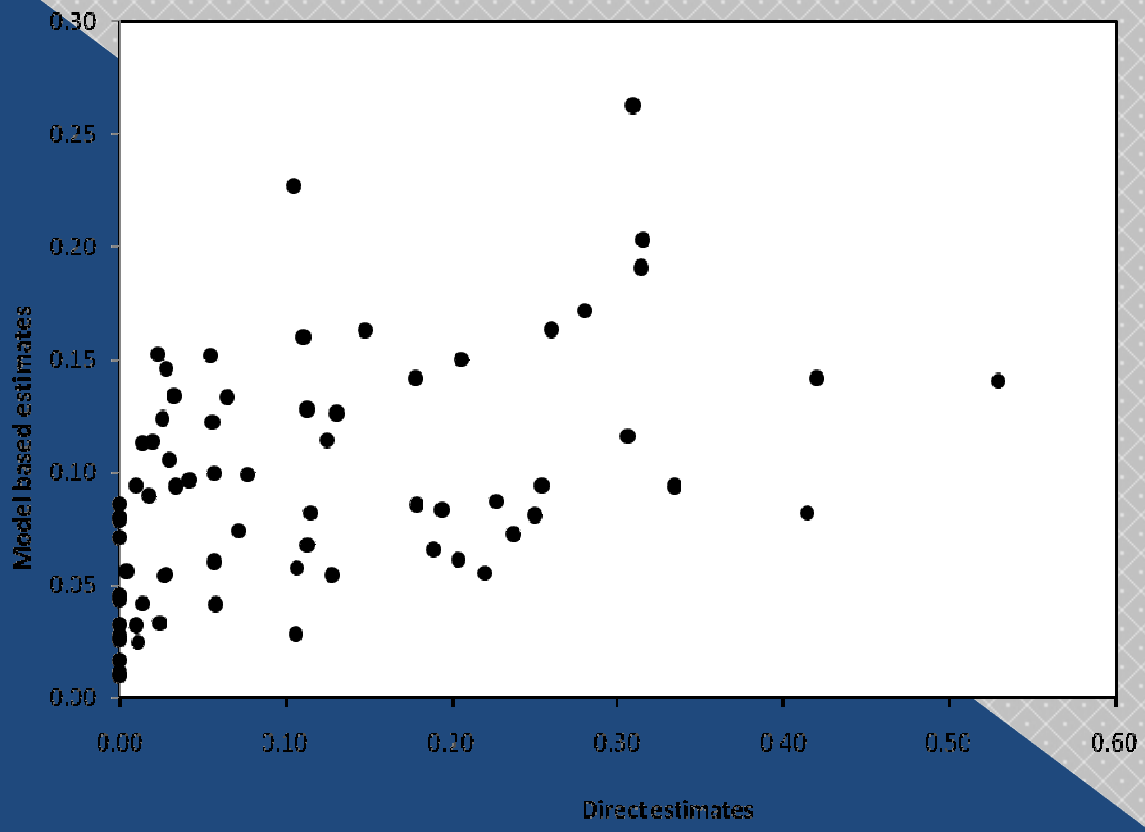


Figure 2. Bias diagnostics plot



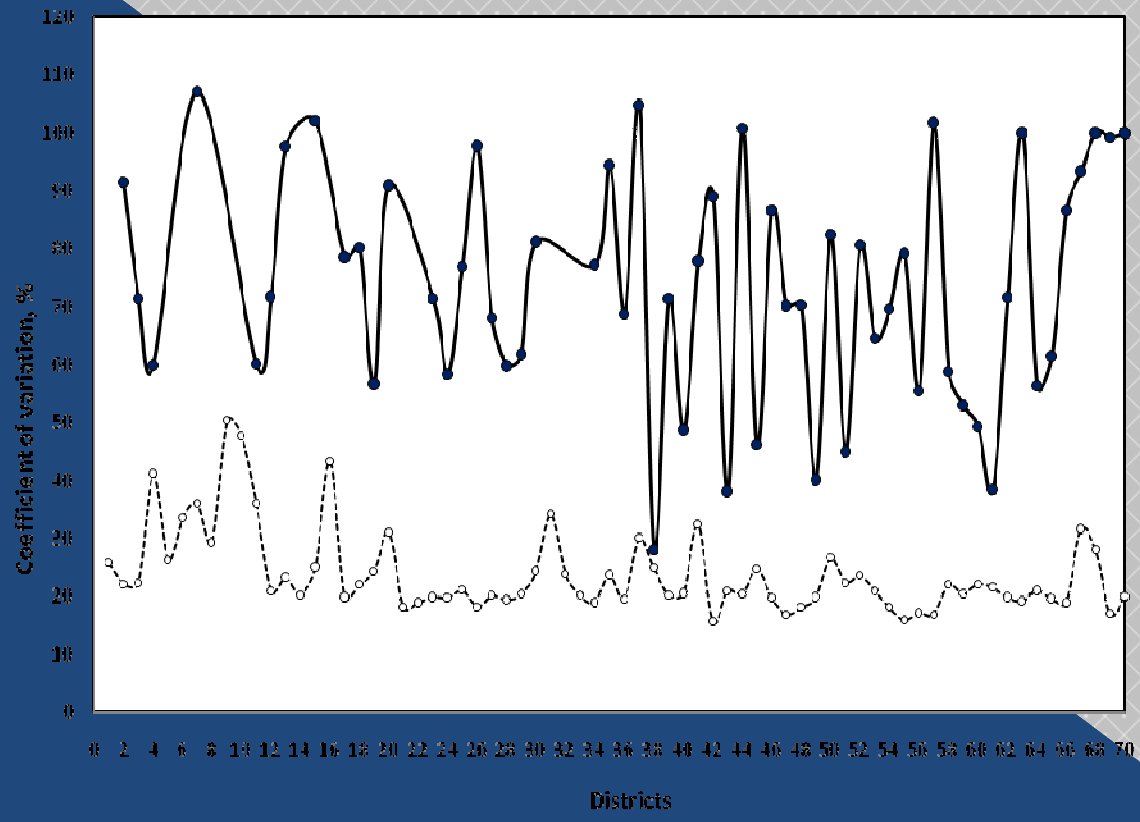
Contd.

- To validate the reliability of the model-based small area estimates the bias diagnostics, coefficient of variation (CV) and the 95 percent confidence intervals were computed
- 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 [5].
- The bias scatter plot of the model-based estimates against the direct estimates is set out in Figure 2.
- The plot show that the model-based estimates are less extreme as compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average.

Contd.

- The CVs were computed 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 to judge how large is 'too large' ([1] and [5]).
- Figure 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.

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



Contd.

Table 2 gives the district-wise 95 percent 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.

For many districts it is not possible to produce the confidence intervals due to unavailability of standard errors.

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

	Direct			Model-based				Direct			Model-based		
District	Estimate	Lower	Upper	Estimate	Lower	Upper	District	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	0.00	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	0.00	0.000	0.01	0.001	0.019	51	0.41	0.042	0.786	0.08	0.045	0.119

Contd.

0.02	-0.01	0.046	0.09	0.054	0.126	52	0.24	-0.145	0.619	0.07	0.038	0.107
0.03	-0.02	0.086	0.13	0.075	0.193	53	0.12	-0.033	0.263	0.08	0.047	0.116
0.11	-0.015	0.241	0.07	0.035	0.101	54	0.11	-0.044	0.27	0.13	0.082	0.174
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
00.0	0.004	0.004	0.06	0.036	0.077	56	0.26	-0.028	0.548	0.16	0.107	0.219
0.00	0.00	0.000	0.08	0.05	0.111	57	0.03	-0.027	0.079	0.12	0.082	0.166
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
0.04	-0.007	0.091	0.1	0.058	0.135	59	0.18	-0.011	0.367	0.14	0.084	0.200
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
0.06	-0.054	0.168	0.1	0.063	0.136	61	0.32	0.073	0.557	0.20	0.116	0.291
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
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
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
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
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
0.00	0.000	0.000	0.07	0.037	0.105	67	0.11	-0.092	0.304	0.03	0.01	0.046
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
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
0.06	-0.05	0.162	0.12	0.064	0.181	70	0.01	-0.010	0.030	0.09	0.056	0.132

Contd.

- 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, the estimated direct proportions were 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.

REFERENCES

1. R. Ambler, D. Caplan, R. Chambers, M. Kovacevic, and S. Wang, *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 Int. Assoc. of Survey Stat., Meeting of the ISI, Seoul, August 2001.
2. G. E. Battese, R. M. Harter, and W. A. Fuller, *An error component model for prediction of county crop areas using survey and satellite data,* J. of the Amer. Stat. Assoc. 83 (1988), pp. 28-36.
3. N. E. Breslow and D. G. Clayton, *Approximate inference in generalized linear mixed models,* J. of the Amer. Stat. Assoc. 88(1993), pp.9-25.
4. R. E. Fay and R. A. Herriot, *Estimation of income from small places: an application of james-stein procedures to census data,* J. of the Amer. Stat. Assoc. 74(1979), pp. 269-277.
5. F.A. Johnson, H. Chandra, J. J. Brown, and S. Padmadas, *District-level estimates of institutional births in ghana: application of small area estimation technique using census and DHS data*, J. of Off. Stat. (2009), to appear

Contd.

6. G.W. Manteiga, M.J. Lombardìa, I. Molina, D. Morales, and L. Santamarìa, *Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed mode*l, Comput. Stat. & Data Anal. 51(2007), pp. 2720-2733.
7. D. Pfeffermann, *Small area estimation: new developments and directions*. Int. Stat. Rev. 70(2002), pp.125-143.
8. J.N.K. Rao, *Small Area Estimation*. Wiley Series in Survey Methodology, John Wiley and Sons Inc, 2003.
9. A.Saei and R. Chambers, *Small area estimation under linear and generalized linear mixed models with time and area effects, W.P. No. M03/15*(2003), Southampton Statistical Sciences Research Institute, University of Southampton, UK.
10. C.E.Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

THANK YOU