J. Ind. Soc. Agril. Statist.
61(3), 2007 : 295-309

Small Area Estimation – A Perspective and Some Applications[1]

A.K. Srivastava
Former Joint Director, Indian Agricultural Statistics Research Institute, New Delhi

PROLOGUE

I feel privileged in getting this opportunity to deliver this prestigious lecture in the memory of Dr. V.G. Panse. For all of us, who started our careers in Agricultural Statistics in mid sixties and later, Dr. Panse has been an ideal icon all along. He was not only a visionary but had the innate capability to transform the vision into reality. I had joined the Indian Agricultural Statistics Research Institute (IASRI, known as Institute of Agricultural Research Statistics (IARS) in those days) in August 1965, first as a student and then as a staff member. During my entire tenure, I have not come across any component of Agricultural Statistics System, which does not have an imprint of Dr. Panse. Whether it is statistics relating to crops, livestock, fisheries, minor crops, cost of cultivation or training in agricultural statistics, Dr. Panse's contributions have been a guiding light for the entire statistical community. His patronage and guidance to IARS and to the Indian Society of Agricultural Statistics (ISAS) has been instrumental in providing a sound base to both these organizations. Besides his role in the national context, his contributions to the Agricultural Statistics Systems of various countries through international organizations like FAO made him a highly respected statistician worldwide.

My choice of the topic on Small Area Estimation (SAE) for this occasion has got a purpose. Research in the area of SAE in the present form has started getting recognition in late sixties. A less recognized fact in this context is that in mid sixties, there was a paper by Panse *et al.* (1966). The results were based on a Scheme on Block Level Estimates, undertaken at IARS under the guidance of Dr. Panse. In this paper, the problem for providing estimates at small area levels (Blocks in this case), was addressed through a classical sample survey approach. Lot of developments in SAE has taken place subsequently. Through the present paper, I propose to share our perspective and some experiences in this area – a tribute to Dr. Panse.

1. INTRODUCTION

Surveys are normally planned with specific populations in view. Quite often, interest also lies in parts of the population known as subpopulations or domains of interest. Domain parameters may be estimated satisfactorily through usual sample survey approach provided the domains get sufficient representation of sampled units in the main sample. Sometimes, the subpopulations or domains are too small to provide reliable direct estimates. The term small domain or area typically refers to the part of a population for which reliable statistics of interest cannot be produced due to certain limitations of the data.

The topic of small area estimation has gained importance in view of growing needs of micro level planning. Demands for reliable Small Area Statistics (SAS) are increasing both from public and private sectors with growing concerns of governments relating to issues of distribution, equity and disparity. The need for statistics at lower levels has been felt for a long time and efforts

[1] *Dr. V.G. Panse Memorial Lecture delivered at 61st Annual Conference of the Indian Society of Agricultural Statistics at Birsa Agricultural University, Kanke, Ranchi (Jharkhand)*

have been made to meet the requirements through some traditional approaches. In a historical perspective, Brackstone (1987) tracked the references of SAS to the eleventh century England and seventeenth century Canada. In late sixties small area estimation got an impetus with the application of synthetic method of estimation in a disability survey by National Center for Health Statistics (NCHS 1968). Purcell and Kish (1979) reviewed the methods for SAS available till that time. Most of the methods were developed in the context of population studies. Applications too were mainly in the field of population/demographic studies. The current emphasis and recent advances in this area have been mainly due to significant advances in statistical data processing. The advances in computing facilities have also provided convenient tools for many theoretical developments in this area. Since traditional sampling theory fails to provide reliable and valid estimates in this situation, many SAE techniques have been developed which make use of information from other sources. They also borrow strength from related or similar areas through explicit and implicit models that connects the small area via supplementary data.

In this lecture we discuss some of the developments and some experiences relating to applications of SAE methods in agriculture and allied fields.

2. EARLY DEVELOPMENTS

2.1 Demographic Methods

Most of the SAE techniques in the early stages were developed in the context of demographic studies. These may be broadly categorized as Symptomatic Accounting Techniques (SAT). Such techniques utilize current data from administrative registers in conjunction with related data from the latest census. One of the most important SAT techniques is the Vital Rate (VR) method which uses birth and death rates as symptomatic variables. The method heavily depends on the assumption that the ratio of birth (death) rates in current year to those in the latest census year for the local area is approximately equal to the corresponding ratios for the larger area. There are several improvements on this method in the form of composite estimators.

One of the main reasons for application of these techniques in population studies was, perhaps the availability of various demographic data, which could be effectively used in adjusting the estimates for small areas.

There are several other traditional SAE methods available in literature. Purcell and Kish (1979) reviewed various methods available till that time. For an early review of other traditional methods, reference may be made to Platek *et al.* (1987). In the following section, we discuss synthetic method of estimation which is by far one of the most widely used SAE methods.

2.2 Synthetic Method

This technique is a simple common sense approach for small area estimation. The name synthetic estimation is credited to National Center for Health Statistics (NCHS 1968). A commonly acceptable description of synthetic estimator due to Gonzalez (1973) is as follows:

An unbiased estimate is obtained from a sample survey for a larger area. When this estimate is used to derive estimates for sub-areas having the same characteristics as the larger area, these estimates are identified as synthetic estimates.

Consider a population of size N, divided into two dimensions with H post-strata and D small areas (domains). The cell $\{(d, h)\ d = 1, \ldots, D;\ h = 1, \ldots, H\}$ consists of N_{dh} units with $N_{d.}$ and $N_{.h}$ as marginal totals. Let Y_{dh} be the total for the characteristic of interest and $Y_{.h} = \sum_d Y_{dh}$ for the cell (d, h). Reliable estimates of post-strata totals $Y_{.h}$ can be calculated from the survey data. We are interested in estimating the domain totals $Y_d = \sum_h Y_{dh}$ or domain mean $\bar{Y}_d = Y_d / N_d$, using known auxiliary variable totals X_{dh}.

The original synthetic estimator for $\bar{Y}_d$ considered at NCHS (1968) was of the form

$$\bar{Y}_d^s = \sum_h \frac{X_{dh}}{X_d} \hat{\bar{Y}}_{.h} \qquad (2.1)$$

where $\hat{\bar{Y}}_{.h}$ is the estimator of mean for h^{th} post-stratum. There are several variations of synthetic estimator. Rao (2002) considered an estimator for Y_d as

$$\hat{Y}_d^s = \sum_h X_{dh} (\hat{Y}_{.h} / \hat{X}_{.h}) \qquad (2.2)$$

where $\hat{Y}_{.h}$ and $\hat{X}_{.h}$ are reliable direct estimates of post-strata totals $Y_{.h}$ and $X_{.h}$ respectively. The bias of synthetic

estimator will be small if the ratios $R_{dh} = Y_{dh}/X_{dh}$ are homogeneous across small areas i.e. $R_{dh} = R_{.h} = Y_{.h}/X_{.h}$ for each h. The design variances of synthetic estimators are likely to be small as they depend on the post strata estimates only. Thus, if the biases are small, synthetic estimates are good enough. However, if the above assumptions for smallness of the biases are not valid then synthetic estimates are risky. An approximate unbiased estimator of Mean Square Error (MSE) of $\hat{Y}_d^s$ is given by

$$\text{MSE}\left(\hat{Y}_d^s\right) = \left(\hat{Y}_d^s - \hat{Y}_d\right)^2 - v\left(\hat{Y}_d\right) \qquad (2.3)$$

where $\hat{Y}_d$ is the direct estimator of Y_d and $v\left(\hat{Y}_d\right)$ is a design unbiased estimator of variance of $\hat{Y}_d$. MSE of synthetic estimator as in (2.3) may be highly unstable.

Synthetic estimators have been very widely used method of estimation. However, their biases have been a matter of concern and attempts have been made to mitigate them through the application of composite estimators.

3. MODEL BASED SAE METHODS

The SAE methods described above are indirect methods in which information from other sources (records, registers etc.) is utilized and strength is borrowed from other similar areas. They are invariably based on certain assumptions which are in the form of implicit models. We now consider some explicit model-based methods which are essentially mixed models and are used in specific situations based on data availability on the response variables of interest. These are (i) area level models where information on response variable is available only at the small area level; and (ii) unit level models where information on the response variable is available at the unit level. These models are described as follow.

3.1 Area Level Models

An area level model has two components:

(i) Direct survey estimator of the parameter based on the sampling design, expressed as

$$\hat{\theta}_d = \theta_d + e_d, \; d = 1, \ldots, D \qquad (3.1)$$

where e_d's are assumed to be independent across small areas with mean zero and known variances χ_d. The model (3.1) is a sampling model and χ_d is a design- based sampling variance.

(ii) A linking model

$$\theta_d = z_d^T\beta + v_d, \; d = 1, \ldots, D \qquad (3.2)$$

where the model errors v_d are assumed to be independent and identically distributed with mean zero and variance σ_d^2. The model variance σ_d^2 is a measure of homogeneity of the areas after accounting for the covariates z_d. Combining (3.1) and (3.2), the resultant mixed linear model is

$$\hat{\theta}_d = z_d^T\beta + v_d + e_d, \; d = 1,\ldots,D \qquad (3.3)$$

Using the data $\{(\hat{\theta}_d, z_d), \; d = 1, \ldots, D\}$, we can obtain estimates θ_d^*, of the realized values θ_d from the model (3.3). Here e_d's and v_d's are design-based and model-based random variables respectively.

Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB) methods have played an important role in the estimation of small area means $\bar{Y}_i$ under model (3.3). EBLUP method has been used in many practical applications. One of the early applications of this method was due to Fay and Herriot (1979). In fact, this method was adopted by the U.S. Bureau of Census in 1974 to form Per Capita Income (PCI) estimates for small places. EBLUP method is applicable for mixed linear models and it does not require normality assumption of the random errors v_d and e_d.

The other methods EB and HB are applicable under specific distributional assumptions. The inferences in HB methods are obtained through posterior distributions. EBLUP and EB are identical under normality assumptions. For EBLUP and EB, an estimator of $\text{MSE}(\tilde{\theta}_d) = E(\tilde{\theta}_d - \theta_d)^2$ is used as a measure of variability of $\tilde{\theta}_d$, where the expectation is with respect to the model (3.3).

EBLUP estimator of θ_d is a composite estimator of the form

$$\theta_d^* = \hat{\gamma}_d \hat{\theta}_d + (1-\hat{\gamma}_d) z_d^T \hat{\beta} \qquad (3.4)$$

where $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \chi_d)$ and $\hat{\beta}$ is the weighted least square estimator of β with weights $(\hat{\sigma}_v^2 + \chi_d)^{-1}$ obtained by regressing θ_d on z_d:

$\hat{\beta} = (\sum_d \hat{\gamma}_d z_d z_d^T)^{-1}(\sum_d \hat{\gamma}_d z_d \hat{\theta}_d)$ and σ_v^2 is an estimator of the variance component σ_v^2. It may be noted that θ_d^* is a linear combination of direct estimator $\hat{\theta}_d$ and the model based regression synthetic estimator $z_d^T\hat{\beta}$, with weights inversely proportional to their respective variances. For the non-sampled areas the EBLUP estimator is given by the regression synthetic estimator itself.

Under model (3.4), the leading term of MSE $(\tilde{\theta}_d)$ is given by $\gamma_d \chi_d$ which shows that the EBLUP estimate can lead to large gains in efficiency over the direct estimate with variance χ_d, when γ_d is small i.e. the model variance σ_v^2 is small relative to the sampling variance χ_d. Choice of good auxiliary data to provide a good model fit is, therefore the key to successful application of the small area estimation technique.

An excellent example of application of this method is in a study on Small Area Estimates of School-Age Children in Poverty (Constance *et al.* (2000)).

3.2 Unit Level Models

Consider a population of N units with d-th small area consisting of N_d units. Let y_{dj} and x_{dj} be the unit level y-value and correlated covariate x-value for j-th unit in the d-th small area. It is assumed that the domain mean $\bar{X}_d$ is known. Consider the following one-folded nested error linear regression model

$$y_{dj} = x_{dj}^T\beta + v_d + e_{dj}, \quad j = 1, ..., N_d; \; d = 1, ..., D \qquad (3.5)$$

where the random small area effects v_d have mean zero and common variance σ_v^2 and are independently distributed. Also, e_{dj} are assumed to be independently distributed with mean zero and variance σ_e^2 and are also independent of area effects v_d. This model was initially considered by Battese *et al.* (1988).

If N_d is large, the population mean $\bar{Y}_d$ is approximately equal to $x_d^T\beta + v_d$. The sample data $\{y_{dj}, x_{dj}\, j = 1, ..., n_d; d = 1, ..., D\}$ is assumed to satisfy the population model (3.5). This happens in equal probability sampling. This will also follow in probability proportional to size sampling when the size measure is taken as the covariate in the model. Assuming $\bar{Y}_d = \bar{X}_d^T\beta + v_d$, the EBLUP estimate of $\bar{Y}_d$ is of the form

$$\bar{y}_d^* = \hat{\gamma}_d[\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T\hat{\beta}] + (1-\hat{\gamma}_d)\bar{X}_d^T\hat{\beta}$$
$$d = 1, ..., D \qquad (3.6)$$

where $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_d^{-1})$ with estimated variance components $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$, and $\hat{\beta}$ is the weighted least square estimate of β. It may be noted that the EBLUP estimator is a composite estimator combining the survey regression estimator with the regression synthetic estimator.

Under model (3.5) for the sample data, the leading term of MSE $(\bar{y}_d^*)$ is given by $\gamma_d(\sigma_e^2/n_d)$, which shows that EBLUP estimator can lead to large gains in efficiency over the survey regression estimator when γ_d is small. Battese *et al.* (1988) applied the nested error regression model to estimate area under corn and soybeans at county level in North-Central Iowa using farm interview data in conjunction with LANDSAT satellite data.

For details of an exhaustive and thorough presentation of small area estimation an excellent reference is the book by Rao (2003). A more recent paper by Jiang and Lahiri (2006) provides an excellent overview and appraisal of mixed model prediction in the context of small area estimation.

4. APPLICATIONS IN AGRICULTURE – SOME EXPERIENCES AT IASRI

As already discussed, early attempts towards the application of small area methods were mainly in population studies. Although the need for estimation of important parameters at small area level in agriculture was always realized but applications have been rather scanty. Some of the applications in Indian context are as follows.

4.1 Crop Yield Estimation - Farmer Appraisal Approach

Crop production and crop yield are the most important parameters of interest. In India, the production and yield estimates are developed at district and higher levels. In early sixties, in order to obtain block (Community Development Blocks) level estimates, attempts were made within the framework of traditional sample survey approach. The technique of double sampling was used with eye estimates as an auxiliary variable (Panse *et al.* (1966), Singh 1968). Due to poor correlations in the study character and the auxiliary variable and some other limitations the approach could not be pursued further at that time. One of the limitations was that it could not be fitted in the approach of estimation being followed in the General Crop Estimation Surveys (GCES). Subsequently the approach has been tried in one of the projects at IASRI with a design conforming to the GCES approach (Sud *et al.* (2001)). However, these attempts are direct estimates and are based on the usual sample survey techniques for improvements of estimators. They do not fit into the indirect estimation of SAE approaches.

4.2 Crop Yield Estimation - Synthetic Method of Estimation

For estimating the crop yields at block level, a study was undertaken at IASRI in which synthetic method of estimation was applied. As described earlier, in synthetic method, the population is distributed in two dimensions with small areas on one side and post-strata (homogeneous groups) on the other side. In case of crops, formation of groups has to be based on characteristics related to crop yields. Such characteristics, in this case, are various inputs like irrigation, fertilizers etc. For application of synthetic estimation, the cell (cells of two dimensional tables) weights are needed. For crop yields, the relevant weights are area under the crop for the cell. It was realized that these weights are not available. Even the group marginal totals are not known in many situations. This was one of the major limitations for application of synthetic approach for crop yield estimation. An approach for estimation of weights under different cells was developed, using the raking ratio method. In this process, data collected in the crop-cutting approach was utilized. It may be remarked that in the process of data collection for crop- cutting approach, lot of ancillary data on various inputs for selected fields is collected. This data was used in conjunction with small area level data for crop areas for estimating the weights. It is not possible to study the performance and the effect of estimating the weights on the basis of sample data. This could be done through simulation studies. The approach was demonstrated for estimation of crop yields at block level for wheat and paddy crops on the basis of data from crop estimation surveys in Haryana State during 1987-88. The results were encouraging with respect to consistency as well as efficiency. However, they are based on certain assumptions which could not be tested and the efficiencies are based on variances which do not account for the biases. In case of failure of assumptions, biases could be serious. This has been a major limitation in the synthetic approach.

4.3 Crop Yield Estimation - An Application to Remote Sensing

The synthetic method of estimation was also used for estimation of crop yield for wheat crop at tehsil level, using remote sensing satellite data (Singh and Goel 2000). Post-strata were formed on the basis of vegetation indices derived from the remote sensing satellite data. Normalized Difference Vegetation Index (NDVI) and Ratio Vegetation Index (RVI) were used as vegetation indices. The crop data pertained to wheat crop from General Crop Estimation Surveys during 1995-96 in Rohtak district of Haryana State while the spectral data of IRS-1B LISS-II for February 17,1996 was taken for vegetation indices. The use of synthetic estimation improved the efficiency of estimators as measured in terms of standard errors. However, ignoring the bias remains a serious limitation.

4.4 Crop Yield Estimation - An Application to Crop Insurance

Crop insurance is a technique of protecting farmers in the event of crop failure due to unforeseen circumstances. In India, it has been in practice since 1985

in the form of a scheme called Comprehensive Crop Insurance Scheme (CCIS). The methodology followed in the scheme is essentially based on area unit approach as described in Dandekar (1985). In this approach, farmer is liable for compensation if there is a shortfall in the actual average yield per hectare in the area as compared to the threshold yield as obtained on the basis of normal yield. Insured farmers are required to pay a small amount as premium in return to their claim for indemnity, in case of loss to their crops due to natural calamities. In CCIS, the insurance was linked to credit system and the coverage was limited to loanee farmers and the scope was confined to crops like rice, wheat, millets, oilseeds and pulses. The area level, identified for the study, was Community Development Blocks (CDB). Crop yield estimation at the specified area unit level (in this case blocks) becomes essential for assessment of losses. The yield assessment was based on the results of General Crop Estimation Surveys. The reporting levels for crop yield estimates in GCES were the districts. For estimating the yields at block level, sample sizes for crop- cutting experiments (CCEs) were increased accordingly.

In 1999-2000, CCIS was replaced by National Agricultural Insurance Scheme (NAIS) or Rashtriya Krishi Bima Yojana (RKBY). Besides increasing the scope of crop coverage, a salient change was that area unit level was identified as Gram Panchayat (GP) level in place of CDB. This had an immediate statistical implication that average yield of the selected crops were needed at the GP level. Increasing the sample sizes for obtaining the GP level estimates was not only cost prohibitive but was likely to have serious repercussion on the quality of data.

An alternative approach was suggested for obtaining the crop yield estimates at GP level. For application of SAE techniques, availability of concomitant variates at small area level or at unit levels is necessary. It was found that meaningful information to be used as concomitant variates at lower level was not available. In the suggested approach, information on crop yields on selected fields was obtained by enquiry from farmers (farmers' appraisal), which was used judiciously for obtaining correction factors. These correction factors were then used to scale down the block level estimates to GP level. One of the risks of the approach could be that the farmers' appraisal might be influenced by the subjective assessment. However, if there is no systematic underestimation or overestimation in the entire area, the correction factors are likely to be free from such effects. The approach was akin to small area estimation, in the sense that block level estimates were scaled down to GP level. Also the advantages of farmers' appraisal (Panse *et al.* (1966)), was embedded in the approach, of course with some improvement in the timing of enquiry from the farmers. A theoretical framework along with some results are available in Sharma *et al.* (2004).

This approach, which was called Small Area Crop Estimation Methodology (SACEM) was tried on pilot basis in the districts of Ratlam (Madhya Pradesh), Thane (Maharashtra), Thiruvallur (Tamil Nadu), Muzaffarnagar (Uttar Pradesh) and Nadia (West Bengal). At present, the method is under investigation for further improvements at IASRI.

5. SOME OTHER APPLICATIONS

There has been a growing concern for the data needs at small area level in every field. There are number of large scale surveys being conducted regularly in the country. Most of these surveys are planned to provide reliable estimates at somewhat higher level. The surveys conducted by NSSO provide estimates at State level. Agricultural surveys provide crop production and yield estimates at district level. Similar situation exists in statistics relating to other fields also, such as health, education etc. Everywhere there is a need for statistics at levels lower than whatever is already available. It is also observed that lot of data are generated for official purposes which is available at different level and at different sources. Most of it is generated for specific purposes and the maintenance of such records is not very satisfactory in many cases. If somebody wants to use such data as ancillary information, its access is a major problem.

5.1 Committee on Small Area Statistics

In 1996, an expert committee on small area statistics under the chairmanship of Prof. J. Roy was set up by Department of Statistics, Ministry of Planning and Programme Implementation, Government of India. Some salient features of the report are described here.

The terms of reference of the committee were as follows:

- To analyze the data implications of the 73rd and 74th amendments to the Indian Constitution and to advise appropriate source agencies therefor.
- To examine the capabilities of existing systems to cope with emerging requirements, identify weaknesses and propose remedial measures to fill up the gaps.
- To consider various alternative methodologies for generating small area statistics and to suggest the methodology most appropriate to Indian conditions.
- To recommend the corresponding organizational and infrastructural requirements for setting up the system.
- To verify specific variables on which small area statistics ought to be generated and the periodicity with which this data should be made available.

The 73rd and 74th Amendment Acts 1992 pertain to the panchayats and municipalities respectively. The amendments provide more authority to these local area level agencies and are instruments of planning and implementation of various development plans at micro level. This naturally puts pressure on data needs for policy formulation at small area levels.

Report of the committee was submitted in April 1997. It was observed that a variety of information at the village level is collected and is available through different sources. The data needs of Panchayati Raj system, the existing data bases and the concerned source agencies, was presented in the form of a comprehensive statement. Collection and maintenance of additional data for the three tier panchayat system were also discussed.

One of the most important suggestions was regarding construction of a register of households at village level. It was recommended that a Register of Households (RH) should be maintained in each village which can serve as a basic source of information on different aspects of each household and each member of that household in a village. The modalities of construction, maintenance and updating of RH, as well as the mechanism of onward transmission of relevant information have been discussed in the report. Advantages of having these records are manifold for future planning activities.

The methodological issues relating to small area statistics were also discussed and the need for pilot studies for applying SAE techniques to some large scale surveys was highlighted. One of the great limitations in the application of SAE techniques is the non availability of related information at lower level. The directions suggested in the report would have gone a long way in addressing the problems related to the requirements of small area statistics. But to the best of my knowledge, the report is yet to receive the attention which it deserves.

6. APPLICATIONS TO NSS DATA

The National Sample Survey Organisation (NSSO) regularly conducts nation wide household surveys on various socio-economic aspects. The results of these surveys are the main source of data requirement for planning purposes. The results are reported at the State level in NSSO publications and the need for district level estimates for several parameters of interest is being realized for quite some time. Moreover, there are several domains of interest for which estimates are needed. Estimates for such parameters are developed through the usual approach of estimating domain parameters. There have been some sporadic attempts to apply SAE techniques. Singh *et al.* (2005) used NSS data for application of Spatio-Temporal Models in Small Area Estimation.

A study was sponsored by National Statistical Commission and conducted by Indian Statistical Institute in collaboration with NSSO (2000). In this study district level estimates for several parameters of interest for socio-economic variables were developed for West Bengal and Tamil Nadu. The NSS data for 51st, 52nd, 53rd and 54th rounds for Annual Consumer Expenditure and Employment and Unemployment Surveys was used. One of the limitations of the study was that the only auxiliary variable used was district population for respective years as projected based on 1971, 1981 and 1991 Censuses. The study only indicated the feasibility of application of SAE techniques for NSS data. The study was only exploratory in nature and there was a need for pursuing it further.

6.1 Application to Consumer Expenditure Survey Data

Sastry (2003) explored the feasibility of using NSS Household Consumer Expenditure Survey Data for estimation of district poverty estimates. The study was,

however, confined to examining the distribution of relative standard errors (RSE) of direct estimates for Monthly Per Capita Expenditure (MPCE) and those of the sample sizes at district level as obtained from 55th round of NSS data. The district level estimates were obtained following the usual approach of estimating domain parameters. It was observed that in rural areas, 451 out of 490 districts (92%) are having RSEs less than 5% only. It was also observed that only 2% districts had RSEs of 10% or more. It showed that district level direct estimates for MPCE were fairly reliable. The problem could exist in further sub-classifications. However, the direct estimates, in conjunction with suitable covariates, may be used for further modeling purposes for application of SAE techniques.

In the 61st round survey of NSSO (July 2004- June 2005), the quinquennial series of consumer expenditure surveys was carried out. Utilizing the household level data of this survey for Uttar Pradesh, we have tried to estimate the district level estimates of MPCE for different land holding classes as well as for all classes combined together. The land holding classes are the standard classification as follows:

Marginal	less than 1 ha.
Small	1 to 2 ha.
Semi-medium	2 to 4 ha.
Medium	4 to 10 ha.
Large	more than 10 ha.

Some of the characteristics of the data and the area under study are as follows:

No. of districts	70
No. of surveyed households (rural)	7,868
Estimated number of households	2,32,57,500
Average holding size	0.86 ha.

The methodology used for small area estimation is the basic area unit model as described in Section 3.1 as above and as followed in Fay and Herriot (1979) approach. The approach has also been used more recently in the study on Small Area Estimate of School Age Children in Poverty (2000). The approach consists of following steps :

- Obtain the direct estimates of the MPCE at the district level using the sampling design of the survey along with the estimated sampling errors.
- Obtain data from administrative records and other sources that are available for all districts to use as predictor variables.
- Specifying and estimating a regression equation that relates the predictor variables to a dependent variable, through mixed model approach with random small area effects.
- Using the estimated regression coefficient and the predictor variables to develop estimate of MPCE for all the districts.
- Using the estimated variances of the direct and model-based estimates, obtain the EBLUP estimator as described in (3.1).

Following predictor variables were used at the district level:

(i) Average holding size
(ii) Growth rate of the population
(iii) Literacy rate
(iv) Proportion of schedule caste and Schedule Tribe population
(v) Composite development index for agriculture
(vi) Composite development index for industry
(vii) Composite socio-economic development index

One of the constraints in getting the predictor variables was that they were obtained from various sources and some of the variables were not available for all the districts. As such the EBLUP estimators could not be obtained for all the districts. In fact, it could be obtained only from 56 districts. For some of the land categories this number was even smaller. Choice of predictor variables and ensuring its availability has been rather difficult.

Some of the results are presented here in the form of tables, and charts.

Table 1. Category-wise average holding size and average MPCE

Category	Average holding size (ha.)	Average MPCE (Rs.)
Marginal	0.40	547.90
Small	1.41	549.09
Semi-medium	2.71	554.09
Medium	5.46	568.90
Large	15.00	—
All	0.86	555.53

The EBLUP estimates of MPCE for different holding size are given in Table 1.

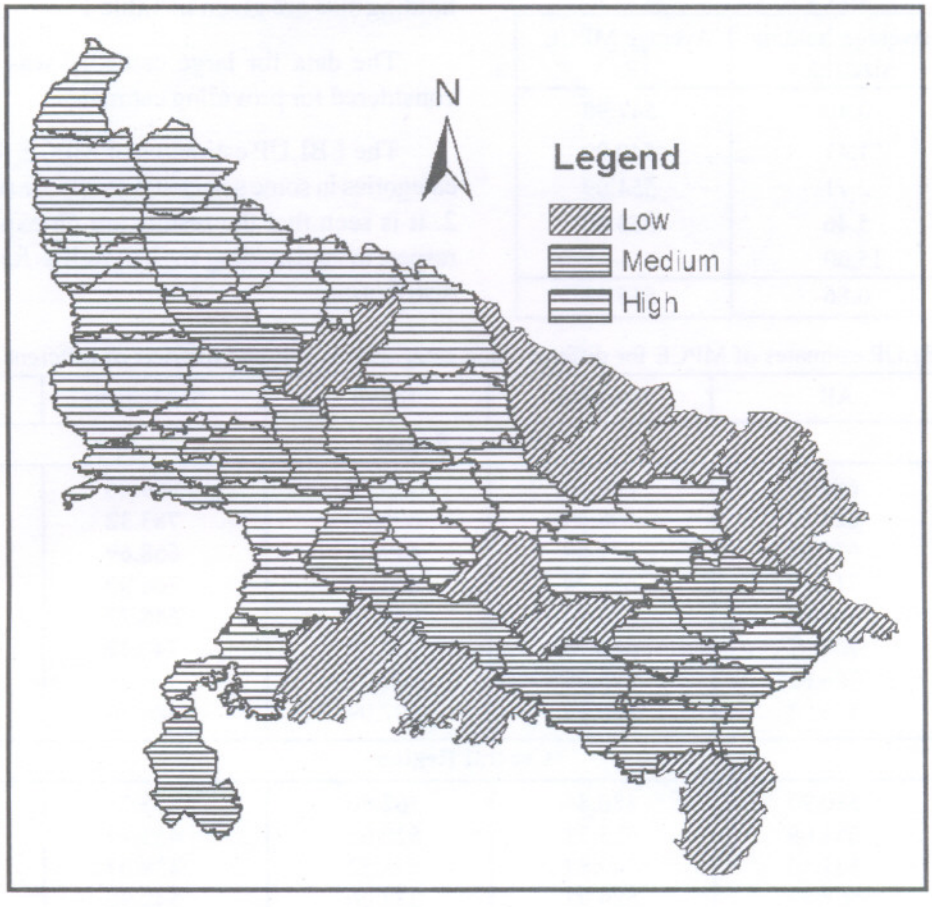
The data for large category was too small to be considered for providing estimates.

The EBLUP estimates of MPCE for different land categories in some selected districts are presented in Table 2. It is seen that the results are on expected lines with respect to land holding sizes as well as for regional/ spatial distribution.

Table 2. EBLUP estimates of MPCE for different land categories in selected districts of different regions

District	All	Marginal	Small	S. Medium	Medium
Western Region					
Saharanpur	688.57	647.55	753.66	809.26	964.44
Muzaffarnagar	621.21	598.49	696.60	783.32	716.20
Bijnor	671.94	672.89	619.46	668.69	777.51
Moradabad	713.72	688.29	644.42	706.90	747.68
Rampur	600.03	577.94	649.38	648.77	829.39
Meerut	706.92	663.76	707.60	740.27	817.87
Bulandshahar	786.62	783.06	699.77	718.36	675.65
Hathras	596.23	579.95	577.09	486.70	671.51
Central Region					
Kheri	580.90	580.86	562.66	455.71	401.90
Sitapur	641.09	625.23	520.65	453.19	496.79
Hardoi	547.55	549.87	573.52	438.07	766.34
Unnao	608.47	589.93	378.26	540.96	803.91
Lucknow	692.20	636.96	583.90	534.60	594.69
Rae Bareli	423.26	433.92	579.13	650.22	339.40
Kanpur Dehat	538.44	520.44	493.39	453.67	649.76
Kanpur Nagar	606.62	591.45	539.16	324.53	449.49
Fatehpur	554.36	563.60	610.52	552.83	490.05
Barabanki	735.31	738.99	504.90	435.54	461.61
Southern Region					
Jalaun	632.98	632.07	314.833	561.69	423.82
Jhansi	630.40	586.30	428.70	456.46	374.85
Lalitpur	543.38	522.78	500.85	523.97	504.10
Mahoba	512.38	512.41	369.88	630.24	373.24
Banda	459.99	459.44	429.18	432.38	502.83
Eastern Region					
Gorakhpur	450.28	443.07	563.09	402.56	561.65
Kushi Nagar	444.78	438.53	437.75	321.36	—
Deoria	444.78	438.53	437.75	321.36	—
Azamgarh	513.44	515.58	536.88	464.08	406.17
Mau	527.33	542.25	415.77	511.03	—
Balia	467.03	457.20	479.13	453.30	461.61
Jaunpur	555.92	550.30	554.45	—	485.87
Ghazipur	406.85	413.07	578.38	—	332.73
Varanasi	514.24	528.91	427.73	—	418.15

An overall picture for district-wise distribution of MPCE levels can be seen from the following map:



Map showing different districts of Uttar Pradesh

Following figures present the CVs of Direct and EBLUP estimators for overall data (Fig.1), marginal (Fig. 2), small (Fig. 3), and semi-medium (Fig. 4) categories. It is seen that for over all data, in most of the cases direct estimators were having CVs less than 10% and in those districts there is hardly any gain due to EBLUP estimation. However, for other categories the direct estimators are not so efficient and gains due to EBLUP estimation are substantial in many districts.

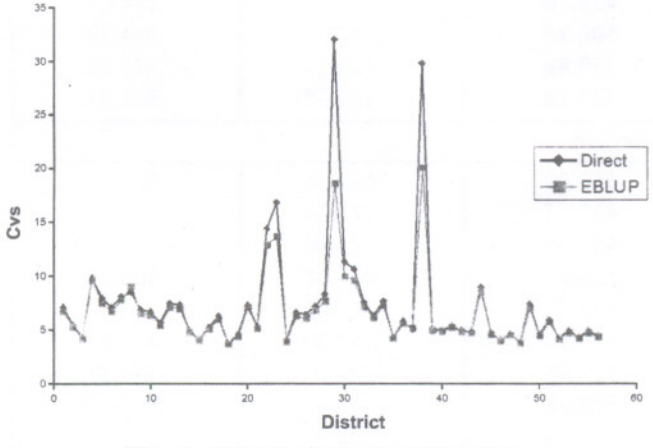


Fig. 1. CVs all categories of holding

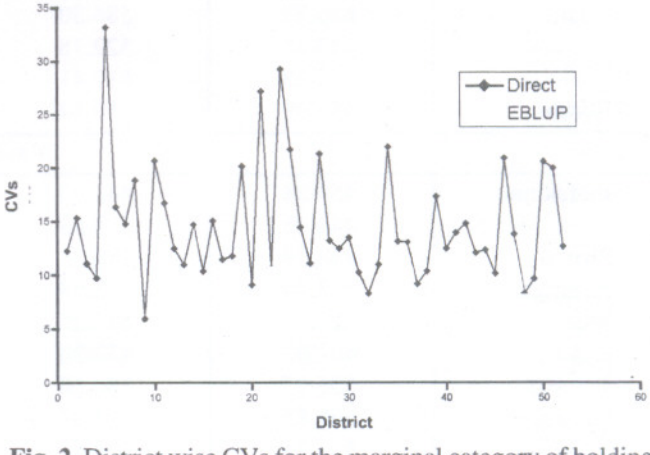


Fig. 2. District wise CVs for the marginal category of holding

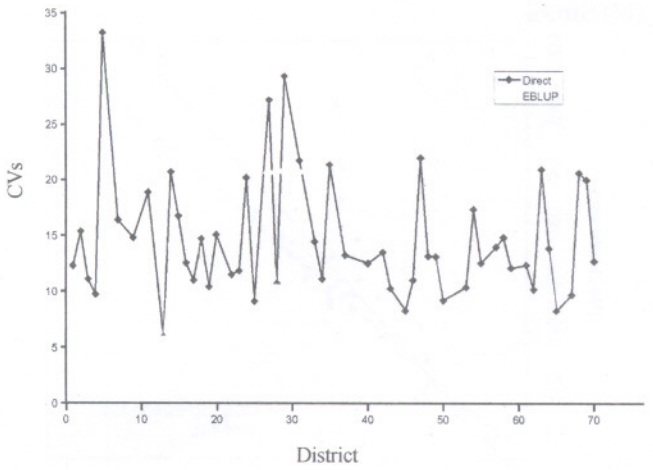


Fig. 3. District wise CVs for the small category of holdings

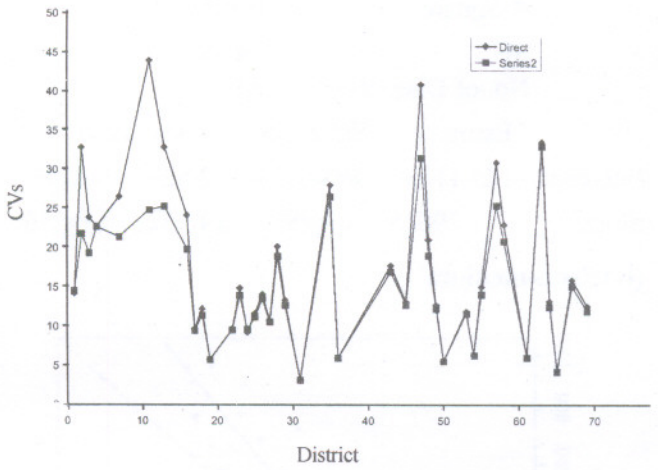


Fig. 4. District wise CVs for the medium category of holdings

SAE Diagnostics: A diagnostic study for the model fit was carried out (courtesy Chambers *et al.* (2007))

Bias Diagnostic: If direct estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If model-based estimates are close to the true values the regression of the direct estimates on the model-based estimates should be similar

- Plot direct estimates on Y-axis and model-based estimates on X-axis
- look for divergence of regression line from Y = X
- test for intercept = 0 and slope = 1

Note : Assumption that direct estimates have similar standard errors.

Goodness of Fit Diagnostic : Model-based estimates should have the same expectations as corresponding direct estimates and should be uncorrelated with them.

Calculate

$$W = \sum_d \left\{ \frac{\left(\text{Direct estimate}_d - \text{Model - based estimate}_d \right)^2}{\hat{\text{Var}}(\text{Direct estimate}_d) + \hat{\text{MSE}}(\text{Model - based estimate}_d)} \right\}$$

and compare with the chi square distribution on D degrees of freedom.

Assumptions

- Unbiased estimators of variance, MSE available
- CLT behaviour for both model-based and direct estimates

Coverage Diagnostic

Suppose $Y \sim N(\theta, \sigma_Y^2)$ and $X \sim N(\theta, \sigma_X^2)$, with Y and X uncorrelated, and $k = (\sigma_Y + \sigma_X)^{-1}\sqrt{\sigma_Y^2 + \sigma_X^2}$.

Then $Y \pm 2k\sigma_Y$ and $X \pm 2k\sigma_X$ should overlap approximately 95 % of the time.

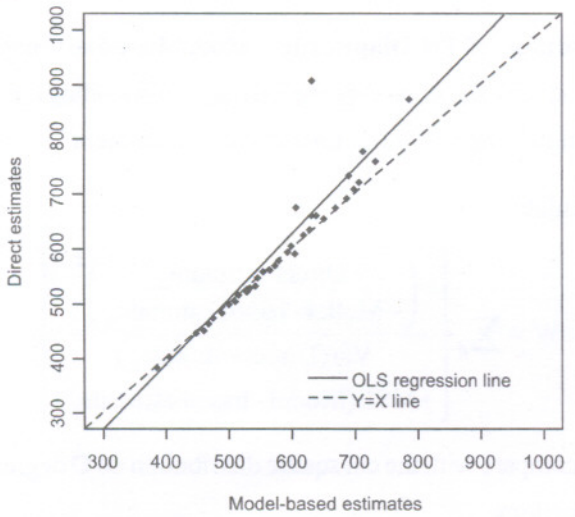
- Form adjusted 95% confidence intervals for small area means based on direct and model-based estimates using the critical values

$$2\frac{\sqrt{\hat{\text{Var}}(\text{direct}) + \hat{\text{MSE}}(\text{model})}}{\sqrt{\hat{\text{Var}}(\text{direct})} + \sqrt{\hat{\text{MSE}}(\text{model})}}$$

- Count the number of times the intervals do not overlap - should be approximately 5%.

Bias Diagnostic

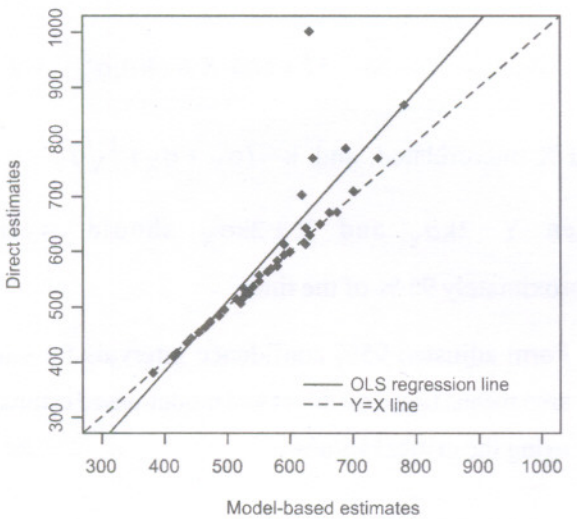
(i) All



R Square	0.6499
SE	81.11
No. of Districts	56

	Estimate	Std Error	t Ratio	Prob > | t |
Intercept	–83.4615	66.7460	–1.25	0.217
Slope	1.1870	0.1186	10.01	6.55e-14

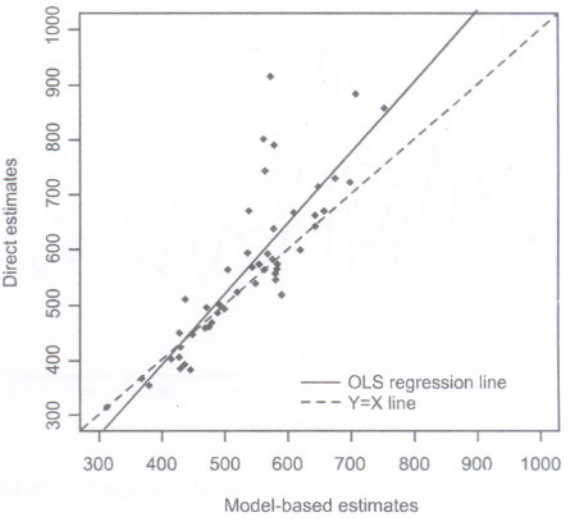
(ii) Marginal



R Square	0.6009
SE	90.94
No. of Districts	54

	Estimate	Std Error	t Ratio	Prob > | t |
Intercept	–127.7015	78.6038	–1.625	.11
Slope	1.2779	0.1417	9.016	2.34e-12

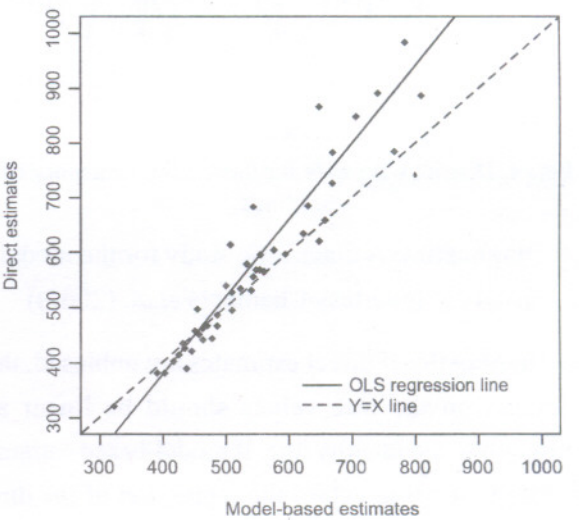
(iii) Small



R Square	0.8059
SE	74.48
No. of Districts	53

	Estimate	Std Error	t Ratio	Prob >| t |
Intercept	–125.11335	49.46334	–2.529	0.0146
Slope	1.29028	0.08867	14.552	2.0e-16

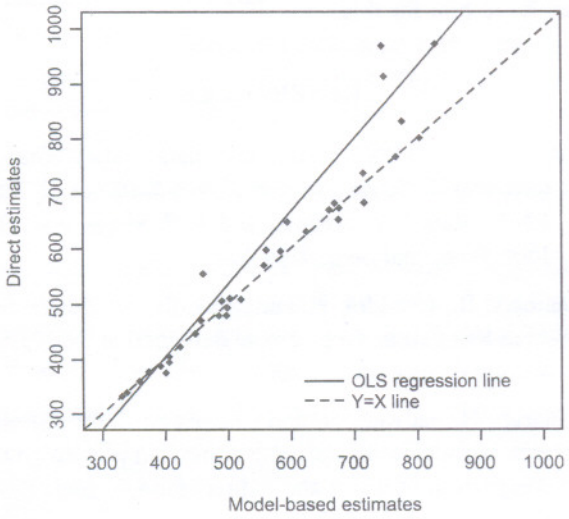
(iv) Semi-medium



R Square	0.8773
SE	61.66
No. of Districts	50

	Estimate	Std Error	t Ratio	Prob > | t |
Intercept	–190.28678	41.90320	–4.541	3.77e-05
Slope	–1.41659	–0.07645	–18.529	2.0e-16

(v) Medium



R Square	0.5357
SE	199.8
No. of Districts	41

	Estimate	Std Error	t Ratio	Prob > | t |
Intercept	–128.503	116.885	–1.099	0.278
Slope	–1.328	–0.198	–6.708	–5.41e-08

Goodness of fit diagnostics

The fit was good for all the categories. The calculated values were smaller than 61.66.

Coverage diagnostics

For all the categories in less than 5% cases the intervals did not overlap.

6.2 Application of SAE to Debt and Investment Surveys

Similar to the above application, district level estimates for the variable – amount of loan outstanding were obtained using data from 59th round of NSS, conducted during the period January to December, 2003 for the State of Uttar Pradesh (Srivastava *et al.* (2006)). The covariates used were

(i) Loan disbursed (in lakhs) kharif

(ii) Loan disbursed (in lakhs) rabi

(iii) Population density

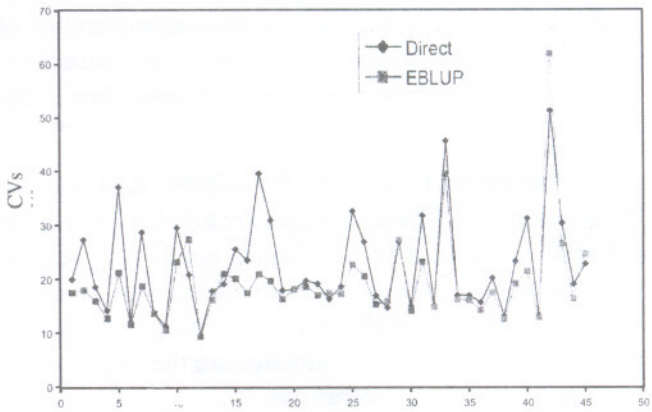
(iv) Rural population SC

(v) Rural population ST

(vi) Percentage irrigated area

We present here only the comparative picture of the efficiencies of EBLUP estimates vs those of direct estimates through following chart.

District-wise CVs for the Direct and EBLUP estimators
(Average amount of loan outstanding)



It is seen that for this characteristic the CVs of direct estimator are generally more than 10 per cent and there are definite gains due to application of EBLUP estimator.

7. FUTURE SCENARIO - SOME CONSIDERATIONS

In the planning process of any country the initial requirements normally pertain to estimates of macro-level parameters. However, with the growth in the development process, requirement of statistics at lower level become more and more important. Small area statistics has become a practical necessity in almost every field of application as far as the data needs are concerned.

There has been phenomenal growth in the theoretical aspects of SAE techniques. Model based estimation has been a turning point in the growth of SAE methods. Depending on the type of data availability, several types of models and corresponding procedures are available. In recent years applications of these methods are also being made in variety of situations. There are instances of large scale studies and corresponding applications such as the study on Small- Area estimates of School- Age Children in Poverty conducted in USA.

In India there have been sporadic attempts for applications. But either they have been based on traditional biased approaches like synthetic method of estimation or the model based applications have been only demonstrative in nature. Data availability, as far as conduct of survey is concerned, is in abundance. There are number of large scale surveys, aimed at macro level estimates. There is enough data which could be used as covariates in different situations. But access to such data is a problem. With computerization facilities, improvements are taking place. With computerization of data from various censuses, access to several covariates at district level and sometimes at even lower levels has become simpler.

Although SAE methods are available, the situations for application of the techniques are also identified, but the application of SAS techniques needs very serious preparatory work. Choice of suitable variates, ensuring their availability, developing and testing the models, validating the small area estimates are the steps, which need very careful considerations.

There are also some cautions needed in application of SAS methods. In this context, as a conclusion a remark by Kalton (1987) seems to be appropriate – I consider that a cautious approach should be adopted to the use of small area estimates and especially to their population by Government Statistical Agencies. When Government Statistical Agencies do produce model dependent small area estimates, they need to distinguish them clearly from conventional sample based estimates. Some small area estimates may be seriously in error and errors in small area may be more apparent to users than errors in national estimates. Before small area estimates can be considered fully credible, carefully conducted evaluation studies are needed to check on the adequacy of the model being used. Sometimes model dependent small area estimators turn out to be of superior quality to sample based estimators and this may make them seem attractive. However, the proper criterion for assessing their quality is whether they are sufficiently accurate for the purpose for which they are to be used.

ACKNOWLEDGEMENTS

I am thankful to ISAS for giving me this opportunity to offer my tributes to Dr. Panse. In the preparation of this paper, Dr. U.C. Sud has helped me a lot in data analysis. I am thankful to him for this.

REFERENCES

Brackstone, G.J. (1987). Small area data: Policy issues and technical challenges. In: *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh, eds., 3-20, John Wiley and Sons, New York.

Chambers, R., Chandra, H. and Tzavidis, N. (2007). Small Area Estimation. Course note delivered at the ICES-III, Montreal, June 18-21, 2007.

Chandrati, H., Salvati, N. and Chambers, R. (2007). Small area estimation for spatially correlated populations – A comparison of direct and indirect models. *Stats. Trans.*, New Series, **8(2)**, 887-906.

Chaudhury, Arijit *et al.* (2000). Report on Small Area Estimation of Socio Economic Variables, November 2000. A study by Indian Statistical Institute in Collaboration with National Sample Survey Organization Sponsored by National Statistical Commission.

Constance, F. Citro and Graham Kalton, Editors (2000). *Small-Area Estimates of School-Age Children in Poverty*. National Academies Press.

CSO (1997). Report of the Expert Committee on Small Area Statistics (Chairman - Prof. J. Roy).

Dandekar, V.M. (1985). Crop insurance in India – A review, 1976-77 to 1984-85. Economic and Political Weekly, *Rev. Agric.*, June 1985.

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.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test,* **15(1)**, 1-96.

Narain, Prem, Rai, S.C. and Shanti Sarup (1995). Regional disparities in the levels of development in Uttar Pradesh. *J. Ind. Soc. Agril. Statist.,* **47(3)**, 288-304.

NSSO Report No. 508 (2006). Level and Pattern of Consumer Expenditure NSS 61st Round. (July 2004-June 2005).

Panse, V.G., Rajagopalan, M. and Pillai, S.S. (1966). Estimation of crop yield for small areas. *Biometrics*, **22(2)**.

Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P. (1987). *Small Area Statistics*. Wiley, New York.

Purcell, N.J. and Kish, L. (1979). Estmates for small domain. *Biometrics*, **35**, 365-384.

Rao, J.N.K. (2002). Small area estimation with applications to agriculture. *Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR)*, 555-564.

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

Sastry, N.S. (2003). District level poverty estimates – Feasibility of using NSS household consumer Expenditure survey data. *Economic and Political Weekly*, January 25, 2003.

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.*, **57**, 26-37.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183-195.

Singh, D. (1968). Double sampling and its application in agriculture. *J. Ind. Soc. Agril. Statist. (Panse Memorial Volume)*.

Srivastava, A.K., Ahuja, D.L., Mathur, D.C. and Sethi, S.C. (1999). Project Report on Estimation of Crop Yields for Small Areas. IASRI publication.

Srivastava, A.K., Sud, U.C. and Chandra, H. (2006). Small area estimation: Some applications. Paper presented in the International Conference on Statistics and Informatics in Agricultural Research organized in the Diamond Jubilee of Indian Society of Agricultural Statistics.

Sud, U.C., Srivastava, A.K., Bathla, H.V.L., Mathur, D.C. and Jha, G.K. (2001). Project Report on Crop Yield Estimation at Block Level using Farmer Estimates. IASRI publication.