Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 307–318

On Overlapping Clusters and Multiplicity Sampling

A.K. Srivastava
Indian Agricultural Statistics Research Institute, New Delhi, India

SUMMARY

For overlapping clusters, the selection probabilities provided by the selection procedures get disturbed due to multiplicity of units. Unbiased multiplicity estimator has been considered and its efficiency compared with alternative estimators. A particular case of overlapping clusters in the form of clustering after selection (C.A.S.) by associating nearby units after selecting a 'key -unit', has been considered and the usual estimator formed by considering the clusters as non-overlapping ones is found to perform satisfactorily in terms of efficiency.

Key words : Overlapping clusters, Multiplicity sampling, Key-units.

1. Introduction

Cluster sampling is a well known technique in which clusters are taken as sampling units. Quite often, clusters are available in the form of natural clusters such as households as cluster of individuals or villages as cluster of households. Sometimes, clusters are formed for sample selection such as cluster of villages. In such cases, overlap of clusters is avoided if non-overlapping clusters are to be formed. In natural clusters, if a unit belongs to more than one cluster, overlap is avoided by properly defining the linking rule so that one unit belongs to only one cluster.

Overlapping clusters are sometimes preferred due to constraints at the time of cluster formation. For example, if a map of area to be surveyed is not available, formation of non-overlapping clusters may not be practicable. Formation of clusters by first selecting 'key-villages' and then associating nearby villages to form the clusters, commonly adopted in agricultural surveys, is an example of such overlapping clusters. Various procedures for cluster formation in the form of clustering before selecting the sample (C.B.S.) and after selecting the sample (C.A.S.) were considered by Goel [3], Goel and Singh [4], Aggarwal and Singh [1] etc.

The probability structure of the sampling design, for non-overlapping clusters as sampling units, gets disturbed when the clusters are overlapping. Even for the simplest case of equal probability sampling, if the overlapping

clusters are treated as non-overlapping, the corresponding estimators are biased and the nature and extent of the bias is not known. Changes in the sampling design due to overlapping nature of the clusters may be accounted for, if the corresponding changes in the probability structure may be ascertained. This is feasible but not always simple. Some of these aspects are investigated in this paper for equal probability samples of clusters. Estimation based on multiplicity of units in various clusters is considered in Section 2. Nature and extent of the bias of the usual estimator, considering the clusters as non-overlapping, is examined in Section 3. An alternative approach based on the varying probability sampling is examined in Section 4. Efficiencies of these estimators are considered in Section 5.

2. *Overlapping Clusters and Multiplicity Sampling*

The concept of multiplicity sampling was developed in the context of health surveys for rare diseases (Birnbaum and Sirken [2]). The concept was further extended to other areas (Sirken ([6], [7], [8]), Sirken and Levy [9], Kalton and Anderson [5]). We consider its relevance in the context of overlapping clusters.

Consider a population of L overlapping clusters. Let N be the total number of elements in the population. Define

$\delta_{i\alpha} = 1$ if i^{th} cluster contains α^{th} unit (U_α, say)

$= 0$ otherwise

Thus, $A = ((\delta_{i\alpha}))$ is a $L \times N$ matrix $(i = 1, \ldots L ; \alpha = 1 \ldots N)$

Further, define

$$s_\alpha = \sum_{i}^{L} \delta_{i\alpha} = \text{multiplicity of } U_\alpha$$

$$M_i = \sum_{\alpha}^{N} \delta_{i\alpha} = \text{cluster size for the } i^{th} \text{ cluster}$$

$$R = \sum_{\alpha}^{N} s_\alpha = \sum_{i}^{L} M_i$$

y_α = y – value for U_α

$$\lambda_i = \sum_{\alpha}^{N} y_\alpha \frac{\delta_{i\alpha}}{s_\alpha}$$

It may be observed that

$$\lambda = \sum_{i}^{L} \lambda_i = \sum_{i}^{L} \sum_{\alpha}^{N} y_\alpha \frac{\delta_{i\alpha}}{s_\alpha} = \sum_{\alpha}^{N} y_\alpha = Y$$

Thus, for estimating population total Y, λ may be estimated on the basis of a sample of l clusters selected from L clusters following any sampling design. For example if l clusters are selected by simple random sampling without replacement, a multiplicity based unbiased estimator of population total Y is

$$\hat{Y}_M = \hat{\lambda} = \frac{L}{l} \sum_{i}^{l} \lambda_i \tag{1}$$

Variance of $\hat{Y}_M$ is evidently

$$V(\hat{Y}_M) = L^2 \left(\frac{1}{l} - \frac{1}{L}\right) \frac{1}{L-1} \left(\sum_{i}^{L} \lambda_i^2 - \frac{\lambda^2}{L}\right) \tag{2}$$

$$\sum_{i}^{L} \lambda_i^2 = \sum_{i}^{L} \left(\sum_{\alpha}^{N} \frac{y_\alpha}{s_\alpha} \delta_{i\alpha}\right)^2$$

$$= \sum_{\alpha}^{N} \frac{y_\alpha^2}{s_\alpha} + \sum_{\alpha \neq \beta}^{N} \sum \frac{s_{\alpha\beta}}{s_\alpha s_\beta} y_\alpha y_\beta$$

where $s_{\alpha\beta} = \sum_{i}^{L} \delta_{i\alpha} \delta_{i\beta}$, and

$$\frac{\lambda^2}{L} = \frac{Y^2}{L} = \frac{1}{L}\left[\sum_{\alpha=1}^{N} y_\alpha^2 + \sum_{\alpha \neq \beta}^{N} \sum y_\alpha y_\beta\right]$$

Thus,

$$V(\hat{Y}_M) = L\left(\frac{1}{l} - \frac{1}{L}\right) \frac{1}{L-1} \left[\sum_{\alpha}^{N} y_\alpha^2 \frac{1 - \bar{s}_\alpha}{\bar{s}_\alpha} + \sum_{\alpha \neq \beta}^{N} \sum y_\alpha y_\beta \left(\frac{\bar{s}_{\alpha\beta} - \bar{s}_\alpha \bar{s}_\beta}{\bar{s}_\alpha \bar{s}_\beta}\right)\right] \tag{3}$$

where $\bar{s}_\alpha = s_\alpha / L$ and $\bar{s}_{\alpha\beta} = s_{\alpha\beta} / L$

Following cases are of interest.

Case (i) $M_i = M$ i.e. cluster sizes are equal.

In this case, $\bar{s}_\alpha$ and $\bar{s}_{\alpha\beta}$ satisfy following properties

$$\sum_{\alpha}^{N} \bar{s}_\alpha = \frac{1}{L} \sum_{i}^{L} \sum_{\alpha}^{N} \delta_{i\alpha} = M$$

$$\sum_{\substack{\beta = 1 \\ (\neq \alpha)}}^{N} \bar{s}_{\alpha\beta} = (M-1)\, \bar{s}_\alpha$$

Thus $\bar{s}_\alpha$ and $\bar{s}_{\alpha\beta}$ behave like inclusion probabilities for α^{th} unit and for pair of α^{th} and β^{th} units respectively for a sample of size M from N. For such a sample, the Horvitz and Thompson estimator of population total Y is given by

$$Y^*_{HT} = \sum_{\alpha}^{M} \frac{y_\alpha}{\bar{s}_\alpha}$$

$$V(Y^*_{HT}) = \sum_{\alpha < \beta} \sum (\bar{s}_\alpha\, \bar{s}_\beta - \bar{s}_{\alpha\beta}) \left(\frac{y_\alpha}{\bar{s}_\alpha} - \frac{y_\beta}{\bar{s}_\beta}\right)^2 \qquad (4)$$

In fact $V(\hat{Y}_M)$ from (3), reduces to

$$V(\hat{Y}_M) = \frac{L - l}{l\,(L-1)}\, V(Y^*_{HT}) \qquad (5)$$

$V(\hat{Y}_M)$ consists of two factors. The first factor arises due to selection of l clusters from L clusters and is similar to finite population correction (f.p.c.). The second factor is the variance term for a H.T. estimator and it reduces to zero when y_α is proportional to s_α. In most of the practical situations y_α is likely to be uncorrelated with s_α and $\hat{Y}_M$, though unbiased, is likely to loose efficiency on this account.

Case (ii) L = N

This case corresponds to a practical situation of cluster formation through selection of 'key villages'. Consider a population of N villages. For cluster formation, n villages are first selected and to each selected key village, (M-1)

nearest villages are attached to form the clusters of M villages each. The corresponding multiplicity estimator of Y is given by

$$\hat{Y}_M = \frac{N}{n} \sum_{i}^{n} \lambda_i \tag{6}$$

$$V(\hat{Y}_M) = \frac{N-n}{n(N-1)} \sum_{\alpha<\beta}^{N} \sum (\bar{s}_\alpha \bar{s}_\beta - \bar{s}_{\alpha\beta}) \left(\frac{y_\alpha}{\bar{s}_\alpha} - \frac{y_\beta}{\bar{s}_\beta}\right)^2 \tag{7}$$

where $\bar{s}_\alpha = s_\alpha / N$ and $\bar{s}_{\alpha\beta} = s_{\alpha\beta} / N$

A practical difficulty in the use of $\hat{Y}_M$ is that λ_i should be known for the selected clusters which in turn requires that multiplicity s_α of all the units in the selected clusters should be known. Fortunately this is not difficult in the case of "key village selection procedure". Assuming that for a key-village, all the associated villages fall in a maximum distance of r kms. (say), enumeration of village falling within a distance of 2r kms. will enable one to know about the number of possible clusters in which the elements of the selected clusters can get associated. Thus, the multiplicity of elements belonging to selected clusters can easily be obtained in this procedure.

3. *Bias and Variance of the Usual Estimator in Key Village Selection Procedure*

Consider the case of N clusters. The usual biased estimator of population total is

$$\hat{Y}_b = \frac{N}{n} \sum_{i}^{n} \bar{Y}_i \tag{8}$$

where $\bar{Y}_i = Y_i / M_i$ is the per element cluster mean for the i^{th} cluster. At this stage we consider clusters of unequal sizes. In this case multiplicity is not taken into account.

$$E(\hat{Y}_b) = \sum_{i}^{N} \bar{Y}_i$$

$$= \sum_{i}^{N} \sum_{\alpha}^{N} \frac{y_\alpha \delta_{i\alpha}}{M_i}$$

$$= \sum_{\alpha}^{N} u_\alpha y_\alpha \qquad \text{where } u_\alpha = \sum_{i}^{N} \frac{\delta_{i\alpha}}{M_i}$$

It is observed that $\sum_{\alpha}^{N} u_\alpha = N$. Thus,

$$\text{Bias}(\hat{Y}_b) = \sum_{\alpha}^{N} u_\alpha y_\alpha - \sum_{\alpha}^{N} y_\alpha$$

$$= \sum_{\alpha}^{N} u_\alpha y_\alpha - \frac{\Sigma u_\alpha \Sigma y_\alpha}{N}$$

$$= (N-1) S_{uy} \qquad (9)$$

where $S_{uy} = \frac{1}{N-1} \{ \sum_{\alpha}^{N} u_\alpha y_\alpha - (\sum_{\alpha}^{N} y_\alpha)/N \}$

Thus, bias vanishes if u and y are uncorrelated. When clusters are of fixed size M i.e., $M_i = M$ for all i, then $u_\alpha = s_\alpha / M$. Thus, the usual estimator of total is unbiased when y's and s's are uncorrelated. In case of 'key village' clustering procedures, the multiplicity s's provide measure for density of the area. In actual practice y values and the multiplicity are seldom likely to be correlated. Therefore, in agricultural surveys, the bias introduced by considering overlapping (obtained through 'key village' procedure) clusters as non-overlapping one is hardly of much significance.

Variance of the estimator $\hat{Y}_b$ is given by

$$V(\hat{Y}_b) = N^2 (\frac{1}{n} - \frac{1}{N}) \frac{1}{N-1} \left[\sum_{i}^{N} \bar{Y}_i^2 - \frac{1}{N} (\sum_{i}^{N} \bar{Y}_i)^2 \right]$$

Assuming $M_i = M$, for all i

$$V(\hat{Y}_b) = \frac{N^2 (N-n)}{M^2 n (N-1)} \left[\sum_{\alpha}^{N} y_\alpha^2 \frac{s_\alpha}{N} (1 - \frac{s_\alpha}{N}) + \sum_{\alpha \neq \beta}^{N} \sum y_\alpha y_\beta (\frac{s_{\alpha\beta}}{N} - \frac{s_\alpha s_\beta}{N^2}) \right]$$

$$= \frac{N^2 (N-n)}{M^2 n (N-1)} \left[\sum_{\alpha}^{N} y_\alpha^2 \bar{s}_\alpha (1 - \bar{s}_\alpha) + \sum_{\alpha \neq \beta}^{N} \sum y_\alpha y_\beta (\bar{s}_{\alpha\beta} - \bar{s}_\alpha \bar{s}_\beta) \right]$$

$$= \frac{N^2 (N-n)}{M^2 n (N-1)} \sum_{\alpha} \sum_{< \beta} (\bar{s}_\alpha \bar{s}_\beta - \bar{s}_{\alpha\beta}) (y_\alpha - y_\beta)^2 \qquad (10)$$

4. *Horvitz-Thompson Estimator for Overlapping Clusters*

For overlapping clusters, the probability structure of original sampling scheme of clusters gets disturbed due to multiplicity of sampling. As such even for the equal probability sampling scheme, the resultant sampling scheme is a varying probability scheme. For the key village scheme in which L=N clusters with M elements in each cluster and n clusters being selected by equal probability sampling, usual Horvitz-Thompson estimator for Y is

$$\hat{Y}_{HT} = \sum_{\alpha} y_\alpha / \pi_\alpha \qquad (11)$$

with variance as

$$V(\hat{Y}_{HT}) = \sum_{\alpha} (1 - \pi_\alpha) y_\alpha^2 / \pi_\alpha + \sum_{\alpha \neq \beta} \sum (\pi_{\alpha\beta} - \pi_\alpha \pi_\beta) y_\alpha y_\beta / \pi_\alpha \pi_\beta \qquad (12)$$

where

$$\pi_\alpha = 1 - \prod_{t=1}^{s_\alpha} \{1 - \frac{n}{N-t+1}\} \qquad (13)$$

$$\pi_{\alpha\beta} = 1 - \prod_{t=1}^{s_{\alpha\beta}} \{1 - \frac{n}{N-t+1}\} \qquad (14)$$

It may be noted that if v is the number of distinct elements in the sample then

$$E(v) = \sum_{\alpha} \pi_\alpha \qquad (15)$$

5. *Efficiency Comparisons*

5.1 *Overlapping versus non-overlapping clusters : a model based comparison*

It is seen that the case of multiplicity sampling reduces to non-overlapping clusters when

$$s_\alpha = 1$$

and $\delta_{i\alpha}\,\delta_{i\beta} = 1, \quad$ if $\alpha, \beta \in i^{th}$ cluster $(i=1, \ldots L)$

$= 0$, otherwise

Thus, the estimator of Y in case of non-overlapping clusters may be compared with the usual $\hat{Y}_b$ when overlapping clusters were considered as non-overlapping clusters. We consider the following model

$$y_\alpha = \mu + \varepsilon_\alpha$$

where $E(\varepsilon_\alpha) = 0, \quad E(\varepsilon_\alpha^2) = \sigma^2 \quad$ and $\quad E(\varepsilon_\alpha\, \varepsilon_\beta) = 0$

This model is suitable under the situation when y's are uncorrelated with multiplicity of U_α. Under this model, $EV(\hat{Y}_b)$ reduces to

$$EV(\hat{Y}_b) = \frac{(N-n)}{n(N-1)} \frac{N^2}{M^2} \{ M - \sum^{N} \bar{s}_\alpha^2 \}\, \sigma^2$$

To compare $\hat{Y}_b$ with the non-overlapping case, consider $L = \frac{N}{M}$ non-overlapping clusters from which l = pn clusters are selected where p is a constant. Expected variance of $\hat{Y}_{NOC}$ reduces to

$$EV(\hat{Y}_{NOC}) = L\left(\frac{1}{pn} - \frac{1}{L}\right) \frac{1}{L-1} \frac{N(N-M)}{M}\, \sigma^2$$

$$= \frac{N - pnM}{pnM}\, N.\sigma^2$$

Now, a little simplification shows that

$$EV(\bar{Y}_b) \gtreqless EV(\hat{Y}_{NOC}) \text{ according as}$$

$$\sum_{\alpha}^{N} \bar{s}_\alpha^2 \gtreqless M\left\{1 - \frac{(N-1)(N-pnM)}{(N-n)Np}\right\}$$

Following cases are of interest

Case (i) Sampling fractions of clusters are same

$$\frac{L}{l} = \frac{N}{n}$$

or $\quad l = \frac{L}{N}\, n$

$$\text{or} \qquad p = \frac{L}{N} = \frac{1}{M}$$

For this value of p,

$$M\left\{1 - \frac{(N-1)\ (N-pnM)}{(N-n)\,Np}\right\} = M\left\{1 - \frac{(N-1)\,M}{N}\right\} < 0, \text{ for } M \geq 2$$

Thus, $\qquad EV(\hat{Y}_b) < EV(\hat{Y}_{NOC})$

Case (ii) Expected number of cluster elements are same

$$lM = E(v)$$

where $E(v)$ is obtained from (13), (14) and (15). For $p = E(v)/nM$

$$M\left\{1 - \frac{(N-1)\,(N-pnM)}{(N-n)\,Np}\right\} = M\left\{1 - \frac{(1-\frac{1}{N})\ (\frac{1}{E(v)} - \frac{1}{N})\ M}{(\frac{1}{n} - \frac{1}{N})}\right\}$$

$$\doteq M\left\{1 - \frac{nM}{E(v)}\right\} \quad \text{for large N}$$

$$\leq 0 \quad \text{as } E(v) \leq nM$$

Thus, in this case $\qquad EV(\hat{Y}_b) \leq EV(\hat{Y}_{NOC})$

Case (iii) Number of clusters are same

$$l = n$$

or $\qquad p = 1$

$$EV(\hat{Y}_b) \gtreqless EV(\hat{Y}_{NOC}) \text{ according as}$$

$$\sum_{\alpha}^{N} \bar{s}_{\alpha}^2 \gtreqless M\left\{1 - \frac{(N-1)\,(N-nM)}{(N-n)\,N}\right\}$$

This condition indicates that for p=1, $\hat{Y}_{NOC}$ is likely to be better than $\hat{Y}_b$. In fact, if some idea about the expression $\Sigma\, \bar{s}_{\alpha}^2$ is available then one can know about values of p (and thereby l) for which $\hat{Y}_{NOC}$ may be superior or inferior to $\hat{Y}$.

5.2 *An empirical comparison*

The population in this illustration consists of 25 villages in a compact area of Rampura phul tehsil of Bhatinda district in Punjab. The overlapping clusters of sizes 3 each were formed following 'key village procedure' from the map for the above mentioned tehsil from District Census Handbook, 1981. The information for each villages in respect of total area of the village $y_{(1)}$, number of households $y_{(2)}$, irrigated area $y_{(3)}$ and multiplicity s_α is given in Table 1.

Samples of five overlapping clusters are considered for comparison purposes. Mean square errors (M.S.E.) and variances of estimators $\hat{Y}_b$, $\hat{Y}_M$ and $\hat{Y}_{HT}$ for the three characters $y_{(1)}$, $y_{(2)}$ and $y_{(3)}$ are presented in Table-2.

Evidently $\hat{Y}_b$ is more efficient as compared to $\hat{Y}_M$ which is better than $\hat{Y}_{HT}$ in this illustration.

Table 1. Clusters for key villages and corresponding $y_{(1)}$, $y_{(2)}$, $y_{(3)}$ and s_α

Key Vill. No.	Cluster			$y_{(1)}$	$y_{(2)}$	$y_{(3)}$	s_α
1	1	2	3	468	197	427	3
2	2	1	4	465	184	398	3
3	3	1	4	1018	402	892	4
4	4	2	3	584	161	366	3
5	5	3	8	1117	372	923	2
6	6	7	8	2014	703	838	3
7	7	6	8	837	299	315	5
8	8	5	7	706	253	584	5
9	9	8	10	402	145	218	2
10	10	9	11	608	220	561	4
11	11	7	10	2161	106	1624	2
12	12	10	13	1295	528	755	2
13	13	12	14	331	149	283	3
14	14	13	15	2094	653	1666	5
15	15	14	16	515	192	374	2
16	16	7	14	2992	1099	1893	2
17	17	19	23	2161	562	711	1
18	18	19	20	679	215	377	2
19	19	18	20	1539	482	1267	4
20	20	21	22	353	90	243	5
21	21	20	22	595	135	454	2
22	22	19	20	534	225	288	3
23	23	14	24	2018	609	522	3
24	24	23	25	4564	360	2892	3
25	25	6	24	2033	535	1768	2

Table 2. M.S.E. and variances $\hat{Y}_b$, $\hat{Y}_M$ and $\hat{Y}_{HT}$ for $y_{(1)}$, $y_{(2)}$ and $y_{(3)}$

MSE/Variance	$y_{(1)}$	$y_{(2)}$	$y_{(3)}$
M.S.E. $(\hat{Y}_b)$	$.1016 \times 10^8$	$.5268 \times 10^6$	$.4851 \times 10^7$
V $(\hat{Y}_M)$	$.6563 \times 10^9$	$.4168 \times 10^8$	$.3460 \times 10^9$
V $(\hat{Y}_{HT})$	$.2990 \times 10^{10}$	$.2409 \times 10^9$	$.1294 \times 10^{10}$

REFERENCES

[1] Aggarwal, D.K. and Singh, P., 1982. On cluster sampling strategies using ancillary information. *Sankhya*, **B44,** 184-192.

[2] Birnbaum, Z.W. and Sirken, M.G., 1965. Design of sample surveys to estimate the prevalance of rare diseases. Three unbiased estimates. *National Centre for Health Statistics*, Series 2, **11**, 1-8.

[3] Goel, B.B.P.S., 1973. Efficiency of certain systems of cluster sampling and its applications. Unpublished thesis for Ph.D., I.A.R.I., New Delhi.

[4] Goel, B.B.P.S. and Singh, D., 1977. On the formation of clusters. *J. Indian Soc. Agric. Statist.*, **29**, 53-68.

[5] Kalton, G. and Anderson, D., 1986. Sampling rare populations. *J. Roy. Statist. Soc.*, **A149**, 65-82.

[6] Sirken, M.G., 1970. Household surveys with multiplicity. *J. Amer. Statist. Assoc.*, **67**, 224–227.

[7] Sirken, M.G., 1972. Variance components of multiplicity estimators. *Biometrics*, **28,** 869-873.

[8] Sirken, M.G., 1974. The counting rule strategy in sample surveys. *Proc. Social Statist. Sec., J. Amer. Statist. Assoc.*, 119-123.

[9] Sirken, M.G. and Levy, P.S., 1974. Multiplicity estimation of proportion based on ratios of random variables. *J. Amer. Statist. Assoc.*, 68-73.