Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 4, Issue 2 (2011): 83-91

SMALL AREA ESTIMATION IN LONGITUDINAL SURVEYS

Bhim Singh[1] and B.V.S. Sisodia[2]
[1]Department of Basic Science, College of Horticulture & Forestry, MPUAT, Jhalarapatan, Jhalawar-326 023, India.
E Mail: bhimsingh1@ gmail.com
[2]Department of Agricultural Statistics, NDUAT, Faizabad-224 229, India.
E Mail: bvs@india.com

Abstract

Longitudinal surveys are very common in sampling over interval of times to estimate the aggregate level of population means at given point of time. In the present paper, the estimation methods have been developed for small area in longitudinal surveys using small area estimation (SAE) techniques. Direct, synthetic and composite estimators have been proposed to estimate the population mean of small area at given point of time. We also provide detailed of mean square errors (MSE) of the proposed estimator. An empirical study is carried out to show the properties of the proposed estimators.

Keywords: Longitudinal surveys, small area estimation, direct estimator, synthetic estimator, composite estimator, means square errors.

1. Introduction

It is a common practice for survey organisations to estimate the population parameter at regular interval of time such as mean or total, which varies with time if characteristics and composition of population are changing over time. Now, in decentralized planning process at Panchayat level, data needs for micro-level become more and more important in order to focus the aggregate as well changes over time. One of the design in sampling over time is a longitudinal surveys in which similar measurements are made on the same sample at different point of time. Such surveys are also concerned with population sub-groups (domain/small area) that have experienced the same event during the same period, for example, cattle in milk calved in a particular week, or of same lactation etc. This all makes difference in estimating average milk yield. Similarly, milk yield also varies considerably in various socio-economic groups of households/farmers. All such populations with multiple observations can be defined as two dimensional populations, units having a spread over space and the observations giving a spread over time. Surveys dealing with multiple observations spread over time are defined as longitudinal surveys. When one is concerned about estimates of population sub-groups in longitudinal surveys, particularly the pattern of changes at individual level as well as aggregate level over time, the techniques of SAE can possibly be employed to achieve the above goal.

The early work in the small area estimation can be dated back to 1966 when Panse *et al.* (1966) and Singh (1968) examined the feasibility of using double sampling for estimation of yield at the block level. Various small area estimation methods have been developed in recent past by Gonzalez (1973), Gonzalez and Singh (1977), Purcell and Kish (1979, 1980), Drew *et al* (1982). Many studies dealing with the small area

estimation problem have been discussed by various authors (Ghosh and Rao, 1994; Rao, 1999; Singh and Sisodia, 2001). Ferrante and Pacei (2004) derived small area estimators as extensions of an estimation strategies proposed by Fuller (1990) for partial overlap sample.

In the present paper it has been attempted to propose the estimates of population mean for small areas in longitudinal surveys. The relative efficiencies of the proposed estimators are studied. It is also illustrated with an empirical data collected through a survey.

2. The Proposed Estimators

Let the population U = {1, 2, …, N } is divided into D non-overlapping small areas such that $\sum_{d=1}^{D} N_d$ = N. Let the population also be divided into G non-overlapping groups, which are considered to be larger than small area, and $\sum_{g=1}^{G} N_g = N$ and also $\sum_{d=1}^{D}\sum_{g=1}^{G} N_{dg} = N$. Let Y be the characteristic of interest and object is to estimate the population mean, $\bar{Y}_d$, for small area U_d, d = 1, 2, …, D, over different period of time in longitudinal surveys. Let us consider that a sample, s, of size n from U is drawn using simple random sampling without replacement. Let s_d denotes the intersection of s and U_d with cardinality n_d, which is a random variable subject to $\sum_{d=1}^{D} n_d = n$. Similarly, s_g denotes the intersection of s and U_g with cardinality n_g s.t. $\sum_{g=1}^{G} n_{.g} = n$. The s_{dg} denotes the intersection of s and U_{dg} with cardinality n_{dg} s.t. $\sum_{d=1}^{D}\sum_{g=1}^{G} n_{dg} = n$.

Moreover, assume that observations on Y are recorded for the selected units n over different interval of times t, t= 1, 2, …, T. Let $Y_{dgk}^{(t)}$ be the value of the character on y k^{th} unit in the $(d,g)^{th}$ cell, where k = 1, 2, …, N_{dg} (the number of population units in $(d, g)^{th}$ cell). The object is to estimate population mean $\bar{Y}_d^{(t)}$ for a specific point time t; t = 1, 2, …, T and d = 1, 2, …, D, such that

$$\bar{Y}_d^{(t)} = \sum_{g=1}^{G}\sum_{k=1}^{N_{dg}} \frac{Y_{dgk}^{(t)}}{N_d} = \sum_{g=1}^{G} \frac{N_{dg}}{N_d}\bar{Y}_{dg}^{(t)} \qquad (1)$$

where $\bar{Y}_{dg}^{(t)} = \sum_{k=1}^{N_{dg}} \frac{Y_{dgk}^{(t)}}{N_{dg}}$

Here, the estimation methods are developed for small area in longitudinal surveys using small area estimation (SAE) techniques. Direct, synthetic and composite estimators are proposed to estimate the population mean for small area over different period of time in longitudinal surveys.

2.1 Direct Estimator

A direct estimator of $\bar{Y}_d^{(t)}$ for t[th] time point is given by

$$\bar{y}_d^{(t)} = \sum_{g=1}^{G} W_{dg} \bar{y}_{dg}^{(t)} \tag{2}$$

where $W_{dg} = \frac{N_{dg}}{N_d}$ and $\bar{y}_{dg}^{(t)} = \sum_{k=1}^{N_{dg}} \frac{y_{dgk}^{(t)}}{n_{dg}}$, i.e. the sample mean for (d, g)th cell.

Since (2) is unbiased estimator of $\bar{Y}_d^{(t)}$ and its variance is given by

$$V(\bar{y}_d^{(t)}) = \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)}$$

It is known that MSE is equal to the variance. So MSE of direct estimator can be written as

$$MSE(\bar{y}_d^{(t)}) = V(\bar{y}_d^{(t)}) = \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} \tag{3}$$

where $S_{dg}^{2(t)} = \frac{1}{(N_{dg}-1)} \sum_{k=1}^{N_{dg}} \left(Y_{dgk}^{(t)} - \bar{Y}_{dg}^{(t)} \right)^2$

2.2 Synthetic Estimator

For direct estimator, it is essential that sample size n is sufficiently large enough. However, even if n_d is large, there is likelihood that n_{dg} may be zero for some U_{dg}'s. In such situation, direct estimator leads to under estimation as well as it may have large variance. If small areas have the same characteristic as the large areas (group), the estimate of same characteristic for large areas, an estimator for large area is used to derive the estimates for small areas. This is equivalent to borrowing strength from larger areas having similar region for the small areas. The estimator developed for small areas borrowing strength from outside of small area but similar region is said to be synthetic estimator or indirect estimator.

Assuming that the groups g's (g = 1, 2, …, G) are similar region for the small area d's (d = 1, 2, …, D), the synthetic estimator of $\bar{Y}_d^{(t)}$ is given by

$$\bar{y}_d^{'(t)} = \sum_{g=1}^{G} W_{dg} \bar{y}_{.g}^{(t)} \tag{4}$$

where, $\bar{y}_{.g}^{(t)} = \sum_{d=1}^{D}\sum_{k=1}^{n_{dg}} y_{dgk}^{(t)} / n_{.g}$ and $n_{.g} = \sum_{d=1}^{D} n_{dg}$

obviously, $\bar{y}_{d}^{'(t)}$ is a biased estimator. The bias of $\bar{y}_{d}^{'(t)}$ is given by

$$B[\bar{y}_{d}^{'(t)}] = \sum_{g=1}^{G} W_{dg}[\bar{Y}_{.g}^{(t)} - \bar{Y}_{d.g}^{(t)}] \tag{5}$$

as $\bar{y}_{.g}^{(t)}$ is a unbiased for $\bar{Y}_{.g}^{(t)}$, the population mean for gth group. Evidently, bias reduces to zero if $\bar{Y}_{.g}^{(t)} = \bar{Y}_{dg}^{(t)}$ for all g = 1, 2, …, G. Therefore, if small areas are similar across the groups, then synthetic estimator (4) is an almost unbiased.

MSE of the synthetic estimator (4) is given by

$$MSE[\bar{y}_{d}^{'(t)}] = \left[\sum_{g=1}^{G} W_{dg}\left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)}\right)\right]^2 + \sum_{g=1}^{G} W_{dg}^2\left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}}\right)S_{g}^{2(t)} \tag{6}$$

where $S_{g}^{2(t)} = \sum_{d=1}^{D}\sum_{k=1}^{N_{dg}}\left(Y_{dgk}^{(t)} - \bar{Y}_{.g}^{(t)}\right)^2 /(N_{.g} - 1)$

2.3 Composite Estimator

The composite estimator of $\bar{Y}_{d}^{(t)}$, by combining (1) and (4), can be given in the following way when ϕ is arbitrary constant, 0 < ϕ <1.

$$\bar{y}_{d}^{''(t)} = \phi\bar{y}_{d}^{(t)} + (1-\phi)\bar{y}_{d}^{'(t)} \tag{7}$$

The bias and MSE of the estimator in equation (7) are respectively given by

$$B(\bar{y}_{d}^{''(t)}) = (1-\phi)\sum_{g=1}^{G} W_{dg}\left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)}\right) \tag{8}$$

$$MSE(\bar{y}_{d}^{''(t)}) = (1-\phi)^2\left(\sum_{g=1}^{G} W_{dg}\left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)}\right)\right)^2 + \phi^2\sum_{g=1}^{G} W_{dg}^2\left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}}\right)S_{dg}^{2(t)}$$

$$+(1-\phi)^2\sum_{g=1}^{G} W_{dg}^2\left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}}\right)S_{g}^{2(t)} + 2\phi(1-\phi)\sum_{g=1}^{G} W_{dg}^2\left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}}\right)S_{g}^{2(t)} \tag{9}$$

Optimum value of ϕ is given by

$$\phi_o = \frac{\left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2}{\left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2 + \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_g^{2(t)}}$$

Substituting this optimum value in equation (9), the minimum MSE of $\bar{y}_d^{"(t)}$ is given by

$$MSE\ (\bar{y}_d^{"(t)})_{\min} = \frac{AB + C(B-C)}{A+B-C}; \qquad (10)$$

where

$$A = \left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2; \quad B = \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} \text{ and}$$

$$C = \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_g^{2(t)}$$

Efficiency Comparisons

In this section, we compare the efficiencies of the proposed estimators as follows:

$$V(\bar{y}_d^{(t)}) - MSE(\bar{y}_d^{'(t)}) > 0,$$

$$\sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_g^{2(t)} - \left[\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right]^2 > 0$$

,

$$\sum_{g=1}^{G} W_{dg}^2 \left(\left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_g^{2(t)} \right) - \left[\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right]^2 > 0 \qquad (11)$$

If the condition (11) is satisfied, then $\bar{y}_d^{'(t)}$ is more efficient than $\bar{y}_d^{(t)}$. Moreover, if $\frac{1}{N_{dg}}$ and $\frac{1}{N_{.g}}$ can be approximated to zero for large N_{dg} and $N_{.g}$, then inequality (11) reduces to

$$\sum_{g=1}^{G} W_{dg}^2 \left(\frac{S_{dg}^{2(t)}}{n_{dg}} - \frac{S_{g}^{2(t)}}{n_{.g}} \right) - \left[\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right]^2 > 0 \qquad (12)$$

Obviously, the second term of (12) is generally expected to be very small as difference between two means, cell mean and corresponding group mean, is always expected to be small because of similar region. Since $n_{dg} < n_{.g}$, and $S_{dg}^{2(t)}$ is also expected to be larger than $S_{.g}^{2(t)}$, in general, therefore, the first term will be generally positive. Hence, the inequality (12) will hold true and synthetic estimator $\bar{y}_d^{\prime(t)}$ is more efficient than direct estimator $\bar{y}_d^{(t)}$, in general.

$$V(\bar{y}_d^{(t)}) - MSE(\bar{y}_d^{\prime\prime(t)}) > 0 ,$$

$$\frac{\left[\sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2 - \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_{g}^{2(t)} \right]^2}{\sum_{g=1}^{G} W_{dg} [\bar{Y}_{.g}^{(t)} - \bar{Y}_{d.g}^{(t)}] + \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2 - \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_{g}^{2(t)}} > 0$$

(13)

If the condition (13) is satisfied, the composite estimator $\bar{y}_d^{\prime\prime(t)}$ is more efficient than direct estimator $\bar{y}_d^{(t)}$.

$$MSE(\bar{y}_d^{\prime(t)}) - MSE(\bar{y}_d^{\prime\prime(t)}) > 0 ,$$

$$\frac{\left[\sum_{g=1}^{G} W_{dg} [\bar{Y}_{.g}^{(t)} - \bar{Y}_{d.g}^{(t)}] \right]^4}{\sum_{g=1}^{G} W_{dg} [\bar{Y}_{.g}^{(t)} - \bar{Y}_{d.g}^{(t)}] + \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) S_{dg}^{2(t)} - \left(\sum_{g=1}^{G} W_{dg} \left(\bar{Y}_{.g}^{(t)} - \bar{Y}_{dg}^{(t)} \right) \right)^2 - \sum_{g=1}^{G} W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}} \right) S_{g}^{2(t)}} > 0$$

(14)

If the condition (14) is satisfied, the composite estimator $\bar{y}_d^{\prime\prime(t)}$ is more efficient than indirect estimator $\bar{y}_d^{\prime(t)}$.

3. Empirical Investigation

In order to illustrate the performance of the estimators developed in previous sections, a survey was conducted in the vicinity of the N.D.U.A&T., Kumarganj, Faizabad to collect the data on milk production of the all farmers having milch cattle.

The population was classified into non-overlapping classes according to land holding of the household, i.e. marginal (<1 ha), small (1-2 ha), medium (2-4 ha) and large farmers (>4 ha). However, there were no large farmers. The population was further classified into non-overlapping classes according to breed of cattle, i.e., exotic and indigenous. The socio-economic classes based on land holding were considered as small area. Due to small number of the units in the higher socio-economic class, only two classes viz; marginal and other farmers were considered. It is essential assumption in small area estimation for borrowing strength from similar regions that small area must have similar characteristics across the larger non-overlapping groups of the population. It is expected in the present empirical study that small areas, i.e. socio-economic classes of farmers will have similar characteristics across the breed of the cattle. Therefore, breeds of the cattle were considered as groups. Under this assumption the study has been carried out. The estimate of the average milk yield per day with respect the marginal and other farmers based on direct, synthetic and composite estimators along with their variances/MSE are shown in Table 1. As expected the average milk yield has decreased over time. There has been, in general, consistency in the estimates of average milk yield obtained by all three estimators. Comparing the variances/MSE of these estimators, the composite estimator has been found to be more accurate. It can also be noted that estimates of variances/MSE were also stable time.

Point of time	Direct estimate		Synthetic estimate		Composite estimate	
	Marginal farmers	Other farmers	Marginal farmers	Other farmers	Marginal farmers	Other farmers
I	2.561 (0.090)	3.231 (1.725)	2.541 (0.064)	2.778 (0.344)	2.541 (0.064)	2.830 (0.321)
II	2.595 (0.103)	2.415 (1.127)	2.660 (0.071)	2.869 (0.410)	2.653 (0.071)	2.993 (0.342)
III	2.237 (0.085)	2.991 (1.006)	2.300 (0.063)	2.523 (0.316)	2.298 (0.062)	2.614 (0.274)
IV	1.984 (0.099)	2.372 (0.524)	1.977 (0.060)	2.156 (0.117)	1.977 (0.060)	2.176 (0.112)
V	1.621 (0.085)	2.660 (0.833)	1.794 (0.082)	1.995 (0.531)	1.712 (0.068)	2.243 (0.366)
VI	1.652 (0.085)	2.545 (0.675)	1.804 (0.073)	1.979 (0.394)	1.744 (0.063)	2.176 (0.282)
VII	1.684 (0.092)	2.421 (0.535)	1.814 (0.068)	1.963 (0.274)	1.776 (0.063)	2.104 (0.210)
VIII	1.593 (0.097)	2.331 (0.568)	1.730 (0.075)	1.887 (0.265)	1.685 (0.069)	2.012 (0.210)

NB: Figures in parentheses indicate variance/MSE.

Table 1: Estimates of average milk yield (kg/day) for marginal and other farmers

Points of time	Relative efficiencies in percentage			
	Marginal farmers		Other farmers	
	Synthetic Estimator	Composite estimator	Synthetic Estimator	Composite estimator
I	139.75	139.76	501.31	538.05
II	144.66	145.54	274.81	329.44
III	134.92	136.07	318.25	367.69
IV	164.73	164.86	449.01	466.61
V	103.79	125.58	156.93	227.78
VI	116.76	136.44	171.54	239.19
VII	135.49	146.11	195.18	255.13
VIII	129.68	139.87	214.74	271.64

Table 2: Relative efficiencies of the proposed estimator

The relative efficiency of the synthetic and composite estimators over the direct estimator has been computed as follows:

$$E_i = \frac{V(\bar{y}_d^{(t)})}{M_i} X\ 100\ ,\ (i = 1, 2)$$

where M_1 and M_2 are MSE of the synthetic and composite estimators respectively.

The computed relative efficiencies are given in Table 2 for both marginal and other farmers at different point of time. We see that the relative efficiencies of the synthetic and composite estimators over direct estimator have been quite appreciable at different points of time in case of marginal farmers as well as other farmers. In general, the relative efficiencies of composite estimator were found comparatively to be more as compared to indirect estimator.

4. Conclusion

In this paper, the estimation methods have been proposed to estimate the population mean for small area in longitudinal surveys based on SAE techniques. The MSE of the proposed estimators have been derived and their relative efficiencies have been worked out theoretically. These theoretical results have also been satisfied by an empirical data collected through a survey. It is found that the composite estimator is more efficient than the synthetic estimator as well as direct estimator.

References

1. Drew, J.D., Singh, M.P. and Choudhury, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. Survey Methodology, 8, p. 17-47.
2. Ferrante, M.R. and Pacei, S. (2004). Small area estimation for longitudinal surveys. Statistical Methods and Applications. 13, p. 327-340.

3. Ghangurde, P.D. and Singh, M.P. (1977). Synthetic estimates in periodic household surveys. Survey Methodology, 3, p. 152-181.

4. Ghosh, M. (1992). Constrained bayes estimation with application. Journal of American Statistical Association, 87, p. 533-540.

5. Ghose, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal. Statistical Science, 9 (1), p. 55-93.

6. Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. Proceedings of the Social Statistics Section. American Statistical Association, p. 33-36.

7. Panse, V.G., Rajagopalan, M. and Pillai, S. (1966). Estimation of crop yields for small areas. Biometrics, 66, p. 374-388.

8. Purcell, N.J. and Kish, L. (1979). Estimation of small domains. Biometrics, 35, p. 365-384.

9. Purcell, N.J. and Kish, L. (1980). Potential estimates for local areas (or domains). International Statistical Review, 48, p. 3-18.

10. Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. Survey Methodology, 25, p. 175–186.

11. Singh, B. and Sisodia, B.V.S. (2001). Small area estimation in longitudinal surveys (M.Sc. Thesis). Department of Agricultural Statistics, NDUAT, Faizabad, INDIA.

12. Singh, D. (1968). Double sampling and its application in agriculture, contributions in statistics and Agricultural Sciences, Presented to Dr. V. G. Panse, Indian Society of Agricultural Statistics, New Delhi.