J. Ind. Soc. Agril. Statist.
60(2), 2006 : 134–143

Spatial Sampling Procedures for Agricultural Surveys using Geographical Information System

Prachi Misra Sahoo, Randhir Singh and Anil Rai
Indian Agricultural Statistics Research Institute, New Delhi
(Received : June, 2006)

SUMMARY

Traditional sampling recommends sampling designs like simple random sampling, stratified sampling and systematic sampling for spatial data which is dependent in nature and exhibits spatial correlation. But these sampling designs when applied to spatial data fail to capture the dependency present in the data raising questions about estimator's efficiency. Thus, if we have a clue of the spatial correlation structure underlying the spatial phenomenon to be sampled, it is desirable to exploit this information in the sampling design. Therefore, there is a need to modify the classical sampling designs to take into account the dependency (spatial correlation) present in the spatial data while selecting sampling units. In this paper, new spatial sampling procedures have been proposed exploiting the functionalities of Geographical Information System (GIS) for handling spatial data. The performance of the proposed spatial sampling procedures have been evaluated and it has been found that these procedures provide more efficient and reliable estimators for the spatial data as compared to those obtained from traditional sampling techniques.

Key words : Spatial data, Spatial sampling, Spatial autocorrelation, Geographical Information System.

1. INTRODUCTION

In agricultural surveys, often the parameter of interest is geographical in nature *i.e.* it pertains to specific location. This type of data termed as spatial data, involves the simultaneous consideration of both locational and attributes information (Goodchild 1986). Spatial data are not independent in nature unlike usual data. This was noticed long ago in thirties when Stephan (1934) observed that the data of geographic units are tied together like bunches of grapes and not separate like balls in an urn so that they cannot be thought of as randomly generated from the classical urn model. It indicates that the generating process underlying the spatial data is constituted by a sequence of random variables, $X_1, X_2, ..., X_n$, which are no more i.i.d. and whose pattern of dependency can be studied.

Sampling design for spatial data have a long tradition starting from Mahalanobis (1940), Quenouille (1949) and Das (1950). Hansen *et al.* (1960), Kish (1965), Mailing (1989), Webster and Oliver (1990) recommended traditional sampling designs like simple random sampling, stratified sampling and systematic sampling for spatial data. However, these designs do not consider the spatial parameter while assigning the probability of selection for the sampling units of the target population. The existence of dependence in the spatial data violates the basic assumption of independence of the traditional sampling procedures. Classical statistical methods when applied to spatial data fail to capture the dependency present in the data raising questions about estimator's sufficiency, bias, efficiency and consistency (Zhang and Griffith 2000). Thus, there is a need to modify the classical sampling designs to take into account the dependency (spatial correlation) present in the spatial data while selecting sampling units.

Hedayat *et al.* (1988) suggested Balanced Sampling design Excluding Contiguous unit technique (BSEC) which excluded the selection of contiguous units in the sample. Extending this idea, Arbia (1990, 1991) proposed a sample selection method named as Dependent Unit

Sequential Technique (DUST), where the spatial correlation based on auxiliary character has been used to assign the probability of selection to each unit in the population. He showed empirically that this technique led approximately to 30 percent gain in efficiency with respect to simple random sampling (Arabia 1993).

Thus, the distinctive feature of spatial data being spatially correlated is now fully recognized as a problem at the stage of data analysis and statistical modeling but almost systematically neglected in the context of survey. This is because of the difficulties faced by the researchers in handling voluminous spatial data. But the recent technological developments in computers have shifted the emphasis of survey research towards newer emerging areas. A major development has been that of integrated software for the capture, storage, analysis and display of spatial information in the form of Geographical Information System (GIS). The new advances in GIS functionalities capable of handling various types of information through their geographic coordinates have facilitated easy processing and mapping of spatial information which has made it a potential tool to change substantially the statistical approaches to the study of geographic reality particularly for survey design.

In view of the above, this study has been taken up with an aim to develop reliable and statistically sound spatial sampling procedures capable of efficiently handling the complexity of correlation structure present in the spatial data by employing GIS technology. These procedures not only give sampling designs for spatial data but also provide efficient and reliable estimates for estimating population means and totals.

2. SPATIAL SAMPLING PROCEDURES

According to Tobler's first law of geography, nearby units are more related as compared to the units, which are far apart (Tobler 1979). Therefore, if a particular unit is selected in the sample, the neighbouring unit is not likely to provide any more relevant additional information about the target population. If such a sampling design could be developed which avoids the selection of such neighbouring units, the duplication of information partly contained in areas already sampled would be reduced to greater extent. This would not only ensure better representation of target population, but also provide efficient estimation procedure under the classical randomization framework. In this way, it economizes the cost without losing the reliability of the estimates. Thus, if we have a clue of the spatial correlation structure underlying the spatial phenomenon to be sampled, it is desirable to exploit this information in the sampling design.

The proposed sampling procedures are GIS based procedures in which the dependency in the spatial data is estimated in the form of spatial correlation using GIS. Further, a method for spatial stratification has been proposed in order to form strata which are spatially homogeneous. The method of sample selection is characterized by variable inclusion probabilities at each draw. The principle underlying the allocation of selection probabilities is that the units which are far apart from the earlier selected units are given higher probability of selection as compared to the units which are nearer. Further, since the size of the units also play major role in case of spatial data, size measure is also used for calculating the probability of selection. Finally, suitable unbiased estimators are proposed. The proposed sampling procedure in detail follows.

2.1 Estimation of Spatial Correlation

The dependency in the spatial data is estimated by spatial correlation. Cliff and Ord (1973) has defined spatial correlation as: "Given a group of mutually exclusive units or individuals in a two dimensional plane, if the presence, absence or degree of a certain characteristic affects the presence, absence or degree of the same characteristic in neighbouring units, then the phenomenon is said to exhibit spatial correlation". Spatial correlation tests whether or not the observed value of a variable at one locality is independent of values of that variable at neighbouring localities. A positive spatial correlation refers to a map pattern where geographic features of similar value tend to form a cluster on a map, whereas a negative spatial correlation indicates a map pattern in which geographic units of similar values scatter throughout the map. When no statistically significant spatial correlation exists, the pattern of spatial distribution is considered to be random.

The classical measure of spatial correlation is given by Moran (1950). If X_i and X_j are the values of X at i^{th} and j^{th} locations, respectively then the spatial correlation β for N observations is given by

$$\beta = \frac{N}{C}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \quad (1)$$

where

$C = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$ $(i \neq j)$, w_{ij} are the weights such that $w_{ij} = 1$, if i and j are neighbours and 0 otherwise.

2.2 Spatial Stratification

Spatial stratification means formation of strata in such a manner that each stratum consists of units which are spatially homogenous. This is done by testing the stationarity of spatial correlation. For this purpose a Monte Carlo significance test given by Brunsdon *et al.* (1998) is applied. A suitable test of significance is used for testing p value. Each time one unit is added to the group and tested for homogeneity. If it is found to be homogenous, it is included in the group otherwise dropped. In this way, all the units are tested one by one in order to form strata which are spatially homogenous.

2.3 Sample Selection and Estimation Procedures

Consider a population of N areal units. Let Y be the character under study and X be the auxiliary character, with known values for each unit. The first lag spatial correlation is denoted by β. In this study, four different sampling schemes and method of estimating population mean has been proposed. These are as described below.

2.3.1 Contiguous Unit Based Spatial Sampling (CUBSS)

Let $\Omega(y_1, y_2, ..., y_N)$ be the set of all the units contained in the population. Let $y_1, y_2, ..., y_n$ be the values of the units drawn at the first, second,... ,n^{th} draw respectively. Let $s_1^*, s_2^*, ..., s_n^*$ denote the sample set which contains the units selected from the N units after first, second, ..., n draws respectively, such that

$s_1^* = \{y_1\}, s_2^* = \{y_1, y_2\}, ..., s_n^* = \{y_1, y_2, y_3, ..., y_n\}$

Let $\alpha_1, \alpha_2, ..., \alpha_n$ be the probabilities of selection of $y_1, y_2, ..., y_n$ respectively.

Selection of first unit in the sample

The first unit is selected by simple random sampling. Clearly, the probability of selecting i^{th} unit at the first draw will be 1/N i.e. $\alpha_{i_1} = \alpha_1 = 1/N$ $\forall\ i = 1, 2, ..., N$ where α_{i_1} is the probability of selecting i^{th} unit in the first draw. It can be seen that the sum of the probabilities at the 1^{st} draw will be unity.

Selection of second unit in the sample

Second unit from remaining N-1 units is selected using following steps

Step 1 : Select a random number from 1 to N-1 (say i).

Step 2 : Select another number at random from 1 to M (say r), where M is the maximum value of the auxiliary character.

Step 3 : Select the unit i if $(r \leq U_{i_2} X_i)$ where U_{i_2} is given as

$U_{i_2} = (1 - \beta^{d_{12}})$

Here, i_2 indicates that i^{th} unit has been selected in the second draw, X_i is the size measure of the i^{th} unit and $d_{12} = 1, 2, 3, ...$ if 1^{st} and 2^{nd} selected units are $1^{st}, 2^{nd}, 3^{rd}$ ... lag neighbours.

Step 4 : Reject unit i and repeat the above process if $(r > U_{i_2} X_i)$.

Let i^{th} unit be selected in the 2^{nd} draw, whereas j^{th} unit is already selected at first draw. Then the probability of selecting i^{th} unit in second draw is

$$\alpha_{i_2} = \alpha_2 = \left(\frac{U_i X_i}{\sum_{\substack{i \subset \Omega \\ i \not\subset s_1^*}}^{N} U_i X_i}\right)$$

where s_1^* is the set of earlier selected units.

Clearly, the sum of the probabilities at the second draw is also unity.

Selection of Subsequent Units

For selecting a sample of size n, the above procedure is repeated till n units are selected with U_{i_i}'s changing at each draw after selection of each unit. The general term U_{i_n} for n^{th} draw is given as

$$U_{i_n} = (1-\beta^{d_{1n}})(1-\beta^{d_{2n}})....\left(1-\beta^{d_{(n-1)n}}\right) \quad (2)$$

Thus, for the case of n^{th} draw

$$\alpha_{n(n\neq 1)} = \left(\frac{U_n X_n}{\sum\limits_{\substack{i\subset\Omega \\ i\not\subset s^*_{n-1}}}^{N} U_i X_i} \right) \quad (3)$$

Clearly, the sum of the probabilities at the n^{th} draw is also unity. It can thus be observed that the probability of selection varies at each draw and sum of probabilities in each draw remains unity.

Estimation Procedure

Let T_1 be the estimator of population mean.

Define

$$d_n = \frac{1}{N}\left[y_1 + y_2 + y_3 + ... + \frac{y_n}{\alpha_n} \right] \quad (4)$$

where α_n is defined in (3)

Then, T_1 is given by

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} d_i \quad (5)$$

It can easily be seen that T_1 is an unbiased estimator of population mean. The form of estimator T_1 is similar to the ordered estimator given by Des Raj (1956 a, b). Thus, on similar lines, the estimate of variance of the estimator T_1 which is also an ordered estimator can be expressed as

$$\hat{V}(T_1) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(d_i - T_1)^2 \quad (6)$$

2.3.2 Stratified contiguous unit based spatial sampling (Stratified CUBSS)

In this sampling design, the method of sample selection described in case of CUBSS is applied in each stratum. The spatial stratification is applied to form the strata. Stratification can also be done on the basis of geographical or administrative boundaries such as districts in the state or tehsils in the district.

Let Ω_h be the set of all the units in the h^{th} stratum. Let $y_{1h}, y_{2h}, ..., y_{n_h h}$ be the values of the unit drawn at the first, second... n^{th} draw respectively from h^{th} stratum. Let L be the total number of strata. Let $s^*_{1h}, s^*_{2h}, ..., s^*_{nh}$ denote the sample set which contains the units selected from N_h units after first, second, ..., n draws in each stratum respectively such that

$$s^*_{1h} = \{y_{1h}\}, s^*_{2h} = \{y_{1h}, y_{2h}\}, ...,$$

$$s^*_{nh} = \{y_{1h}, y_{2h}, y_{3h}, ..., y_{nh}\}$$

Let $\alpha_{1h}, \alpha_{2h}, ..., \alpha_{nh}$ be the probabilities of selection of $y_{1h}, y_{2h}, ..., y_{nh}$ in the h^{th} stratum respectively.

Selection of first unit in the sample

The first unit in each stratum is selected by simple random sampling. Clearly, the probability of selecting i^{th} unit at the first draw in h^{th} stratum will be $1/N_h$.

$$\alpha_{ih_1} = \alpha_{1h} = 1/N_h \qquad \forall \quad i = 1, 2, ..., N$$

where α_{ih_1} is the probability of selecting i^{th} unit in the first draw in the h^{th} stratum.

Selection of subsequent units in the sample

Second unit from remaining N_h-1 units is selected from each stratum using following steps

Step 1 : Select a random number from 1 to N_h-1 (say i).

Step 2 : Select another number at random from 1 to M_h (say r), where M_h is the maximum value of the auxiliary character in the h^{th} stratum.

Step 3 : Select the unit i if $\left(r \le U_{ih_2} X_{ih}\right)$

where $U_{ih_2} = (1-\beta^{d_{12h}})$ and X_{ih} be the size measure of the i^{th} unit in the h^{th} stratum.

Here, ih_2 denotes that i^{th} unit is selected in the second draw in the h^{th} stratum and

$d_{12h} = 1, 2, 3, \ldots$ if 1st and 2nd selected units in the h-th stratum are 1st, 2nd, 3rd, ... lag neighbour.

Step 4 : Reject the unit i and repeat the above process if $(r > U_{ih_2} X_{ih})$.

It can be easily seen that the sum of the probabilities at the second draw is unity in each stratum. For selecting a sample of size n, the above procedure is repeated till n_h units are selected with U_{i_1}'s changing after selection of each unit in each stratum. The U_{ih_n} is given as

$$U_{ih_n} = (1-\beta^{d_{1n}})(1-\beta^{d_{2n}})\ldots\left(1-\beta^{d_{(n-1)n}}\right)$$

Thus, for the case of n_h^{th} draw

$$\alpha_{nh_n} = \left(\frac{U_{ih}X_{ih}}{\sum\limits_{\substack{i \subset \Omega_h \\ i \not\subset s_{(n-1)h}}}^{N} U_{ih}X_{ih}} \right) \tag{7}$$

It can thus be observed that the probability of selection varies at each draw and sum of probabilities after every draw remains unity in each stratum.

Estimation Procedure

An appropriate estimator for the population mean is obtained by suitably combining the stratum wise estimators of the character under study. Let T_2 be the estimator of the population mean obtained by applying stratified CUBSS. Let us define

$$d_{jh} = \frac{1}{N_h}\left[y_{1h} + y_{2h} + \cdots + y_{(n_h-1)h} + \frac{y_{nh_h}}{\alpha_{nh_h}} \right] \tag{8}$$

where α_{nh_h} is given by (7) and

$$\bar{y}_h = \sum_{j=1}^{n_h} \frac{d_{jh}}{n_h} \tag{9}$$

is the sample mean of h-th stratum, then T_2 is given by

$$T_2 = \sum_{h=1}^{L} W_h \bar{y}_h \tag{10}$$

Obviously, T_2 is an unbiased estimator of $\overline{Y}$. Further, according to equation (6), an estimate of variance of T_2 can easily be written as

$$\hat{V}(T_2) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \left[\sum_{j=1}^{n_h} (d_{jh} - \bar{y}_h)^2 \right] \tag{11}$$

2.3.3 Modified contiguous unit based spatial sampling (MCUBSS)

It is well known that probability proportional to size (pps) sampling is more efficient than simple random sampling (srs). Thus, it is expected that by applying pps sampling in the selection of first unit, the performance of CUBSS and Stratified CUBSS may be improved. Keeping this in view, CUBSS and Stratified CUBSS are modified for selection of first unit in the sample.

The first unit is selected by the method of probability proportional to size sampling. It is known that the probability of selecting any unit by pps is given by X_i/X, such that $X = \sum_{i=1}^{N} X_i$. Clearly, the probability of selecting i-th unit at the first draw will be $\alpha_{i_1} = \alpha_1 = X_i/X \;\; \forall \;\; i = 1, 2, \ldots, N$ where α_{i_1} is the probability of selecting i-th unit in the first draw. The sum of the probabilities at the 1st draw will be

$$\sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \frac{X_i}{X} = 1$$

The second and subsequent units are drawn in a similar manner as mentioned in Section 2.3.1 for CUBSS.

Estimation Procedure

Let the unbiased estimator of population mean obtained by modified CUBSS technique be denoted by T_3. Alike T_1 the estimator T_3 can be given as

$$T_3 = \sum_{i}^{n} d_i \Big/ n \tag{12}$$

where d_n and α_n are given be equation (4) and (3) respectively.

The estimate of variance of the estimator T_3 will also be similar to that of T_1 as given in equation (6) on similar lines of T_1 is given by

$$\hat{V}(T_3) = \frac{1}{n(n-1)} \sum_{i}^{n} (d_i - T_3)^2 \tag{13}$$

Same sampling technique can now be applied in each strata which gives rise to stratified modified CUBSS.

2.3.4 Stratified modified CUBSS (StMCUBSS)

The population is divided into homogenous strata on the basis of spatial correlation or the administrative boundaries are considered as strata. Sample is then selected from each stratum using modified CUBSS technique. In this method, the first unit in the sample is selected by probability proportional to size in each stratum. The second and subsequent units are drawn in a similar manner as mentioned in Section 2.3.2 for Stratified CUBSS. All the notations are similar to the one mentioned in Section 2.3.2. The unbiased estimator of population mean denoted by T_4 is given as

$$T_4 = \sum_{h=1}^{L} W_h \bar{y}_h$$

where $\bar{y}_h$ is given by equations (9).

To obtain the sampling variance of T_4, it may be noted that sampling is carried out independently in each stratum. Hence, an estimate of variance of T_4 can be written as

$$\hat{V}(T_4) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \left[\sum_{j=1}^{n_h} (d_{jh} - \bar{y}_h)^2 \right]$$

3. EMPIRICAL STUDY FOR IMPLEMENTATION OF SPATIAL SAMPLING PROCEDURES USING GIS

In order to evaluate the performance of the proposed sampling procedures and to compare their efficiencies with the traditional sampling techniques, a simulation study has been carried out in Rohtak district of Haryana to estimate the irrigated area in the district. The irrigated area (Y) which has to be estimated has been treated as the character under study and the total cultivated area (X) which is highly correlated with the character under study has been taken as the auxiliary character for the study. The major steps of implementing spatial sampling procedures using GIS are shown in Fig. 3 and are described as below.

3.1 Data Preparation

Two types of data were required for the study viz. spatial and attribute data. These were procured from District Handbook of Census (DHC) of Rohtak of the year 1991. The spatial data were procured from DHC in the form of five tehsil maps having village boundaries. These hard copy maps have been converted into digital coordinates by the process of digitization using SUMRGRID V digitizer, configured with GIS software PC ARC/INFO. The digitization process has been carried out using ARCEDIT module of the software. The digitized maps of five tehsils have been transformed to a common scale and merged/ mosaiced to obtain a single map of the district having all the village boundaries. The digitized map with district and village boundaries is shown in Fig. 2. The smallest unit in the map is village which is represented as polygon. Each village or polygon

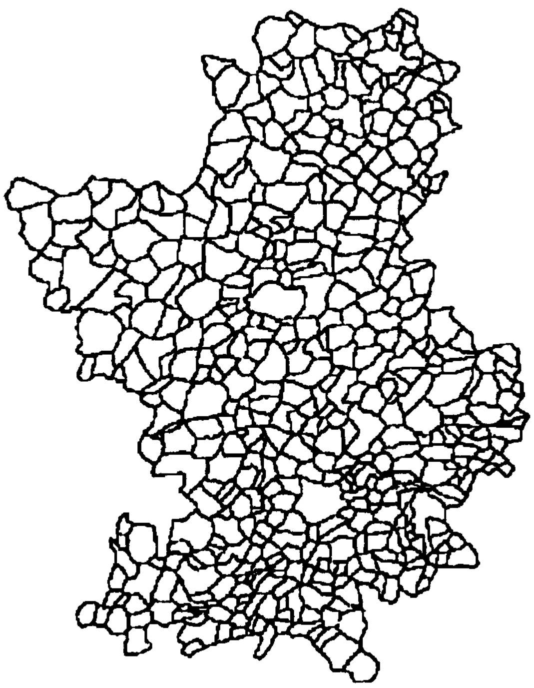

Fig. 2. Digitized map of Rohtak district

carries a unique identification number (ID number). The district consists of 492 villages and thus we have 492 polygons in the map. Two data files in Date Base File (DBF) format namely Arc Attribute Table (AAT) and Polygon Attribute Table (PAT) were automatically generated during the process of digitization. AAT file contains the spatial data and PAT file contains the information regarding this spatial data. The attribute data includes the data of irrigated area and cultivated area for each village. The data for these variables is stored for each village along with its unique ID number in a separate database format file (DBF). The attribute data is linked with the spatial data through this unique ID number using the PAT file in TABLE module of the software PC ARC/INFO. In this way a complete database having both spatial data in the form of maps and attribute information regarding each village has been developed using GIS.

3.2 Development of Algorithms

Number of algorithms have been developed in computer programming language FORTRAN 90. These are

- Algorithm for identification of neighbours
- Algorithm for calculation of spatial correlation
- Algorithm for spatial stratification
- Algorithm for sample selection for various sampling designs like CUBSS, StCUBSS, MCUBSS, StMCUBSS, SRSWR, SRSWOR, StSRSWOR, DUST and StDUST
- Algorithms for obtaining estimators of the above said sampling designs

AAT and PAT files obtained through GIS has been used as the input files for implementing these algorithms.

3.3 Identification of Neighbours

The AAT file having spatial information regarding each polygon has been used as the input file for identification of neighbours. The total number of neighbours and the neighbour of each polygon is identified for all 492 polygon using the algorithm for identification of neighbours. This information is stored in a separate file, which has been used for calculating the spatial correlation. The first, second and third lag neighbours are shown in Fig. 3.

3.4 Calculation of Spatial Correlation and Spatial Stratification

The spatial correlation has been computed using the expression given in equation (1). The value of the spatial correlation for each polygon is obtained and overall spatial correlation for entire map is also

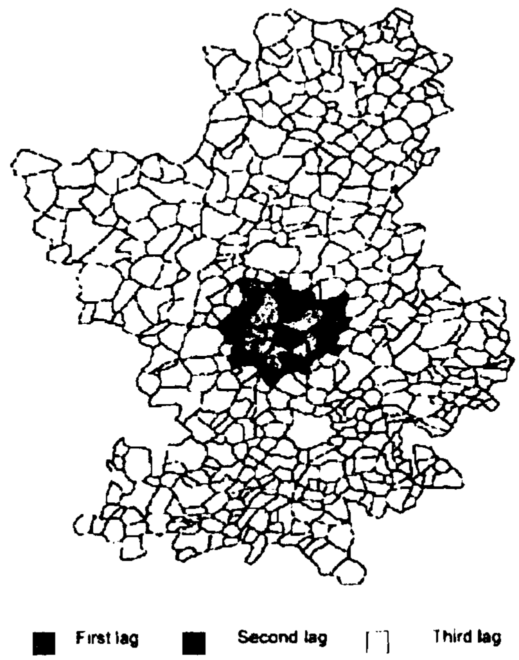

Fig. 3. Contiguity based neighbours at different lags

calculated. The value of overall spatial correlation comes out to be 0.41.

Spatial stratification is done according to the procedure mentioned in Section 2.2. The PAT file and the file containing the spatial correlation of each village have been used for formation of strata. Since the data is highly spatially correlated the entire map came out to be one single stratum.

3.5 Sample Selection and Estimation

One thousand samples of different sample sizes 30, 50, 75, 100 has been selected using the proposed sampling procedures explained in Section 2.3. The sample selection of the conventional sampling procedures have been done using standard methods.

Besides the proposed estimators T_1, T_2, T_3 and T_4, corresponding to the sampling techniques CUBSS, Stratified CUBSS, Modified CUBSS and Stratified modified CUBSS respectively, mentioned in Section 2.3, the estimators of traditional sampling techniques have also been considered for comparing the efficiency of the proposed estimators. These estimators are (i) Simple Random Sampling with Replacement (T_5), (ii) Simple Random Sampling Without Replacement (T_6), (iii) Stratified SRSWOR (T_7), (iv) Dependent Unit

Table 1. The estimator of population means, its variance and estimate of variance for different sampling techniques

Sampling Technique	Estimator	Variance	Estimate of Variance
SRSWR	$T_5 = \sum_i^n y_i / n$	$V(T_5) = \frac{\sigma^2}{n}$	$\hat{V}(T_5) = \frac{s^2}{n}$
SRSWOR	$T_6 = \sum_i^n y_i / n$	$V(T_6) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$	$\hat{V}(T_6) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2$
Stratified SRSWOR	$T_7 = \sum_{h=1}^{L} W_h \bar{y}_h$	$V(T_7) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 W_h^2$	$\hat{V}(T_7) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2 W_h^2$
DUST	$T_8 = \sum_i^n y_i / n$	$V(T_8) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$	$\hat{V}(T_8) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2$
Stratified DUST	$T_9 = \sum_{h=1}^{L} W_h \bar{y}_h$	$V(T_9) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 W_h^2$	$\hat{V}(T_9) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2 W_h^2$

Sequential Technique (DUST) T_8 (Arbia, 1993) and (v) Stratified DUST (T_9). In case of Stratified CUBSS, Modified Stratified CUBSS, Stratified SRSWOR and Stratified DUST, the strata are formed on the basis of administrative boundaries i.e. the five tehsils in the district have been considered as strata.

The estimator of population mean, its variance and estimate of variance for these sampling techniques (T_5, T_6, T_7, T_8 and T_9) are given in Table 1.

3.6 Criteria for Comparison of Different Estimators

To compare the performance of the proposed spatial sampling scheme with the various existing sampling schemes, the percentage relative bias (RB), relative efficiency (RE) as compared to the estimator based on SRSWOR and coefficient of variation (CV) have been calculated using the following formulae.

Percent Relative Bias : $RB = (T_i - \bar{Y})/\bar{Y} \times 100$ where T_i is the sample mean for i-th estimator; $i = 1,2,...9$ and $\bar{Y}$ is the population mean.

Relative Efficiency : RE for the estimator T_i ($i = 1, 2, ...9$) as compared to the estimator T_6 is given by $RE = V(T_6)/V(T_i)$ where $V(T_6)$ is the variance of the estimator (T_6) based on SRSWOR and $V(T_i)$ is the variance of the estimator T_i, for $i = 1, 2, 3, 4, 5, 7, 8, 9$.

Coefficient of Variation : $CV = \left(\sqrt{V(T_i)}/T_i\right) \times 100$

4. RESULTS AND DISCUSSION

Different measures for comparing the performance of various estimators are given in Table 2. Following points has been observed from the results of the study

- The results clearly indicate that the percent relative bias is very low and ranges from 0.003 to 0.74 for various estimators, which suggest that all the estimators are almost unbiased.
- Considering relative efficiency, it is observed that there is remarkable gain in efficiency for all the proposed estimators as compared to the traditional estimators. The obvious reason for this is that in case of traditional sampling techniques like Simple Random Sampling with Replacement, Simple Random Sampling without Replacement and Stratified SRSWOR the spatial dependence present in the data in not taken into account. In case of DUST and Stratified DUST though spatial dependence is accounted for at the stage of sample selection but the size measure in not included which is also an important variable in case of spatial data. Further, in case of T_8, Arbia (1993) used the

Table 2. The percent relative bias, relative efficiency and coefficient of variation for different estimators for different sample sizes

Sample size	Estimators	Percent relative bias	Relative efficiency	Coefficient of variation
n = 30	T_1	0.59	4.24	6.37
	T_2	0.19	6.26	5.32
	T_3	0.09	8.38	4.75
	T_4	0.14	11.81	4.70
	T_5	0.05	0.88	14.25
	T_6	0.28	-	13.38
	T_7	0.30	1.19	12.21
	T_8	0.74	0.88	14.13
	T_9	0.54	0.90	10.73
n = 50	T_1	0.17	5.75	4.40
	T_2	0.01	7.34	3.94
	T_3	0.09	9.02	3.62
	T_4	0.12	11.29	3.54
	T_5	0.68	0.84	11.07
	T_6	0.19	-	10.66
	T_7	0.62	1.33	9.19
	T_8	0.73	1.04	10.39
	T_9	0.50	1.05	7.85
n = 75	T_1	0.42	5.72	3.36
	T_2	0.11	6.56	3.18
	T_3	0.23	7.42	3.03
	T_4	0.13	11.35	2.81
	T_5	0.15	0.83	8.96
	T_6	0.26	-	8.19
	T_7	0.05	1.35	7.02
	T_8	0.23	0.95	8.37
	T_9	0.18	1.15	5.79
n = 100	T_1	0.05	7.40	2.84
	T_2	0.003	8.61	2.70
	T_3	0.09	8.79	2.63
	T_4	0.05	11.63	2.28
	T_5	0.00	0.98	7.63
	T_6	0.11	-	7.27
	T_7	0.03	1.68	5.95
	T_8	0.16	1.29	6.79
	T_9	0.01	1.51	6.29

estimator of simple random sampling whereas in case of proposed sampling in case of DUST procedures an unbiased estimator is developed in which the component of spatial correlation is included in the probability of selection.

- Further, among the proposed estimators, clearly T_4 is most efficient estimator followed by T_3, T_2 and T_1 respectively. T_4 is the estimator obtained using Stratified CUBSS i.e. first unit is selected by pps sampling and stratification is also being done. It has also been observed that T_3 is more efficient than T_1 and T_2 due to the fact that T_3 is based on probability proportional to size sampling whereas in case of T_1 and T_2 first unit is selected by simple random sampling. Further, between T_1 and T_2, T_2 is better than T_1 as stratification is done in case of T_1. Thus, it is clearly observed that all the results obtained are in accordance with the principles of sampling theory.
- Taking into consideration the stability aspect it has been observed that the coefficient of variation of the proposed estimators (T_1, T_2, T_3 and T_4) ranges from 2.28 to 6.37 whereas in case of existing estimators (T_5, T_6, T_7, T_8 and T_9) the range is 5.95 to 14.25. This indicates that the difference in the range of coefficient of variation of the proposed estimators and that of existing estimators is very high. The proposed estimators are highly stable as compared to the existing estimators.
- The study has been conducted for four different sample sizes i.e. 30, 50, 75,100. From the results it has been observed that there is not much gain in efficiency by increasing the sample size to 75 and 100. Thus, the sample size of 50 i.e. a sampling fraction of about 10% will provide quite efficient estimates.

5. CONCLUSION

In this study, new spatial sampling procedures for spatial data have been proposed. These procedures have been developed using the potential of GIS in handling spatial data. The results of the study pointed out that by using the proposed GIS based spatial sampling procedure considerable gain in efficiency of the estimators could be achieved. The proposed spatial sampling procedures provide more efficient, stable and reliable estimates as compared to the traditional sampling procedures. In fact, by using the proposed spatial sampling procedures a better allocation of resources could be achieved leading to higher levels of accuracy of the estimates for a given sample size, or, alternatively to a smaller sample size at a constant level of accuracy.

REFERENCES

Arbia, G. (1990). Sampling dependent spatial units. *Workshop on Spatial Statistics*. Commission on Mathematical Modeling of the ICU, Boston.

Arbia, G. (1991). GIS–based sampling design procedures. *Proc. of EGIS*, 1991, **1**, EGIS Foundation Utrecht.

Arbia, G. (1993). The use of GIS in spatial statistical surveys. *Internat. Statist. Rev.*, **61(2)**, 339-359.

Brunsdon, C., Fotheringham, S. and Martin, C. (1998). Geographically weighted regression - modeling spatial non-stationarity. *The Statistician*, **47(3)**, 431-443.

Cliff, A.D. and Ord, J.K. (1973). Spatial autocorrelation. *Pion*, London.

Das, A.C. (1950). Two dimensional systematic sampling and the associated stratified and random sampling. *Sankhya*, **10**, 95-108.

Goodchild, M.F. (1986). *Spatial Autocorrelation*. CATMOG, (Norwich : GeoBooks).

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1960). *Sample Survey Methods and Theory*. Wiley, New York.

Hedayat, A.S., Rao, C.R. and Stufken, J. (1988a). Sampling plans excluding contiguous units. *J. Statist. Plann. Inf.*, **19**, 159-170.

Hedayat, A.S., Rao, C.R. and Stufken, J. (1988b). *Designs in Survey Sampling avoiding Contiguous Units*. In : Handbook of Statistics P.R. Krishnaiah and C.R. Rao Eds., **6**, Elsevier Science Publishers B.V., 575-583.

Kish, L. (1965). *Survey Sampling*. Wiley, New York.

Mahalanobis, P.C. (1940). A report on the Bihar crop survey. 1943-44. *Sankhya*, **7**, 29-118.

Mailing, D.H. (1989). *Measurement from Maps*. Pergamon Press.

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17.

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.*, **20**, 355-375.

Raj, D. (1956a). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Assoc.*, **51**, 269-284.

Raj, D. (1956b). A note on the determination of optimum probabilities in sampling without replacement. *Sankhya*, **17**, 197-200.

Stephan, F.F. (1934). Sampling errors and the interpretation of social data ordered in time and space. *J. Amer. Statist. Assoc.*, **29**, 165-166.

Tobler, W.R. (1979). *Cellular Geography*. In : Gale S. Olsson G. (eds) Philosophy in Geography. Dordrecht, Reidel 379-386.

Webster, R. and Oliver, M.A. (1990). *Statistical Methods in Soil and Land Resources Survey*. Oxford Univ. Press.

Zhang, Z. and Griffith, D.A. (2000). Integrating GIS components and spatial statistical analysis in DBMSs. *Int. Jour. G.I.S.*, **14(6)**, 543-566.