Jour. Ind. Soc. Ag. Statistics
Vol. XL, No. 1 (1988), pp. 60—69

SYMPOSIUM* ON SMALL AREA STATISTICS

CHAIRMAN : DR. M. N. MURTHY, *Honorary Director, Applied Statistics Centre, Madras*

CONVENERS : DR. O. P. KATHURIA, *Head (SSM and ASD), IASRI, New Delhi*
DR. A. K. SRIVASTAVA, *Scientist S-3, IASRI, New Delhi*

A symposium on Small Area Statistics was organised on 17th December, 1987 under the Chairmanship of Dr. M. N. Murthy, Honorary Director, Applied Statistics Centre, Madras and former Director, Asian Statistical Insitute, Tokyo.

Welcoming the Chairman Professor Prem Narain, Secretary, Indian Society of Agricultural Statistics and Director, IASRI referred to the contributions made by Dr. M. N. Murthy in the theory of sample surveys and its applications at the Indian Statistical Institute, Calcutta and later as the Director of the Asian Statistical Institute, Tokyo. It was, therefore, most befitting to have Dr. Murthy to chair the Symposium on this important topic of small area statistics.

In his opening remarks, the Chairman pointed out the need for statistics at disaggregated levels for planning of development programmes and for assessing achievements. He stated that the data available on auxiliary characters in the officical records or from a previous census or survey can provide a useful base for deriving estimates at (say) block or tehsil level. He cited the example of Japan where a good infrastructure base has been developed to provide statistics at disaggregated levels using grids of 1 sq km area as basic statistical units.

Five papers were presented at the symposium dealing mostly with review of various methods of small area statistics and a few approaches were attempted in the Indian context.

*Organised on 17th December, 1987 during the 41st Annual Conference of the Society at C. T. R. I., Rajahmundry (A.P).

The following recommendations emerged out of the deliberations at the symposium :

1. The requirement of statistics at disaggregated levels should be kept in view while planning large scale data collection efforts.

2. Relevance of the concept of statistical mapping as a technique for providing small area statistics should be explored. Use of remote sensing technique in this context may be an added advantage.

3. Theoretical investigations are needed in various types of errors such as sampling, non-sampling and model errors in the study of various small area estimation techniques. The robustness of the techniques also needs to be examined.

4. Documentation of empirical investigations and practical experience in applying the techniques of small area statistics estimation to real-life problems will be most useful.

Prof. Narain thanked the Chairman for conducting the symposium and the convenors for their help in organising the symposium.

Detailed summaries of papers presented at the symposium are as follows :

1. Estimation for Small Domains—Application in Demography

K. S. NATARAJAN
Office of the Registrar General of India, New Delhi

In recent years demands are increasing for production of timely statistics for administrative units like district and tehsil in various fields like health and family welfare. For example, a large number of new schemes for maternal and child health care as also family planning have been introduced recently. These programmes have set goals at national level. However, to monitor the achievement very little statistics are available at district and lower levels. Similar demands for data on employment and unemployment are also raised from time to time.

While at national and state level, sample surveys have been able to meet the growing demands to a great extent, such surveys have not been able to release useable estimates for smaller administrative units. In recent years a number of techniques have been developed to provide estimates for small domains. The paper reviews the methods available in the Demographic literature. These methods can be applied to obtain small area statistics in the field of population provided reliable statistics on civil registration and other administrative statistics are available at

district and lower level. Unfortunately, in India these statistics are not available.

The methodology adopted in other countries has been applied to one union territory Goa, Daman and Diu where civil registration is good. It is found that estimates derived appear to be reasonable. Considering the available data it has been pointed out that annual enrolment figures, data on ration card holders and data on number of married females in the re-productive age group together with improved civil registration would be the statistics required in the field of population for application of methods developed to estimate population at small area level. Unfortunately, statistics on these items are either deficient or not available on timely basis.

2. Small Area Estimation

T. J. RAO
Indian Statistical Institute, Calcutta

In large scale sample surveys, the main problem is estimation of parameters of interest for bigger areas or larger subgroups. However, if it is required to obtain estimates for small areas using the general methodology, the resulting conclusions are usually found to be not very satisfactory due to difficulties in identifying the sample units belonging to the small areas and the errors involved in building up the estimates from the sample data. Usually, the governments are interested in obtaining estimates for smaller geographical or administrative divisions as well as for individuals belonging to a certain group classification.

The methods of small area estimation can be classified into thrre types. First, the 'ratio based methods' include the simple and intutive approach called 'synthetic estimation' which uses the sample survey data for a larger area to obtain estimates for small area having the same characteristics as the large area. Here the estimates may become badly biased if there is a departure from these assumptions. The 'regression based methods' such as 'symptomatic regression' and 'sample regression' have emerged from the oldest 'symptomatic accounting technique.' These methods are found to be suitable for reasonable relationships between study and symptomatic variables, stable growth over a period and good auxiliary data. Finally, the 'other methods' include the 'base unit method' which consists of dividing the survey frame into constituent geographic units called 'base units' at a level lower than the small area of interest and then building up an estimate, the 'structure preserving estimate' and estimates based on the 'predictive approach',

From the reviews and appraisals of these methods it can be seen that there is no single approach that can be best used in all situations. If for a population, a technique that is most suitable is decided depending on the assumptions involved, data available and structure etc., then we suggest the following appreach : First obtain the small area estimate $\hat{Y}_s$ for small area s based on the technique decided. Next for the area b at a level just bigger than the small area for which exact information (say census) is available, obtain an estimate, say $\hat{Y}_b$ as if the area b itself is a 'small' area using information at a level higher than b. Then, use the known information on Y_b to correct the estimate $\hat{Y}_s$ by a ratio or regression adjustment. If Y_b is not known, data on a related variable X_b may be used. This technique can be fruitfully applied along with synthetic method. One question that is of importance in large scale sample surveys is the determination of sample size so as to get stable estimates for small areas such as percentage of workers in single digit industrial division or proportion of rural unemployed etc.

The small area classification depends on the type of the sampling frame chosen. In certain cases, it may be proper to define the classification in terms of the total of a known auxiliary variable highly related to the study variable, as for example a 'small' domain may be defined to comprise of a sub-total between two well-defined fractions of the population total of the auxiliary variable. Thus efforts should be made to obtain a frame for the units of the small areas of interest and the time and money spent is well worth in view of the non-availability of a proper logical methodology for small area estimation. With the advent of super computer networks and with the computer awareness among the people and governments in several countries, it is perhaps not too difficult to have the computerised dissemination of small area statistics. Durbin, recently, points out that computerised small area statistics from the 1981 U.K. Census giving details of social and economic data for enumeration districts of the size of about 150 households each are available from the Office of the Population Census and Surveys (OPCS). Also, the Manpower service commission's on-line system and the Post Code Address File in the U.K. have continuous update on the local area statistics. We finally remark that the problem of small area estimation should be further explored *vis-a vis* the use of computers for dissemination of small area statistics.

3. Woodruff's Technique of Estimating Small Area Statistics

S. G. PRABHU-AJGAONKAR
Marathwada University, Aurangabad 4310004

A sample is selected to estimate characteristics pertaining to a population from which it has been drawn. There exists auxiliary information. On many occasions estimators are requireed for a part of population or for population units. Hansen, Hurwitz and Madow (Ref : Sample Survey Methods and Theory, vol I, 1953) were the first to consider this problem in connection with radio listeners programme sample survey. They formulated a regression estimator and employed it to predict values for the unknown population units.

A simple random sample of n primary units is drawn without replacement from a population consisting of N primary units. From the ith sampled primary units, a further sample of secondary units is selected and an unbiased eastimator $\hat{Y}_i$, of Y_i, the total of Y-characteristic for the ith primary unit, is obtained. The object is to estimate T_y, the population total for the Y-characteristic. X's is an auxiliary characteristic, information about which is available for each sampling unit. If $\bar{y}$ and $\bar{x}$ are sample means based on n primary units for the Y and X characteristics respectively, then the best linear unbiased estimator of the population total, T_y is given by

$$Y = N[\bar{y} + \beta(\bar{X} - \bar{x})]$$

where β is the regression coefficient and $\bar{X}$ is the population mean for the X-characteristic.

Woodruff's Method

Let $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ be the values of Y-characteristic corresponding to n primary population units that are included in the sample. Woodruff (Ref : Use of regression technique to produce area breakdowns of the monthly National Estimates of Retail Trade, *Jour. American Stat. Asso.* **61**, 1966) suggested that if the ith population unit is included in the sample, then an estimator of Y_i, of the Y-characteristic for the ith primary unit, is

$$\bar{Y}_{(i)} = \frac{N}{n}\left[\hat{Y}_{(i)} - \beta X_{(i)}\left(1 - \frac{n}{N}\right)\right]$$

If the jth population unit is not included in the sample, then an estimator of $\bar{Y}_j$, the population total of the Y-characteristic for the jth primary unit, is

$$\bar{Y}_j = \beta X_j.$$

The variance of Y is estimated as follows:

$$S_y^2 = \frac{n}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2$$

where $z_i = Y_i + \beta(\bar{X} - x_i)$

and $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$.

The variance for a part of population is estimated from this expression giving z_i a value zero if the ith population unit is not included in the sample.

4 Estimation for Small Areas Under Crop Insurance

P. NARAIN AND SHIVTAR SINGH
Indian Agricultural Statistics Research Institute, New Delhi 110012

Introduction

The Comprehensive Crop Insurance Scheme launched from Kharif 1985 covers wheat, paddy, millets, oilseeds and pulses. The time series data on average annual yield of a crop in an area, generated from crop estimation surveys conducted by the State Governments are being utilised for working out the guaranteed yield and the actual average yield for settlement of indemnity claim in the area. The Scheme operates on homogeneous area basis. The area may be a taluk or a block or a mandal or even a Panchayat for which precise estimates of average yield per hectare are required. It is common knowledge that the Crop Estimation Surveys were designed to provide reliable estimates of production/productivity at the aggregated administrative levels like the state or districts. However, for Crop Insurance, reliable estimates at the disaggregated levels for small domains viz. block/mandal/panchayat are required to make the Scheme attractive and beneficial to the farmers. This calls for either increasing the number of crop cutting experiments or looking into the existing small domain estimation techniques for their suitability to provide estimates at the disaggregated levels. The first alternative appears to be prohibitive in cost (monetary as well as technical manpower). The "base unit method" proposed by Kalsbeek (1973) as reported by Purcell and Kish in *Biometrics*, 35, 1979 under small domain estimation techniques appears to be an appropriate alternative. But, this requires the

availability of symptomatic variable(s) at the base unit level. In this paper, some symptomatic variables which can be used for the purpose of obtaining estimates for small domains are suggested.

The Base Unit Method

This method proposed by Kalsbeek (1973) consists in splitting up the small domains (blocks/mandals) into smaller base units (villages), then classifying each of these units into one of 'k' groups of base units for which estimates can be obtained from a sample survey. The small domain estimates are then formed by taking weighted combinations of these base estimates.

The survey frame is divided into constituent geographic units (villages) which are termed base units. Survey data are needed for a sample of these base units. On the basis of the auxiliary information and the information on the variable of interest for these base units, they are grouped into 'k' homogeneous groups with a suitable clustering algorithm. The local areas of interest (blocks/mandals) are also broken down into constituent base units, and each unit is then classified by use of auxiliary information available, into one of 'k' groups. The estimation procedure is then as follows.

An estimate for each of the 'k' groups of base units is formed by taking a weighted average of the estimates for the sample base units that constitute each group. That is,

$$\bar{x}_g = \sum_{i=1}^{n_g} w_{gi}\, \bar{x}_{gi} \qquad g = 1, 2 \ldots, k$$

where

n_g is the number of sample base units in the gth group of base units,

w_{gi} is the weight,

$\bar{x}_{gi}$ is the sample estimate of the mean (yield/ha) for the ith base unit in the gth group,

$\bar{x}_g$ is the weighted average (mean) for the gth group of base units.

The final estimate $\bar{x}_h$ for the hth small domain (block) is given by

$$\bar{x}_h = \sum_{i=1}^{k} O_{gh}\, \bar{x}_g$$

where O_{gh} represents the composition of the kth groups in the hth block.

Basis of Grouping the Base Untis

In the temporary settled states, the villagewise information on percentage of area irrigated, size of holding of the individual farmers, area under high yielding varieties, number of bovines, number of persons, number of bullocks/tractors, etc., is available. Similar information for the permanently settled states on sample basis is also available. The above information can be used to club the base units (villages) into groups, for which estimates on sample basis can be obtained.

5. On Some Methods of Small Area Estimation

O. P. KATHURIA
Indian Agricultural Statistics Research Institute, New Delhi-110012

Introduction

The data collected in a sample survey, besides providing estimates at the aggregated levels, may also provide estimates of the population parameters at disaggregated levels for different segments or domains. While in the selection of sample it may not be practically feasible to give adequate representation to each segment of the population, estimates for different segments or domains may become necessary subsequently from planning point of view. Data available on auxiliary characters from previous census or survey may prove useful in developing estimates for the main character for segments or domains. However, development of a suitable estimation technique for small areas is called for.

Broad Areas of Application of the Technique

The concept of small area estimation is not new. Panse and others (Ref: Estimation of crop yields for small areas; *Biometrics*, 1966) examined the feasibility of using double sampling for estimation of yield at the block level. The technique consisted in selecting a large sample of villages and fields from the block for eye estimation of yield prior to the harvest and combining it with results of crop cutting experiments conducted on a sub-sample for obtaining estimates of average yield at the block level. The technique envisaged presence of a strong positive correlation between eye estimates and the actual yield harvested. However, this approach could not succeed due to various physical constraints. Other applications of small area estimation are in demography, health and labour force. Comprehensive reviews of various methods of small-area estimation have been made by Purcell and Kish (Ref : Estimation for small

domains; *Biometrics*, 35, 1979; and Post censal estimates for local areas or domains; *Int. Stat. Review*, 48, 1980) and Srivastava and Singh (Ref : Small area estimation—A review; IASRI Symposium, 1987).

In the crop sector, various types of small area statistics that may be needed are estimates of average yield at block/tehsil level, estimates of average yield of irrigated and unirrigated areas of major high yielding varieties of crops etc. Similarly, in the livestock sector where sample surveys are being conducted for estimation of major livestock products at the state level, statistics may be needed for estimation of the products at the district or lower administrative level. Another example may be the need for intercensal estimates of different species of livestock.

Small Area Estimation Using Double Sampling

Consider a situation where a population of N fields is available growing (say) wheat crop of which N_i fields are growing ith variety ($i = 1, 2, \ldots, k$). We wish to estimate the average area $\bar{X}_{N_i}$ and the average yield $\hat{Y}_{N_i}$.

Select a preliminary sample of n' fields out of N with SRSWOR on which only the character X, i.e. area alone is measured. A sub-sample of n units is selected out of n' on which yield y is recorded. Let n'_i and n_i be the number of units out of n' and n respectively which belong to the ith sub-area. Obviously n'_i and n_i are both random variables. Let $\bar{y}_{n'_i}$ $\hat{y}_{n'_i}$ and $\bar{x}_{n'_i}$, $\bar{x}_{n_i}$ be the sample means of Y and X based on n'_i and n_i, units respectively. We assume that n' and n are sufficiently large such that each sub-area is represented in the sample. It is easy to show that the sample estimators $\bar{x}_{n'_i}$ and $\bar{x}_{n_i}$ are both unbiased estimators of X_{N_i}.

Estimation of $\bar{Y}_{N_i}$: Two possible estimators, using auxiliary character X are : (i) Regression estimator and (ii) Ratio estimator.

The Regression estimator $\bar{y}_{di}$ for the ith sub area may be written as

$$\bar{y}_{di} = \bar{y}_{ni} + b_i (\bar{x}_{n'_i} - \bar{x}_{ni})$$

where b_i is the least squares regression coefficient of y on x in the ith sub area computed on n_i units.

The ratio estimator $\bar{y}_{Rdi}$ may similarly be written as

$$\bar{y}_{Rdi} = \frac{\bar{y}_{ni}}{\bar{x}_{ni}} \bar{x}_{n'_i}$$

$\bar{y}_{di}$ and $\bar{y}_{Rdi}$ are both biased estimators. However, to the first order of approximation the bias in both the estimators becomes negligible.

The variance of $\bar{y}_{di}$ is given by

$$V(\bar{y}_{di}) = \frac{N}{N_i}\left[\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n} - \frac{1}{n'}\right)\rho_i^2\right] S_{y_i}^2$$

where ρ_i is the population correlation coefficient between x and y and $S_{y_i}^2$ is the mean sum of squares of y in the ith sub-area.

Similarly, for the ratio estimator $\bar{y}_{rdi}$'

$$V(y_{Rdi}) = \bar{Y}_{N_i}^2 \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{N}{N_i}\left[\frac{S_{xi}^2}{\bar{X}_{N_i}^2} - \frac{2S_{xyi}}{\bar{y}_{N_i}\bar{X}_{N_i}}\right] + S_{yi}^2 \frac{(N-n)}{N_i n}$$

Data Requirement for Small Area Statistics (SAS) in Agriculture

The applicability of various SAS methods depends upon the availability of related data sources. The availability of a sound infrastructure and an established data base system in India makes it possible to obtain small area statistics in many sectors of the economy. The various publications on human, agricultural and livestock census provide detailed information of these populations at different levels. Similarly, the village as a unit is a source of a variety of information. The village records provide a fairly detailed information about the constitution of the population, land utilisation statistics and the livestock statistics. The various types of regular surveys conducted by the Central and State Governmnts also provide reliable estimates of parameters at higher aggregated level.It is, therefore, possible to apply many of the small area estimation techniques to obtain estimates at various disaggregated levels. However, small area parameters are likely to be subject to large variation and may, therefore, not be stable over time. Some small area estimates may turn out to be seriously in error and errors in small area estimates may be more apparent than errors in aggregated estimates. It will, therefore, be worthwhile to examine the robustness of small area estimation techniques.