International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 7, No. 1 (May-2020)

http://dx.doi.org/10.12785/ijcts/070101

Simulation-Cum-Regression (SICURE) Method of Estimation for the Small Domains

G.C. Tikkiwal[1], Piyush Kant Rai[2] and Teena Goyal[3]

[1]*Department of Mathematics and Statistics, J.N.V.U, Jodhpur, Rajasthan, India*
[2]*Department of Statistics, Banaras Hindu University, B.H.U., Varanasi, India*
[3]*Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan, India*

Received Oct 13, 2018, Revised August, 2019, Accepted February 6, 2020 2020, Published May 1, 2020

Abstract: When we talk about small domain problems, the sample regression method is extensively used to build up Small Area Statistics. It is brought out that this method is a special case of the Theory of Successive Sampling. This paper defines and discusses Simulation-Cum-Regression (SICURE) model approach to meet the scarcity of small area statistics. We also demonstrated the gain in efficiency due to this method and the details of its application to obtain the average yield per hectare of the crops for small areas. Also, we compared the performances of these estimators in terms of absolute relative bias and simulated relative standard errors which are computed by a simulation study of crop yield simulated data at Tehsil levels considered as the small domain.

Keywords: Regression Estimator, Small Area Estimation, SICURE Model, Synthetic Estimator, Successive Sampling, Absolute Relative Bias, Simulated Relative Standard Error, Simulation

1. INTRODUCTION

The modeling for small area estimation has been extensively discussed in a volume edited by Platek, Rao, Särndal and Singh (1987). This volume contains all the invited papers presented at the International Symposium on Small Area Statistics (SAS) held at Ottawa in May 1985. The extensive coverage in this volume of the work done in SAS-field is very enlightening. This volume contains three broad categories of models: (i) Structure-preserving estimation (SPREE) models (ii) Synthetic models and (iii) Regression models. The first two models are non-stochastic in nature, while the third one is generally stochastic and makes use of the method of least squares. Schaible et al (1979) and Heeringa (1981) have given a case study where synthetic estimators perform badly. Feeney (1987) suggested that SPREE estimators, not much in general can be applied. However if we can obtain sufficient auxiliary information through secondary or primary sources along with the characters under study the regression estimators are easy to construct, they are unique and are statistically amenable.

In regression estimators for small domains, we use survey data collected generally through a stratified multi-stage design. An early example of this is that of Erickson (1974). For estimating population changes in local areas, he used a two-stage sampling design, the PSU's (primary sampling units) being counties or groups of counties in the United States. While regressing on auxiliary variables, whether this regression should be at the primary stage level or at a lower level is a matter one should decide in advance. Apparently, this point has not been touched in the literature on small area statistics (SAS). Light on this is thrown by discussing some results in multi-stage sampling on successive occasions, Tikkiwal (1980, 1982) and Tikkiwal, and Gupta (1991). Further, it is seen that the early

E-mail address: gctikkiwal@yahoo.com, raipiyush5@gmail.com, teenagoyalnewai@gmail.com

seminal results of Erickson (1974) in the said example are covered by the first situation dealing with two-stage successive sampling with replacement of PSU's only.

The regression models are an extension of the regression and double sampling procedures considered by Watson (1937) and Tikkiwal (1960) in survey sampling while estimating variances of estimators of parameters of small domains, it is necessary to keep in mind that the survey population is essentially finite. We have to obtain the necessary formulae for the same. This is done under a normal model using concepts like 'unbiased in an extended sense' by Tikkiwal (1960), a term synonym with the term 'model unbiased ness' given by Brewer (1963) and Royall (1970). Such concepts also provide consistency criteria for stochastic models. This criterion in addition to other validation criteria has been provided by McCullagh and Zidek (1987), Sec. 2, pp. 64-66; and Cronkhite (1987), Sec. 7, pp.160-174, reported in the said theory volume by Platek et al. (1987). The agricultural crop estimation surveys in India and various developing countries have been carried out from the 1940s or so, it has been discussed by Mahalanobis (1946) and Sukhatme and Agarwal, (1946-47, 47-48). Such surveys provide somewhat reliable estimates at the state level but not necessarily at district and panchayat Samiti levels. This is demonstrated by data from Rajasthan, one of the states in India.

2. SIMULATION-CUM-REGRESSION (SICURE) MODELING

We know that the picture of the availability of reliable statistics varies from country to country. Generally, it is poor in developing countries as compared to that in developed countries. For our purpose, we assume that some data is available or can be made available with some efforts for the small areas. But these data are not enough to provide stable coefficients in the estimating equation to be used for providing reliable statistics for the small areas. This difficulty can be overcome by simulating enough more data through the analysis of available data under an appropriate model and then using the estimating equation. The estimating equation is generally a regression equation. Therefore, we refer to such modeling as Simulation-Cum-Regression (SICURE) modeling. We now proceed to demonstrate how SICURE modeling can be used in a given situation by considering Agricultural Crop Statistics in different countries.

3. SICURE-MODELING FOR CROP ESTIMATION IN SMALL AREAS

The crop Estimation surveys are being conducted in India and other developing countries through the survey methodology developed mostly in the 1940s described by Mahalanobis (1946), Sukhatme and Agarwal (1946-47, 47-48). Earlier there was some pioneering work is done by Hubback (1927) in this area, to which a reference was made by Prof. R.A. Fisher in his memorandum dated 2nd March 1945 addressed to the Imperial Council of Agricultural Research, Government of India. At present, crop statistics is built up in different states of India through crop estimation surveys at state and district levels. They can also be prepared at the tehsil (or taluka) level as tehsils are the strata in these surveys. These statistics are somewhat reliable at the state level, but not so at the district level. Certainly, they will not be reliable at the tehsil level.

The sampling design for crop estimation surveys is stratified multistage random sampling in which tehsils constitute the strata, villages in the tehsils at the first-stage sampling units, fields within villages as second-stage sampling units and plots of a specified size in selected fields as the ultimate units of sampling. The random sampling at the first two stages consists of using the SRSWOR scheme and at the last stage, a modification of the SRSWOR scheme. The reliability of crop statistics is measured in terms of percentage standard error of the estimated average yield per hectare. The theoretical consideration under the normality model suggests that this percentage should be anywhere from 1.2 to 2.6 so that the estimated average yield does not differ from the true average yield by more than 5 percent with 95 percent confidence (Tikkiwal, 1991). If this percent is higher, then the error in the estimated yield proportionately increases and thus the reliability of the estimated yield proportionately decreases.

The picture of the reliability of Statistics of different crops in Rajasthan, the State of India over the five years 1981-82 to 1985-86 is obtained, and given crop-wise percentage standard error of the state-level average yield per hectare for thirteen crops and district level average yield per hectare for three important crops: Maize, Paddy, and Wheat. It is noted that:

1. The average reliability for the wheat crop at the state level is as expected. But for other crops, it is not so. Thus, we need to enlarge our sample surveys for other crops to have the desired level of reliability at the state level itself.

2. The situation is much worse for district-level estimates. Even to keep their reliability somewhat closer to the state level, we have to enlarge our sample surveys considerably, which will have prohibitive cost. We can meet both the situations by first enlarging the sample size through simulation and then making use of auxiliary information on fertilizer use etc. through SICURE modeling without much additional cost. We explain below in detail how this can be affected.

There are districts where there is little variation in percentage standard error of average yield of a crop from year to year. In some others, where there is variation, it can be reduced by analyzing fertilizer data along with average yield data of corresponding years. Even after this, there may remain districts where this variation may persist. Thus, for simulation of yield in a given year at village level for a given district, we have two situations (i) A district belongs to the category 1, where it is not proper to pool the analysis of data of the current year with those of previous years for simulation purposes and (ii) A district belongs to the category 2 where it is proper to do so.

For both the situations, let *n* be the number of villages selected from a given tehsil of a given district in a given year and let *m* be the number of fields selected from a village selected in the sample. In each of the selected fields, let a plot of a specified size be selected in a specified manner. We now deal with the first situation.

The traditional estimator of average yield at tehsil level is obtained as

$$\bar{y}_{\bullet\bullet} = \frac{1}{n\,m}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

$$\bar{y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i\bullet} \qquad (3.1)$$

where $y_{i,j}$ is the yield per hectare of the randomly laid out plot in the j-th selected field of i-th selected village in the given tehsil. Let us consider a random effect model as

$$y_{ij} = \mu + V_i + F_{ij} \qquad (3.2)$$

$$NID(0,\delta_v^2)$$

For all i, j where v_i the $NID(0,\delta_f^2)$ random effect of i[th] village is for a given i and F_{ij} the random effect of j-th field in the i-th village is for a given i. Under the above model, the analysis of variance table for the survey data on yield provides mean sum of squares as

$$m\sum_{i=1}^{n}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2 = m\sum_{i=1}^{n}\left[\left(V_i - \bar{V}\right) + \left(\bar{F}_{i\bullet} - \bar{F}_{\bullet\bullet}\right)\right]^2$$

where $\bar{V}$, $\bar{F}_{i\bullet}$ and $\bar{F}_{\bullet\bullet}$ are averages for village and field effects like those for the y variate. Therefore,

$$E\left(m\sum_{i=1}^{n}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2\right) = m(n-1)\delta_v^2 + (n-1)\delta_f^2$$

Thus, $$E[\,M.S.S.due\ to\ villages] = m\,\delta_v^2 + \delta_f^2 \qquad (3.3)$$

Also,

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\bar{y}_{ij} - \bar{y}_{i\bullet}\right)^2\right] = n(m-1)\delta_f^2$$

which gives $$E\left[M.S.S.due\ to\ fields\ within\ villages\right] = \delta_f^2 \qquad (3.4)$$

Let $\hat{\delta}_f^2$ = M.S.S. due to fields within villages. Then from (3.4) we note that $\hat{\delta}_f^2$ is an unbiased estimator of δ_f^2. Let $\hat{\delta}_v^2$ be given by

$$\hat{\delta}_v^2 = \frac{\text{M.S.S. due to villages-M.S.S. due to fields within villages}}{\text{m}} \tag{3.5}$$

Thus, we see from (3.3) that $\hat{\delta}_v^2$ is also an unbiased estimator of δ_v^2. Now for obtaining an appropriate simulation technique at tehsil level, we note that $E\left(y_{ij}\middle| i\right) = \mu + V_i$. Thus, the simulated value of i-th village of a given tehsil can be obtain as

$$\hat{y}_{i\bullet} = \mu + V_i \tag{3.5a}$$

When the i-th village is a sample village in the tehsil, then

$$E\left(\bar{y}_{i\bullet}\middle| i\right) = \mu + V_i \tag{3.5b}$$

Thus, $\bar{y}_{i\bullet}$ can be taken as simulated value for an i[th] village of the tehsil if it is in the sample. For those villages of the tehsil, which are not in the sample, we can unbiasedly estimate μ by $\bar{y}_{\bullet\bullet}$. Thus, $\bar{y}_{\bullet\bullet} + V_i$ where V_i the normal variate value from $N(0, \delta_v^2)$ can be taken as the simulated value for the i-th village that is not in the sample. Since δ_v^2 is not known, we estimate the same by the analysis of variance.

Similarly, $y_{ij} = \mu + V_i + F_{ij}$ for all i, j, we can simulate a given y_{ij}, from the simulated yield of the corresponding i-th village in which the j-th fields lies, by using

$$\hat{y}_{ij} = \bar{y}_{i\bullet} + F_{ij} = \mu + V_i + F_{ij} \tag{3.5c}$$

where F_{ij} is a normal variate value from $N(0, \delta_f^2)$. Since δ_f^2 is not known, we estimate δ_f^2 by (3.4) once again through analysis of variance.

Thus, the comprehensive simulation plan at the tehsil level for determining the average crop yields of different villages in the tehsil and crop yields of cultivator's fields in them is contained in the following formulae. The simulated crop yield of the i[th] village of the tehsil is given by

$\hat{y}_{is} = \bar{y}_{i\bullet}$ if i-th selected village is the sample village;

$= \bar{y}_{\bullet\bullet} + V_i$ if i-th selected village is not a sample village.

The simulated yield of j-th cultivator's field in the i-th village of the tehsil is given by $\hat{y}_{ij} = \bar{y}_{i\bullet} + F_{ij}$

$\hat{y}_{ij} = \bar{y}_{i\bullet} + F_{ij}$; for j-th non selected field in the i-th selected village;

$= \bar{y}_{\bullet\bullet} + V_i + F_{ij}$; for j-th field in the i-th non-selected village.

4. SICURE-MODELING & GAIN IN EFFICIENCY AT LOWER LEVEL

Initially we select n villages and then let us select some more villages say n_1 from the particular tehsil, by the SRSWOR scheme, so as to make the total selected villages as $n'(>n)$. If we simulate the yield of additional selected villages in number $n'-n$ and then take the average yield from these selected villages, then we have the estimator

$$\bar{y}_{n_1} = \frac{1}{n'-n}\sum_{i=1}^{n'-n}\hat{y}_{is}$$
$$= \frac{1}{n'-n}\sum_{i=1}^{n'-n}\left(\bar{y}_{\bullet\bullet}+V_i\right)$$

Now, the average yield from the total selected villages is

$$\bar{y}_{n'} = \frac{n\bar{y}_n + n_1\bar{y}_{n_1}}{n+n_1}$$

$$\bar{y}_{n'} = \bar{y}_{\bullet\bullet} + \frac{1}{n'}\sum_{i=1}^{n'-n}V_i \qquad (4.1)$$

As auxiliary information plays a very important role in the estimation therefore, we build up another estimator at the tehsil level using auxiliary information, such as the area of the crop grown, fertilizer used or any physical measurement on plants before the crop is harvested. For all i let A_i denote the value of the i-th village of a particular tehsil for particular auxiliary character. Without any loss of generality, we take $A_1, A_2, \ldots, A_{n'}$ as the values of n' villages selected in the sample for the auxiliary character. Let

$$\bar{y}_r = \bar{y}_{n'} + b(\bar{A}_N - \bar{A}_{n'}) \qquad (4.2)$$

with

$$b = \frac{\sum_{i=1}^{n'}(\hat{y}_i - \bar{y}_{n'})(\bar{A}_i - \bar{A}_{n'})}{\sum_{i=1}^{n'}(\bar{A}_i - \bar{A}_{n'})^2}$$

Here the variance of $\bar{y}_r$ and a consistent and asymptotically unbiased estimator of the variance of $\bar{y}_r$ can be obtain as

$$V(\bar{y}_r) = \left(\frac{1}{n'} - \frac{1}{N}\right)\left(1-\rho^2\right)S_b^2 \qquad (4.3)$$

And

$$\hat{V}(\bar{y}_r) = \left(\frac{1}{n'} - \frac{1}{N}\right)\left(1-\hat{\rho}^2\right)s_b^2 \qquad (4.4)$$

with

$$\hat{\rho} = \frac{\sum_{i=1}^{n'} (\hat{y}_i - \bar{y}_{n'})(A_i - \bar{A}_{n'})}{\sqrt{\sum_{i=1}^{n'} (\hat{y}_i - \bar{y}_{n'})^2} \sqrt{\sum_{i=1}^{n'} (A_i - \bar{A}_{n'})^2}}$$

To see the gain due to SICURE modeling consider the following expression of the difference of the estimates of variances

$$\hat{V}(\bar{y}_{\bullet\bullet}) - \hat{V}(\bar{y}_r) = \frac{N-n}{Nn} s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2 - \left(\frac{1}{n'} - \frac{1}{N}\right)(1-\hat{\rho}^2)s_b^2$$

$$\hat{V}(\bar{y}_{\bullet\bullet}) - \hat{V}(\bar{y}_r) = s_b^2\left(\frac{N-n}{Nn} - \frac{N-n'}{Nn'}(1-\hat{\rho}^2)\right) + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2$$

$$= s_b^2\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n'} - \frac{1}{N}\right)(1-\hat{\rho}^2) + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2$$

$$= s_b^2\left(\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n'} - \frac{1}{N}\right)(1-\hat{\rho}^2)\right) + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2$$

If N and M will be quite large, then the above expression will become,

$$\hat{V}(\bar{y}_{\bullet\bullet}) - \hat{V}(\bar{y}_r) \cong s_b^2\left(\frac{1}{n} - \frac{1}{n'}(1-\hat{\rho}^2)\right)$$

$$= \frac{1}{n}(s_b^2) - s_b^2\frac{(1-\hat{\rho}^2)}{n'} \tag{4.5}$$

Thus, the approximate percentage reduction for large N in estimated variance due to the use of SICURE modeling is

$$\frac{\left(\frac{s_b^2}{n}\right) - \left(\frac{1-\hat{\rho}^2}{n'} s_b^2\right)}{\left(\frac{s_b^2}{n}\right)} \times 100 \tag{4.6}$$

Here, for an idea of the reduction in the variance, let $n' \geq 2n$ i.e. we simulate observations at least equal to the sample size. Also, correlation coefficient is taken greater than or equal to half. Then from (4.6) we have at least 62.5 percent reduction in variance. Since, generally the correlation coefficient is expected to be much higher than half and since we can easily simulate many more observations than n, the gain in efficiency at tehsil level would be considerable in using SICURE modeling. This gain can be carried forward to the district and other higher levels.

Further, the same result can be shown by taking the expressions of variance of $\bar{y}_{\bullet\bullet}$ and $\bar{y}_r$ i.e. let there be N PSUs and let each PSU consist of M SSUs (Seconday Sampling Units). Let $\bar{Y}_{NM}$ denote the mean of the population to be estimated through a sample of n PSU's and m SSU's in each of the n sample PSU's drawn with

the help of SRSWOR scheme at both the stages. Then, in the notation parallel to previous sections we know that $\bar{y}_{\bullet\bullet}$ is the design unbiased estimator of $\bar{Y}_{NM}$ and its variance is given by

$$V(\bar{y}_{\bullet\bullet}) = \frac{N-n}{Nn} S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2 \tag{4.7}$$

Let us consider the expression (4.3) and (4.7) and compute

$$V(\bar{y}_{\bullet\bullet}) - V(\bar{y}_r) = \left[\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n'} - \frac{1}{N}\right)\left(1-\rho^2\right)\right] S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2$$

Here, $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$, If N and M will be quite large, then the above expression will become,

$$V(\bar{y}_{\bullet\bullet}) - V(\bar{y}_r) = \left(\frac{S_b^2}{n} - \frac{S_b^2}{n'}\left(1-\rho^2\right)\right) \tag{4.8}$$

Thus, the approximate percentage reduction for large N, in variance due to the use of SICURE modeling is obtain as

$$\frac{\left(\frac{S_b^2}{n}\right) - \left(\frac{1-\rho^2}{n'} S_b^2\right)}{\left(\frac{S_b^2}{n}\right)} \times 100 \tag{4.9}$$

5. SICURE-MODELING FOR CROP ESTIMATION IN SMALL AREAS WITH UNEQUAL FIRST STAGE UNITS

Now, we will discuss the case for the first-stage units of unequal size when simple random sampling is employed at each stage. Let us denote

M_i = the number of second-stage units in the i-th first stage unit, for i=1, 2, …, N

$M_0 = \sum_{i=1}^{N} M_i$, the total number of second-stage units in the population,

m_i = the number of second-stage units to be selected from the i-th first-stage unit, if it is in the sample,

$m_0 = \sum_{i=1}^{n} m_i$, the number of second-stage units in the sample

Several estimates of the population mean can be formed. The simplest is the mean of the first-stage unit means in the sample, i.e.

$$\bar{y}_{s_2} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i(m_i)} \tag{5.1}$$

where the summation runs over the first-stage units in the sample, and $\bar{y}_{i(m_i)}$ represents the arithmetic mean of the m_i selected second-stage units in the i-th first-stage unit.

A second estimate, to be denoted by $\bar{y}'_{s_2}$ is based on the first-stage unit totals and is given by

$$\bar{y}'_{s_2} = \frac{1}{n}\sum_{i=1}^{n} u_i \, \bar{y}_{i(m_i)} \tag{5.2}$$

where, $u_i = \frac{M_i}{\bar{M}}$

A third estimate $\bar{y}''_{s_2}$ is the ratio-estimate given by

$$\bar{y}''_{s_2} = \frac{\sum_{i=1}^{n} u_i \, \bar{y}_{i(m_i)}}{\sum_{i=1}^{n} u_i}$$

Also, it can be written as

$$\bar{y}''_{s_2} = \frac{n\,\bar{y}'_{s_2}}{\sum_{i=1}^{n} u_i} = \frac{\bar{y}'_{s_2}}{\bar{u}_n} \tag{5.3}$$

More generally we may consider a ratio estimate of the population mean by letting x_{ij} be the value of an auxiliary variable x corresponding to the value y_{ij} of y, the variable under study. Let

$$\bar{X}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{N} u_i \, \bar{X}_{i\bullet}$$

and

$$\bar{x}'_{s_2} = \frac{1}{n}\sum_{i=1}^{n} u_i \, \bar{x}_{i(m_i)}$$

Then, the general ratio estimate of the population mean is defined as

$$\bar{y}_{RS_2} = \frac{\bar{y}'_{s_2}}{\bar{x}'_{s_2}} \bar{X}_{\bullet\bullet} \tag{5.4}$$

Among the above three estimators the estimator $\bar{y}'_{s_2}$ is an unbiased estimator for the population mean in the case of two stage sampling with unequal first stage unit (cf. Sukhatme and Sukhatme (1997)). Thus, let us consider estimate (5.2) for the estimation purpose. For the assumed model

$$y_{ij} = \mu + V_i + F_{ij}$$

We have,

$$\bar{y}_{i\bullet} = \mu + V_i + \bar{F}_{i\bullet}$$

and

$$\bar{y}_{\bullet\bullet} = \frac{1}{n}\left(\mu\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} u_i V_i + \sum_{i=1}^{n} u_i \bar{F}_{i\bullet} \right) \tag{5.5}$$

The analysis of variance for Agricultural Crop Yield per Hectare Data provides the mean sum of squares due to village as,

$$\sum_{i=1}^{n} m_i \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{n} m_i \left[\mu\left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right) + \left(V_i - \frac{1}{n}\sum_{i=1}^{n} u_i V_i\right) + \left(\bar{F}_{i\bullet} - \frac{1}{n}\sum_{i=1}^{n} u_i \bar{F}_{i\bullet}\right)\right]^2 \tag{5.6}$$

Therefore,

$$E\left(\sum_{i=1}^{n} m_i \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2\right) = \left[E\sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 + E\sum_{i=1}^{n} m_i \left(V_i - \frac{1}{n}\sum_{i=1}^{n} u_i V_i\right)^2 + E\sum_{i=1}^{n} m_i \left(\bar{F}_{i\bullet} - \frac{1}{n}\sum_{i=1}^{n} u_i \bar{F}_{i\bullet}\right)^2 \right]$$

$$E\left(\sum_{i=1}^{n} m_i \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2\right) = \left[\sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 + \sum_{i=1}^{n} m_i E\left(V_i^2\right) + \frac{1}{n^2} n \sum_{i=1}^{n} u_i^2 E\left(V_i^2\right) + \sum_{i=1}^{n} m_i E\left(\bar{F}_{i\bullet}^2\right) + \frac{1}{n^2} n \sum_{i=1}^{n} u_i^2 E\left(\bar{F}_{i\bullet}^2\right) \right]$$

$$= \sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 + \sum_{i=1}^{n} m_i \delta_v^2 + \frac{1}{n}\sum_{i=1}^{n} u_i^2 \delta_v^2 + \sum_{i=1}^{n} m_i \frac{\delta_f^2}{m_i} + \frac{1}{n}\sum_{i=1}^{n} u_i^2 \frac{\delta_f^2}{m_i}$$

$$= \sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 + \delta_v^2 \left(\sum_{i=1}^{n} m_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2\right) + \delta_f^2 \left(n + \frac{1}{n}\sum_{i=1}^{n} \frac{u_i^2}{m_i}\right) \tag{5.7}$$

As, $F_{ij} \sim N\left(0, \delta_f^2\right)$ thus, $\bar{F}_{i\bullet} \sim N\left(0, \delta_f^2/m_i\right)$. Also, $\bar{F}_{i\bullet\bullet} \sim N\left(0, \left(\delta_f^2/n\right)^2 \sum_{i=1}^{n} \frac{1}{m_i}\right)$

Thus,

$$E\left[M.S.S.due\ to\ villages\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 + \delta_v^2 \left(\sum_{i=1}^{n} m_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2\right) + \delta_f^2 \left(n + \frac{1}{n}\sum_{i=1}^{n} \frac{u_i^2}{m_i}\right) \right] \tag{5.8}$$

Also,

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} \left(y_{ij} - \bar{y}_{i\bullet}\right)^2\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mu + V_i + F_{ij} - \mu - V_i - \bar{F}_{i\bullet}\right)^2\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} E\left(F_{ij}\right)^2 + \sum_{i=1}^{n} m_i E\left(\overline{F}_{i\bullet}\right)^2 - 2\left\{\sum_{i=1}^{n} m_i E\left(\overline{F}_{i\bullet}\right)^2\right\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} \delta_f^2 - \sum_{i=1}^{n} m_i \frac{\delta_f^2}{m_i} = \sum_{i=1}^{n} m_i\, \delta_f^2 - n\delta_f^2$$

$$= \left(\sum_{i=1}^{n} m_i - n\right)\delta_f^2 \tag{5.9}$$

which gives

$$E\left[\text{M.S.S. due to fields within villages}\right] = \delta_f^2 \tag{5.10}$$

Let $\hat{\delta}_f^2$ = M.S.S due to fields within villages. Then from (5.10) note that $\hat{\delta}_f^2$ is an unbiased estimator of δ_f^2. Also $\hat{\delta}_v^2$ can be obtain using the expression (5.8) and (5.10) as

$$\hat{\delta}_v^2 = \frac{1}{\left(\sum_{i=1}^{n} m_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2\right)}\left[\begin{array}{l}(n-1)E\left(M.S.S.due\ to\ villages\right)\\ -E\left(M.S.S.\ due\ to\ fields\ within\ villages\right)\\ *\left(n + \frac{1}{n}\sum_{i=1}^{n}\frac{u_i^2}{m_i}\right) - \sum_{i=1}^{n} m_i \mu^2 \left(1 - \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2\end{array}\right] \tag{5.11}$$

Now to compare the efficiency of the estimator of SICURE model with $\overline{y}'_{s_2}$ let us compute the difference of their variances as

$$V(\overline{y}'_{s_2}) - V(\overline{y}_r) = \left(\frac{1}{n} - \frac{1}{N}\right)S_b'^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2 - \left(\frac{1}{n'} - \frac{1}{N}\right)\left(1-\rho^2\right)S_b^2 \tag{5.12}$$

When N is sufficiently large, the difference can be obtain as

$$V(\overline{y}'_{s_2}) - V(\overline{y}_r) = \frac{S_b'^2}{n} - \frac{S_b^2}{n'}\left(1-\rho^2\right) \tag{5.13}$$

Thus, the approximate percentage reduction for large N, in variance due to the use of SICURE modeling is obtain

$$\frac{\left(\frac{S_b'^2}{n}\right) - \left(\frac{1-\rho^2}{n'} S_b^2\right)}{\left(\frac{S_b'^2}{n}\right)} \times 100 \tag{5.14}$$

6. Estimators Consider for Simulation Study under Two Stage Sampling Design with Equal First Stage Units

For our simulation study we consider and present the following traditional and SICURE Model based estimators of population mean of small domain, under two-stage sampling scheme.

(1) $$\hat{T}_1 = \bar{y}_{\bullet\bullet} = \frac{1}{mn}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right) = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i\bullet}$$

(2) $$\hat{T}_2 = \bar{y}_r = \bar{y}_{n'} + b(\bar{A}_N - \bar{A}_{n'})$$

where $$b = \frac{\sum_{i=1}^{n'}(\hat{y}_i - \bar{y}_{n'})(\bar{A}_i - \bar{A}_{n'})}{\sum_{i=1}^{n'}(\bar{A}_i - \bar{A}_{n'})^2}$$

and $$\bar{y}_{n'} = \bar{y}_{\bullet\bullet} + \frac{1}{n'}\sum_{i=1}^{n'-n} V_i$$

Here $n' = n + n_1$ and $V_i \sim N(0, \hat{\delta}_v^2)$

where

$$\hat{\delta}_v^2 = \frac{\text{M.S.S. due to villages-M.S.S. due to fields within villages}}{\text{m}}$$

Details of Simulation Study

In order to compare the performances of traditional estimators for small areas to the SICURE model estimator for the simulated crop production statistics for the tehsil level, we conduct a simulation study. We have taken villages as sampling units and 2000 independent two stage samples of different sizes are selected from the population of 150 villages with 50 fields each from a particular tehsil. For both small area estimators under consideration and for each sample size we compute Percentage Absolute Relative Bias (ARB) and Percentage Simulated Relative Standard Error (SRSE) as defined below.

$$ARB(\hat{T}_i) = \frac{\left|\frac{1}{2000}\sum_{s=1}^{2000}\hat{T}_i^s - T\right|}{T}$$

and

$$SRSE(\hat{T}_i) = \frac{\sqrt{ASE(\hat{T}_i)}}{(T)}$$

where $$ASE(\hat{T}_i) = \frac{1}{2000}\sum_{s=1}^{2000}(\hat{T}_i^s - T)^2$$ and $$E(\hat{T}_i) = \frac{1}{2000}\sum_{s=1}^{2000}\hat{T}_i^s$$

Here, subscript 'i' is used for a particular small area estimator ($i=1,2$). The simulation and computation work is done using MATLAB software.

TABLE 6.1 ABSOLUTE RELATIVE BIAS (%) & SIMULATED RELATIVE STANDARD ERROR (%) OF THE ESTIMATORS (AND) FOR TEHSIL UNDER TWO STAGE SAMPLING DESIGN

n	n_1	m	$\hat{T}_1$($\hat{T}_2$)	ARB		SRSE	
				$\hat{T}_1$	$\hat{T}_2$	$\hat{T}_1$	$\hat{T}_2$
30	30	10	-0.0738	0.004	0.0189	0.9727	0.8871
		15	-0.0478	0.0385	0.0215	0.9329	0.8763
		20	-0.072	0.0085	0.0261	0.9502	0.8647
20	20	10	-0.1212	0.0008	0.0204	1.2372	1.1271
		15	-0.1505	0.0049	0.0028	1.2203	1.0795
		20	-0.1001	0.011	0.0379	1.189	1.0949
10	20	10	-0.0317	0.0042	0.0184	1.8433	1.825
		15	-0.0412	0.0092	0.0476	1.7864	1.7624
		20	-0.0801	0.0042	0.0104	1.8022	1.7538
5	10	10	0.4621	0.0215	0.0505	2.5319	2.7204
		15	0.2842	0.0618	0.0707	2.5565	2.6739
		20	0.4799	0.0621	0.0042	2.5652	2.7576

CONCLUSIONS AND RECOMMENDATIONS

The Table 6.1 shows the Percentage Absolute Relative Bias and Percentage Simulated Relative Standard Error for the estimators $\hat{T}_1$ and $\hat{T}_2$ under different sample sizes 30, 20, 10, 5 in the first stage and 10, 15, 20 at second stage units for small domains. Followings are some notable points:

As per our observations in Table 6.1 for the sample size 30 the Percentage ARB of for tehsil varies from 0.004 to 0.0385 while that of it varies from 0.0186 to 0.0261. As regards the Percentage SRSE it varies from 0.93 to 0.97 and 0.86 to 0.88 for the estimators and respectively which shows the considerable decrease in SRSE values of

The same result is also seen for the sample sizes 20 and 10 i.e. in some cases the ARB values of is smaller than the ARB of , while in some cases has smaller ARB than . But for the SRSE values are concern is consistently smaller than in all samples other than sample 5.

REFERENCES

[1] Brewer, K.W.R. (1963). Ratio estimation in Finite Populations-Some Results Deducible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics*, 93-105.

[2] Cronkhite, F.R. (1987). Use of Regression Techniques for Developing State and Area Employment and Unemployment Estimates. In: Small Area Statistics (eds. R. Platek, J.N.K.Rao, C.E. Sarndal and M.P. Singh), Wiley, New York, 162-174.

[3] Erickson, E.P. (1974). A regression method for estimation population changes for local areas. Journal of the American Statistical Association, 69, 867-875.

[4] Feeney, G.A. (1987). The Estimation of the no. of Unemployed at the Small Area Level.

[5] Small Area Statistics. In: Small Area Statistics (eds. R. Platek, J.N.K.Rao, C.E. Sarndal and M.P. Singh), Wiley, New York, 198-218.

[6] Heeringa, S.G. (1981). Small Area Estimation Prospects for the Survey of Income and Program Participation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 133-138.

[7] Mahalanobis, P.C. (1946). Sample Surveys of Crop Yields in India. Sankhya, 269-280.

[8] McCullagh, P. and Zidek, J.V. (1987). Regression Methods and Performance Criteria for Small Area Estimation, in Small Area Statistics (eds. R. Platek, J.N.K.Rao, C.E. Sarndal and M.P. Singh), Wiley, New York, 62-74.

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

[10] Royall, R. A. (1977). Statistical Theory of Small Area Estimates- Use of Prediction Models. Unpublished Report Prepared under Contract from the National Center for Health Statistics.

[11] Schaible, W.L., Brock, D.B., Casady, R.J. and Schnack, G.A. (1979). Small Area Estimation: An empirical comparison of conventional and synthetic estimators for states. Vital and Health Statistics, Series 2, No. 82, National Centre for Health Statistics, dept. of Health, Education, and welfare, Hyattsville, M.D.

[12] Sukhatme, P.V. and Agarwal, O.P. (1946-47, 47-48). Report on the crop cutting survey on wheat by the random sampling method in Delhi. Indian Council of agriculture research, New Delhi.

[13] Tikkiwal, B.D. (1960). On the theory of classical regression and double sampling estimation. Journal of Royal Statistical Society, Series B, 131-138.

[14] Tikkiwal, B.D. (1980). Successive Sampling-a review. An invited paper as a principal speaker, proceeding of the 42th session of the International Statistical Institute, 367-387.

[15] Tikkiwal, B.D. (1982). On conceptual and theoretical framework for survey sampling. An unpublished note as invited discussant in the 43rd biennial session of the International Statistical Institute at Buenos Aires, Argentina.

[16] Tikkiwal, B.D. (1991). Modeling through survey data for small domains. The keynote address at the symposium on modeling held at Kurukshetra University, March 7-9, 1991.

[17] Tikkiwal, G.C. and Gupta, A.K. (1991). Estimation of population mean under successive sampling when various weights and regression coefficients are unknown. *Biometrical Journal*, **33**, 529-538.

[18] Tikkiwal, G.C. and Ghiya, A. (2000). A generalized class of synthetic estimators with application to crop acreage estimation for small domains. Biometrical Journal, 42, 865-876.

[19] Watson, D.J. (1937). The estimation of leaf area in field crops. Journal of Agricultural Science, 474-483.

APPENDIX

TABLE 1. THE PERCENTAGE STANDARD ERROR OF DISTRICT-WISE AVERAGE YIELD PER HECTARE FOR THE YEARS 1981-86 IN CASE OF MAIZE, PADDY AND WHEAT CROPS

District/Year	1981-82			1982-83			1983-84			1984-85			1985-86		
	Maize	Paddy	Wheat	Maize	Paddy	Wheat	Maize	Paddy	Wheat	Maize	Paddy	Wheat	Maize	Paddy	Wheat
1. Ajmer	21.43	-	29.56	18.41	-	11.3	10.38	-	8.29	11.18	-	15.7	27.78	-	9.17
2. Alwar	-	-	10.25	13.24	-	3.92	19.3	-	7.82	8.2	-	9.75	41	-	5.95
3. Banswara	14.71	12.12	16.9	11.02	22.5	15	8.26	12.3	11.86	14.63	12.49	14.7	26.21	45.5	11.7
4. Barmer	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5. Bharatpur	-	42.12	14.56	-	9.79	6.43	-	5.61	7.33	-	9.88	7.32	-	15.2	5.06
6. Bhilwara	7.09	-	13.95	14.56	-	7.42	9.46	-	12.88	7.53	-	13	13.09	-	10.5
7. Bikaner	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8. Bundi	21.97	23.2	15.58	31.17	8.17	9.91	12.44	20.8	53.08	12.61	17.06	8.46	13.38	11.2	9.01
9. Chittor	8.59	18.38	14.01	7.04	31.4	11.3	10.29	11.3	13.08	6.54	18.7	9.42	9.07	26.3	5.01
10. Churu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11. Dholpur	-	-	-	-	-	-	-	5.27	11.8	-	47.52	11.7	-	36.3	10.7
12. Dungarpur	18.27	15.59	20.71	16.00	21.00	20.8	12.92	7.84	17.76	12.18	11.59	16	31.18	63	11.4
13. Gangnagar	-	10.21	6.02	-	8.13	4.54	-	9.31	4.64	-	8.78	7.5	-	5.99	3.92
14. Jaipur	26.5	-	11.67	25.18	-	10.7	21.47	-	7.02	15.27	-	8.47	53.61	-	3.84
15. Jaisalmer	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16. Jalore	-	-	4.37	-	-	8.44	-	-	11.26	-	-	13.8	-	-	7.26
17. Jhalawar	9.04	-	20.36	17.92	-	10.1	6.87	-	10.68	8.8	-	14.6	12.9	-	12.3
18. Jhunjhunu	-	-	-	-	-	-	-	-	-	-	-	12.9	-	-	13.9
19. Jhodhpur	-	-	27.08	-	-	18.4	-	-	28.24	-	-	14.6	-	-	13.5
20. Kota	12.22	12.9	8.65	25.6	27.1	6.65	9.74	20.4	6.01	8.46	16.3	13.6	26.4	16.3	5.43
21. Nagaur	-	-	45.06	-	-	13	-	m -	19.14	-	-	8.95	-	-	8.46
22. Pali	31.13	-	15.4	12.79	-	12.8	10.14	-	11.13	15.9	-	7.25	33.88	-	6.95
23. S. M.Pur	-	22.1	14.86	-	24.7	8.87	-	17.2	13.49	-	22.82	7.5	-	24	3.96
24. Sikar	-	-	com JP	-	-	21.3	-	-	21.92	-	-	7.8	-	-	9.01
25. Sirohi	18.1	-	com Jr	14.4	-	2.05	13.35	-	14.17	9.64	-	12.7	21.65	-	7.22
26. Tonk	23.59	-	21.41	19.02	-	12	15.16	-	11.14	10.35	-	9.24	25.13	-	8.44
27. Udaipur	8.51	14.83	7.92	9.15	16.7	11.5	9.28	13.3	6.99	6.01	13.28	3.73	10.43	30.4	4.6
State-Level Percentage Error	3.98	6.54	3.64	4	5.42	2.26	3.68	4.77	2.81	2.82	5.04	2.46	4.87	4.72	1.51

Source: Relevant Reports on General Crop Estimation Survey in Rajasthan, Board of Revenue, Ajmer, Raj.

Thus, we find that the gain in efficiency at the tehsil level is considerable by using SICURE modeling. This gain can be carried forward to the district and other higher levels.