J. Ind. Soc. Agril. Statist.
61(2), 2007 : 249-254

Small Area Estimation - An Application to National Sample Survey Data

A.K. Srivastava, U.C. Sud[1] and Hukum Chandra[1]
Indian Society of Agricultural Statistics, New Delhi

SUMMARY

This article uses already available small area estimation techniques to derive district level estimates of amount of loan outstanding per household using data from the 2002-03 Debt-Investment Survey of National Sample Survey Organization (NSSO) for the rural areas of Uttar Pradesh. Fay and Herriot model (Fay and Herriot, 1979) has been used to obtain the model-based district level estimates. The diagnostic analysis shows that the model-based estimates are reasonably reliable and representative of the districts to which they belong.

Key words: Empirical best linear unbiased predictor, Small area estimation, National Sample Survey Organization.

1. INTRODUCTION

For planned development of a country, information on various aspects of economy is required to be collected on regular basis. The information can be collected through Census i.e. complete enumeration of the population under study. However, the conduct of Census is very time consuming, involves massive operations requiring huge resources, besides, being subject to large errors. Consequently, these can only be conducted after fairly long time gaps, which vary from country to country. In India, while the Population and Economic Censuses are conducted every 10 years, the Agricultural and Livestock Censuses are conducted every five years. For obtaining information during the intervening periods, large scale sample surveys are resorted so that reliable, timely and adequate information on the parameters of interest from large populations can be provided. In India, National Sample Survey Organisation (NSSO) carries out country wide surveys on various socio-economic parameters related to the national economy such as follow up enterprise surveys of Economic Census, Annual Survey of Industries, supervision of Area enumeration and Crop Estimation surveys conducted by the state agencies so that appropriate data can be made available for policy planning and decision making on various issues of national importance. Similarly, the crop cutting experiments are organized by the Directorate of Economics and Statistics for estimation of yield rates of various crops under the scheme of General Crop Estimation Surveys (GCES). While the sample sizes for the surveys conducted by the NSSO are fixed in such a manner that it is possible to get reasonably precise estimates at the State level, the sample size in the GCES are adequate to provide estimates at the District level. Due to the emphasis on micro-level planning reliable estimates of various parameters of interest are being demanded by the administrators and policy planners at the small area level. A small area in the context of NSSO surveys may be a district while it may be a Community Development Block/ Gram Panchayat in case of GCES. In view of the demand for reliable statistics at the local level there is a burst of activity in the area of Small Area Estimation (SAE) technique. Newer techniques are increasingly being developed using tools of statistical inference and linear model. Simultaneously, attempts are also being made to apply these techniques so that precise estimates are available at the small/local area level. In many

[1] *Indian Agricultural Statistics Research Institute, New Delhi*

countries, SAE techniques are extensively used to produce the lower area level estimates, e.g. in United Kingdom the estimate of unemployment levels and rates for their Local Authority Districts (Ambler *et al.* 2001) and in United States the estimates of poor school-age children at County level (Citro and Kalton, 2000). In India also, attempts have been made to use SAE techniques for various purposes (Sharma *et al.* 2004).

The growing demand for small area statistics in recent years has increased the popularity of SAE techniques. In this context model-based methods are widely used (Rao, 2003, Chapter 2). The underlying idea is to use statistical models to link the variable of interest with auxiliary information to define the model-based estimator for small areas. Since the area-specific direct estimators do not provide adequate precision, for generating estimates for small areas it is necessary to employ model-based estimators that "borrow strength" from the related area. Small area model based techniques 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; these relate small area direct estimators to area-specific covariates (Fay and Herriot, 1979), and (ii) nested error unit level regression models, employed originally by Battese *et al.* (1988) for predicting areas under corn and soybean in 12 counties of the state of Iowa in the USA, these models relate the unit values of a study variable to unit-specific covariates.

The purpose of the study is to apply already available SAE technique. To achieve this we used NSS and Agriculture Census (1995-96) data to produce precise district level estimates. In particular, we employed an area level small area model to compute the empirical best linear unbiased predictor estimates and its mean squared error estimates because covariates, collected from Agriculture Census, are available at area level. Throughout this paper district and small area (or area) is used interchangeably.

2. THE EMPIRICAL BEST LINEAR UNBIASED PREDICTOR FOR SMALL AREAS

In the small area estimation method used here the covariates are collected from the Census which are available at district level. Here districts are small area of interest. Widely used 'area level random effects model' is used because the auxiliary information is available only at the area level. This 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 USA, often referred to as Fay and Herriot model (hereafter FH model). In area level model there are two components:

(i) The direct survey estimate of the parameter based on the sampling design, expressed as

$$Y_d = y_d + e_d, \quad d = 1,...,D \qquad (1)$$

where D is total number of small areas that constitute our finite population, y_d are unobserved small area means (i.e., our parameter of interest), Y_d are observed direct survey estimators (the sample mean in our case) and the e_d's are independent sampling errors of survey estimate with $E(e_d/y_d) = 0$ and $V(e_d/y_d) = v_d$. Model (1) is a sampling model and v_d is a design-based sampling variance.

(ii) A linking model

$$y_d = z_d^T\beta + u_d \ , \ d=1,..... D \qquad (2)$$

where z_d denotes p-vector of area (or district) level covariates, β is a p-vector of unknown fixed-effect coefficients and u_d are random effects (also called the model errors), assumed to be independent and identically distributed with

$E(u_d) = 0$ and $V(u_d) = \sigma_u^2$.

Combining (1) and (2), we obtain the model

$$Y_d = z_d^T\beta + u_d + e_d \ , \ d=1,..... D \qquad (3)$$

Clearly, model (3) integrates a model dependent random effect u_d and a sampling error e_d with the two errors being independent. Model (3) is a special case of the linear mixed model. For known variance σ_u^2, assuming model (3) holds, the Best Linear Unbiased Predictor (BLUP) for y_d (Henderson, 1963) is given by

$$\begin{aligned}\tilde{y}_d &= z_d^T\hat{\beta}_{GLS} + \gamma_d(Y_d - z_d^T\hat{\beta}_{GLS}) \\ &= \gamma_d Y_d + (1-\gamma_d)z_d^T\hat{\beta}_{GLS} \qquad (4)\end{aligned}$$

where $\gamma_d = \sigma_u^2/(v_d + \sigma_u^2)$ and

$$\hat{\beta}_{GLS} = (\sum\nolimits_d (v_d + \sigma_u^2)^{-1} z_d z_d^T)^{-1} (\sum\nolimits_d (v_d + \sigma_u^2)^{-1} z_d Y_d)$$

is the generalised least square estimate of β. In practice, the variance σ_u^2 is usually unknown and it is replaced by sample estimates, $\hat{\sigma}_u^2$ (in equation (4) and $\hat{\beta}_{GLS}$) yielding the corresponding Empirical BLUP (EBLUP) denoted by $\hat{y}_d$. We note that the EBLUP $\hat{y}_d$ is a linear combination of a direct estimate Y_d and the model dependent regression synthetic estimate $z_d^T\hat{\beta}_{GLS}$, with weights given by $\hat{\gamma}_d^1$. Here $\hat{\gamma}_d$ is called 'shrinkage factor' since it 'shrinks' the direct estimator towards the synthetic estimator $z_d^T\hat{\beta}_{GLS}$ (Rao 2003, Chapter 5).

Turning to mean squared error (MSE) estimation, if β and σ_u^2 are also known, the variance of the BLUP (4) is given as $\text{Var}\left[\hat{y}_d(\sigma_u^2,\beta)\right] = \gamma_d v_d = g_{1d}$

In practice, β and σ_u^2 are estimated from the sample data and substituted for the true values, giving rise to the EBLUP. A naïve variance estimator is obtained by replacing σ_u^2 by $\hat{\sigma}_u^2$ in g_{1d}. This estimator ignores the variability of $\hat{\sigma}_u^2$ and hence underestimates the true variance. Prasad and Rao (1990), extending the work of Kackar and Harville (1984) approximate the true prediction MSE of the EBLUP under normality of the two error terms and for the case where σ_u^2 is estimated by the ANOVA (fitting of constants) method as

$$\text{MSE}[\hat{y}_d(\hat{\sigma}_u^2,\hat{\beta}_{GLS})] = g_{1d} + g_{2d} + g_{3d} \quad (5)$$

where $g_{2d} = (1-\gamma_d)^2 z_d^T \text{Var}(\hat{\beta}_{GLS}) z_d$ with $\text{Var}(\hat{\beta}_{GLS}) = \left(\sum_d (v_d + \sigma_u^2)^{-1} z_d z_d^T\right)^{-1}$ is the excess in MSE due to estimation of β and $g_{3d} = [\sigma_{Di}^4/(\sigma_{Di}^2 + \sigma_u^2)^3] \times \text{Var}(\hat{\sigma}_u^2)$ is the excess in MSE due to estimation of σ_u^2. The neglected terms in the approximation are of order o(1/D). Building on the approximation, Prasad and Rao (1990) derive a MSE estimator of (5) with bias of order o(1/D) as

$$\text{MSE}[\hat{y}_d(\hat{\sigma}_u^2,\hat{\beta}_{GLS})] = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2) \quad (6)$$

where $g_{kd}(\hat{\sigma}_u^2)$ is obtained from g_{kd} by substituting $\hat{\sigma}_u^2$ for σ_u^2, k =1,2,3. The MSE estimator (6) is robust with respect to departures from normality of the random area effects u_d (but not the sampling errors e_d) (Lahiri and Rao, 1995). Here, standard error of the EBLUP is calculated as square root of MSE. Note that the leading term in (6) is $g_{1d} = \gamma_d v_d$ so for the small values of γ_d (i.e., the model variance σ_u^2 is small relative to the sampling variance v_d), $\text{MSE}[\hat{y}_d(\hat{\sigma}_u^2,\hat{\beta}_{GLS})] < v_d = V_D(Y_D)$ illustrating the possible gains from using the model dependent estimator. Further, the availabilty of good auxiliary data is a key to successful application of the small area technique since this provides a basis for good model fit. An excellent example of application of this method is given by Citro and Kalton (2000).

3. EMPIRICAL STUDY

The theory described in the previous section has been applied to develop district level estimates using the NSSO data. For this purpose we have used NSSO 59th round data for rural areas on Debt and Investment survey conducted for the calendar year 2003 in the State of Uttar Pradesh (UP). The sampling design used in this survey was one of stratified multi-stage random sampling with districts as strata, villages as first stage units and households as the second stage units. The variable used for the study was average amount of loan outstanding per household (A household is defined to be indebted if it has outstanding loan as on 30.6.2002). Thus the parameter of interest was average amount of loan outstanding per household at the district level. For the purpose of implementation of EBLUP the following district level covariates, which were available from the Agriculture Census (1995-96) were used: (1) Area under semi-medium category of holding, (2) Area under medium category of holding, (3) Area under large category of holding, (4) Number of large holdings, (5) Rural scheduled caste population, and (6) Percentage irrigated area. The State of Uttar Pradesh has 70 districts. Due to non-availability of data on the covariates for all the districts the analysis of the data was restricted to only 45 districts. The analysis was carried out using SAS and EBLUP of average amount of loan outstanding/household was obtained.

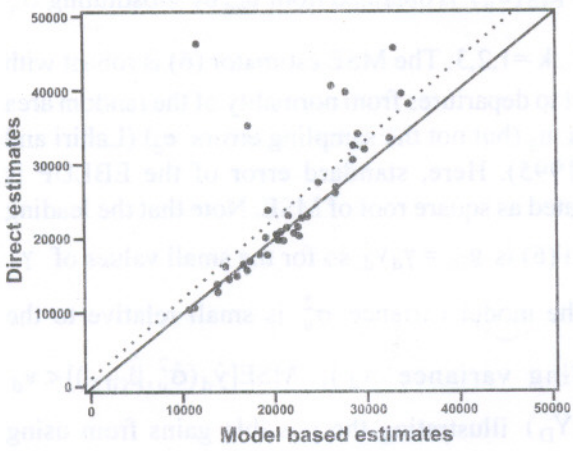

Fig 1 (a)

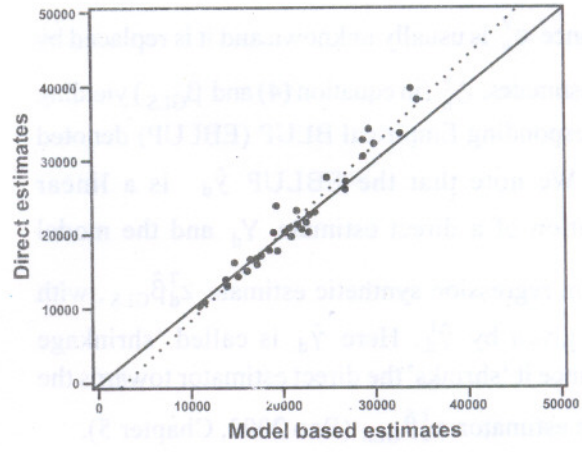

Fig 1 (b)

Fig 1. Bias diagnostics plot showing the ordinary least square regression line (dash line) and y = x line (solid line). The left plot include all estimates while the right exclude five outlying estimates.

3.1 Diagnostics for Small Area Estimates

The aim of this diagnostics procedure is to validate the reliability of the model-based small area estimates versus direct survey estimates. The diagnostic procedures used are (1) bias diagnostics, (2) goodness of fit diagnostics, (3) coverage diagnostics and (4) coefficient of variation diagnostics.

3.1.1 Bias Diagnostics

The bias diagnostic is used to assess the deviation of the model-based estimates from the direct survey estimates. The model-based estimates are expected to be biased predictors of the direct estimates. The model-based estimators will be unbiased predictors of the direct survey estimates if the relationship between the variable of interest and the auxiliary variables have been misspecified or misestimated. Where the relationship has not been misspecified or misestimated, a linear relationship of the type y = x is expected between the direct survey estimates and the model-based estimates. Fig. 1 shows the bias scatter plot of the direct survey estimates against the model-based estimates with the fitted regression line and the y = x line. The value of R^2 for the ordinary least square (OLS) regression line is 0.51. Further, we observe that OLS regression line is deviating from the y = x line. This is because of a (few outlying direct estimates. Excluding these five extremely outlying estimates, the OLS regression line is very close to y = x line with R^2 equal to 0.95, Fig. 1(b).

3.1.2 Coverage Diagnostics

The coverage diagnostics measure the overlap between the 95% confidence intervals of the direct survey estimates and those of the model-based estimates. This diagnostics is aimed at evaluating the validity of the confidence intervals generated by the model-based procedure. Let X and Y be two independent random variables, with the same mean but different standard deviations σ_x and σ_Y respectively and z(a) be such that the probability that a standard normal variable takes values greater than $z(\alpha)$ is $\alpha/2$. Then for a probability α that the two intervals $X \pm z(\beta)\sigma_X$ and $Y \pm z(\beta)\sigma_Y$ do not overlap can be defined as

$$z(\beta) = z(\alpha)\left(1+\frac{\sigma_X}{\sigma_Y}\right)^{-1}\sqrt{1+\frac{\sigma_X^2}{\sigma_Y^2}}$$

To compute $z(\beta)$, $z(\alpha)$ is set at 1.96, σ_x is the estimated standard error of the model-based estimates and σ_Y is the estimated standard error of the direct estimate. $z(\beta)$ is then used to compute the overlap proportion between the direct estimates and the model-based estimates. It is recommended that non-coverage total should not exceed 5%. In our case, there is 100% coverage between the intervals of the model-based estimates and direct survey estimates. This indicates that the method is statistically acceptable.

3.1.3 Goodness of Fit Diagnostics

The goodness of fit diagnostics test whether the model-based estimates are close to the direct estimates. In other words, one could ask – does the geographical variation in the auxiliary variables explain the observed variation in the variable of interest? The approach uses Wald goodness of fit statistic to test whether there are significant differences between the expected values of the model-based estimates and the direct estimates. This diagnostic is carried out by computing the differences between the model-based and direct estimates which are then squared and weighted inversely by their variances and summed over all the domains. This test statistic is then compared to a chi-square distribution with degrees of freedom equal to the number of small areas, in our case districts, in the population. This provides a parametric significance test of bias of model-based estimates relative to their precision. The estimated goodness of fit statistic in the final model was 9.45 with 45 degrees of freedom and corresponding test statistics from the table was 61.66 indicating that anything larger than 61.66 is significant. This shows that these model-based estimates are statistically acceptable (Chambers *et al.* 2007).

3.1.4 Coefficient of Variation

The Coefficient of Variation (CV) is a statistical measure of the dispersion which provides unit free measure of reliability for the estimate. The CV is the ratio of the standard deviation of the estimate to its mean and expressed as a percentage. Estimates with large CVs are considered unreliable. Fig. 2 shows the CV plots for the model-based and direct estimates. It can be seen from the plots that the CVs from the model-based estimates are more stable than the CVs from the direct estimates.

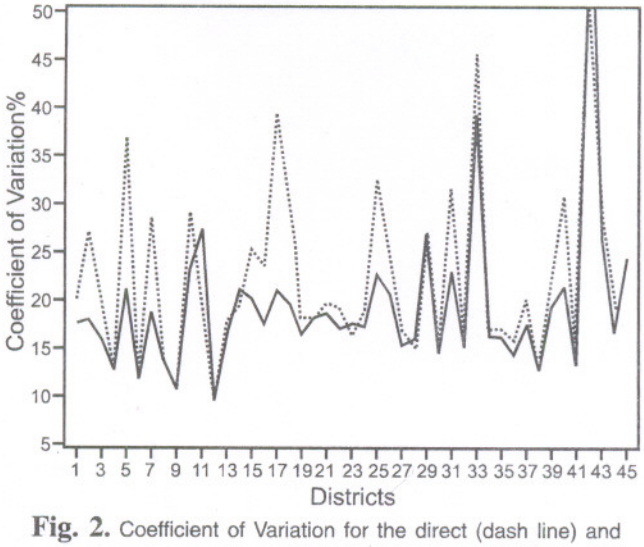

Fig. 2. Coefficient of Variation for the direct (dash line) and model-based (solid line) estimates.

4. CONCLUSION

The model-based method has been found to be very effective for developing district level estimates of average amount of loan outstanding per household. For most of the districts the reduction in coefficient of variation is quite evident. However, the diagnostics results presented in previous section show only marginal gains in the model-based estimates. This was expected since we used 1995-96 Agriculture Census data (the latest census data was not available) for collecting information on the covariates. Due to this we could not get very high correlation between the study variable and the covariates. We already indicated that the success of model-based SAE methods lie in the correct specification of the underlying model and availability of good covariates. This possibly explains the aberration in the diagnostics results.

ACKNOWLEDGEMENTS

The authors express sincere thanks to Dr. V.K. Mahajan, Principal Scientist, IASRI, New Delhi-12, for making computer program for data analysis.

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. *Proc. of the International Association of Survey Statisticians, Meeting of the International Statistics Institute,* 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.*, **95**, 1127-1142.

Chambers, R., Chandra, H. and Tzavidis, N. (2007). Small Area Estimation Course Notes. *Third International Conference on Establishment Surveys, Montreal, Canada*, June **18-21**, 2007.

Citro, C. and Kalton, G. (2000). Small-area estimates of school-age children in poverty. In: *Evaluation of current methodology* (National Research Council), Nat. Acad. Press, Washington DC .

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

Henderson, C.R. (1963). Selection Index and expected genetic advance. In: *Statistical Genetics and Plant Breeding*, eds. W.D. Hanson and H.F. Robinson, National Academic of Sciences- National Research Council, Washington, DC 141-163.

Kackar, R.N. and Harville, D.A.(1984). Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *J. Amer. Statist. Assoc.*, **79**, 853-862.

Lahiri, P. and Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **90**, 758-766.

Prasad, N.G.N and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-71.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.

Sharma, S.D., Srivastava, A.K. and Sud, U. C.(2004). Small area crop estimation methodology for crop yield estimates at Gram Panchayat level. *J. Ind. Soc. Agril. Statist.*, 26-38.