Estimation of District Level Poor Households in the State of Uttar Pradesh in India by Combining NSSO Survey and Census Data- An Application of Small Area Estimation

Chandra, Hukum
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi-110012, India
E-mail: hchandra@iasri.res.in

Sud, Umesh
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi-110012, India
E-mail: ucsud@iasri.res.in

Salvati, Nicola
University of Pisa,
Pisa, Italy
E-mail:salvati@ec.unipi.it

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 (see [7] and [8]). 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 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 [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. In this paper we use SAE 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. 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

Two types of variables are required for this analysis. (1) The variable of interest for which small area estimates are required is drawn from the HCE 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. (2) The auxiliary (covariates) variables are drawn from the Population Census 2001. There were more than 100 covariates available 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.

3. METHODOLOGY

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 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

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

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 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 $\pi_d = \exp(\eta_d)\{1+\exp(\eta_d)\}^{-1}$. 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 [8]. Such a model was originally 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. 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 (see [3]) and suitable for discrete, particularly binary variable. It is noteworthy that the FH model is not applicable in such cases. Saei and Chambers [9] and Manteiga et al. [6] 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:

$$E\left(y_{sd}\,|u_d\right) = n_d\pi_d = n_d\left[\exp(\mathbf{x}_d'\boldsymbol{\beta}+u_d)\left(1+\exp(\mathbf{x}_d'\boldsymbol{\beta}+u_d)\right)^{-1}\right] \qquad (2)$$

$$E\left(y_{rd}\,|u_d\right) = \left(N_d - n_d\right)\pi_d = \left(N_d - n_d\right)\left[\exp(\mathbf{x}_d'\boldsymbol{\beta}+u_d)\left(1+\exp(\mathbf{x}_d'\boldsymbol{\beta}+u_d)\right)^{-1}\right]. \qquad (3)$$

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,

$$\hat{T}_d = y_{sd} + \hat{y}_{rd} = y_{sd} + \left(N_d - n_d\right)\left[\exp(\mathbf{x}_d'\hat{\boldsymbol{\beta}}+\hat{u}_d)\left(1+\exp(\mathbf{x}_d'\hat{\boldsymbol{\beta}}+\hat{u}_d)\right)^{-1}\right]. \qquad (4)$$

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

$$\hat{p}_d = \hat{T}_d N_d^{-1} = N_d^{-1}\left\{y_{sd} + \left(N_d - n_d\right)\left[\exp(\mathbf{x}_d'\hat{\boldsymbol{\beta}} + \hat{u}_d)\left(1 + \exp(\mathbf{x}_d'\hat{\boldsymbol{\beta}} + \hat{u}_d)\right)^{-1}\right]\right\}. \quad (5)$$

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 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 REML estimation of ϕ to estimate these unknown parameters. Detailed description of the approach can be followed from [6, 9].

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 MSE estimate of (5) under model (1) is (see [6, 9]) 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 ([8]). 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 $\hat{\mathbf{V}}_{sd} = diag\left\{n_d\hat{p}_d(1-\hat{p}_d)\right\}$ and $\hat{\mathbf{V}}_{rd} = diag\left\{(N_d - n_d)\hat{p}_d(1-\hat{p}_d)\right\}$, the diagonal matrices defined by the corresponding variances of the sample and non-sample part respectively. $\mathbf{A} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}$, $\mathbf{B} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}\boldsymbol{X}_r - \mathbf{A}\hat{\boldsymbol{T}}_s\hat{\mathbf{V}}_{sd}\boldsymbol{X}_s$ and $\hat{\boldsymbol{T}}_s = \left(\phi^{-1}\mathbf{I}_D + \hat{\mathbf{V}}_{sd}\right)^{-1}$, where $\boldsymbol{X}_s$ and $\boldsymbol{X}_r$ are the sample and non-sample part of covariates and $\mathbf{I}_D$ is an identity matrix of order *D.* We further write $\hat{\mathbf{T}}_{(1)} = \left\{\boldsymbol{X}_s'\hat{\mathbf{V}}_{sd}\boldsymbol{X}_s - \boldsymbol{X}_s'\hat{\mathbf{V}}_{sd}\hat{\boldsymbol{T}}_s\hat{\mathbf{V}}_{sd}\boldsymbol{X}_s\right\}^{-1}$ and $\hat{\mathbf{T}}_{(2)} = \hat{\boldsymbol{T}}_s + \hat{\boldsymbol{T}}_s\hat{\mathbf{V}}_{sd}\boldsymbol{X}_s\hat{\mathbf{T}}_{(1)}\boldsymbol{X}_s'\hat{\mathbf{V}}_{sd}'\hat{\boldsymbol{T}}_s$. With these notations, assuming model (1) holds, the various components of equation (6) are

$$m_1(\hat{\phi}) = \mathbf{A}\hat{\boldsymbol{T}}_s\mathbf{A}', \quad m_2(\hat{\phi}) = \mathbf{B}\hat{\mathbf{T}}_{(1)}\mathbf{B}', \text{ and } \quad m_3(\hat{\phi}) = trace\left(\hat{\nabla}_i\hat{\boldsymbol{\Sigma}}\hat{\nabla}_j' v(\hat{\phi})\right) \quad \text{with } \hat{\boldsymbol{\Sigma}} = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}'.$$

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 $\Delta = \mathbf{A}\hat{\boldsymbol{T}}_s$ and $\hat{\nabla}_i = \partial(\Delta_i)/\partial\phi\big|_{\phi=\hat{\phi}} = \partial(\boldsymbol{A}_i\hat{\boldsymbol{T}}_s)/\partial\phi\big|_{\phi=\hat{\phi}}$, where $\boldsymbol{A}_i$ is the i^{th} row of the matrix $\boldsymbol{A}$.

4. RESULTS AND DISCUSSIONS

Generally two types of diagnostics procedures are tested in SAE, 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 $u_d\,(d = 1,...,D)$ 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 Figure 1. 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.

To validate the reliability of the model-based small area estimates we used the bias diagnostics, coefficient of variation (CV) and computed the 95 percent 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 [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. 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’ ([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. In Table 2 we present the districts-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. Note that for many districts we can even not produce the confidence intervals due to unavailability of standard errors.

The small area estimates diagnostic measures clearly depict that the model-based estimates are reliable and more stable than the corresponding direct estimates (Figure 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. 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 ([6 and 9]), 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.

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), 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), 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.

[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. 5, 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), S3RI, University of Southampton, UK.

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