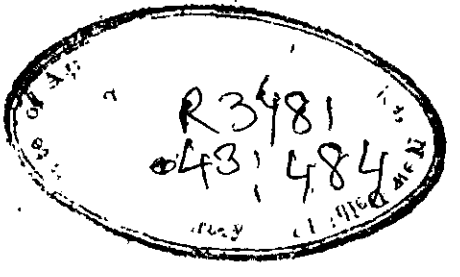
U 95

ORDERED CLUSTER SAMPLING

G. C. CHAWLA



DE-04

C.

INSTITUTE OF AGRICULTURAL RESEARCH STATISTICS
(I.C.A.R.)
NEW DELHI - 12

ORDERED CLUSTER SAMPLING

G. C. CHAWLA

Dissertation submitted in fulfilment of the
requirements for the Award of Diploma
in Agricultural and Animal Husbandry
Statistics, of the Institute of
Agricultural Research
Statistics
(I. C.A. R.)
New Delhi-12

1974

ACKNOWLEDGEMENTS

I have great pleasure in expressing my deep sense of gratitude to Shri M. Rajagopalan, Statistician-cum-Associate Professor, Institute of Agricultural Research Statistics, New Delhi, for suggesting the problem, his valuable guidance, keen interest and constant encouragement during the course of investigation and of preparation of this thesis.

I am also highly grateful to Dr. D. Singh, Director, Institute of Agricultural Research Statistics, Indian Council of Agricultural Research for providing me with adequate research facilities for this work.

Lastly, I take opportunity to thank Shri Prem Kumar for having typed my thesis.

(G. C. CHAWLA)

CONTENTS

CHAPTER - I

INTRODUCTION

1.1. Cluster Sampling: Sample survey techniques are intended to estimate the unknown measures of characters of a population. The elements constituting a sample from the population on which the informati is collected, are selected by probability sampling techniques. Sometimes controls like stratification are also imposed before the sample is drawn. The units in the sample may be selected with equal or unequal probabilities. Depending upon the availability of auxiliary information ratio or regression methods are used to improve the performance of the estimator. In many situations the sample units could be approached only through a multi-stage sampling design.

All these techniques pre-suppose that the elements in the population are themselves the sampling units. However, some situations arise in survey practice, where due to operational or other consideration the information cannot be obtained from each individual element of the population but from only a group of elements put together. Under such situations, the Statistician has no choice but to treat these groups of elements as sampling units. These groups of elements are termed as clusters of elements and the procedure of sampling clusters instead of elements is termed as cluster sampling.

Choice of selecting clusters of elements is unavoidable in many of the situations encountered in survey practices. These clusters could have been formed into well defined groups for which the frame is

available. For instance, in household consumption surveys, it is natural and convenient to select the households as sampling units instead of selecting individuals constituting the household, even though interest lies in estimating the main performance of characters per individual of the population. In these situations it is not possible to list out all the individuals in the population and observe the characteristic on them. It is for more convenient to select the household and it may generally be possible to obtain information for the entire household rather than for each individual of household. Similar is the case for certain characters such as milk production and egg production in the rural areas where it could be possible to obtain information from the group of animals in the household rather than to obtain the production record in each of the animal in the household.

When natural clusters are not available, Statistician is constrained to form artificial clusters of the elements for operational convenience and for reduction in the cost of the survey. The Statistician is concerned not only with the problem of choosing the best design which may provide the estimate with least possible error but also concerned with the economic ways to arrive at these estimates. Therefore, he has to strike a balance between the variance of the estimate and also the cost of the survey, either minimising the one for fixing the other component or vice - versa. Towards this end he forms clusters on his own. He may specify the dimension of the plot which is a group of plants in case of field crop surveys or he may choose a specified number of

neighbouring trees for estimating production in case of horticulturing crop or he may choose a set up of neighbouring villages or neighbouring households in case of characters like milk production and egg production etc. The object of this clustering is obviously to reduce the time of enumeration of the elements and there-by to reduce the cost of the survey. So by adopting cluster sampling design we can over-come administrative difficulties, intervals of time milking operations and effective supervision by taking adjacent villages.

1.2. Over-lapping Clusters: Some times it is easy to demarcate the clusters of elements in the population and the list of such clusters would be used for sampling. For example, a list of households could easily be obtained rather than the list of persons in the household. Such clusters as are already existing in the population can be termed as natural clusters. But in many situations clusters are to be formed by some criteria, by grouping elements of the population. This could be termed as artificial clusters. For instance, if a map of villages is available, clusters of villages of a given size could be formed in such a way that they are adjacent or within a specified distance from each other. The formation and number of such clusters very much depend upon the convenient starting point for the formation of clusters. In such a situation, an element of the population may belong to more than one such artificial cluster. Such clusters are termed as over-lapping clusters.

1.3. Review of Work Done in Non-over-lapping Clusters: The current theory on cluster sampling basically is developed for non-over-lapping clusters or when the frame of clusters is already available before selection whether the clusters are of equal or of unequal size. When the population consists of N clusters of M elements each and a sample of n clusters is drawn from it by the method of simple random sampling (SRS), the relative efficiency of a cluster as the unit of sampling compared with that of an element is given by

$$RE = \frac{S^2}{M S_b^2} \qquad \ldots\ldots (1.3.1)$$

where $S^2 = \frac{1}{NM-1} \sum_{i}^{N} \sum_{j}^{M} (y_{ij} - \bar{Y}_{..})^2$ = Mean sum of squares between elements of the population and

$S_b^2 = \frac{1}{N-1} \sum_{i}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2$ = Mean sum of squares between means of clusters,

y_{ij} = Value of the character for the j-th element ($j = 1, 2, \ldots, M$) in the i-th cluster ($i = 1, 2, \ldots, N$),

$\bar{y}_{i.} = \frac{1}{M} \sum_{j}^{M} y_{ij}$ = mean per element for the i-th cluster,

$\bar{Y}_{..} = \frac{1}{NM} \sum_{i}^{N} \sum_{j}^{M} y_{ij}$ = mean per element in the population.

It can be seen from (1.3.1) that the efficiency of cluster sampling increases as the mean square between clusters decreases and the cluster sampling will be more efficient if the clusters formed are such that the variation between cluster means is as small as possible while the variability within cluster $\bar{S}_w^2$ is as large as possible. If the clusters are formed by grouping together random samples of M elements from a population of NM elements then

$$E(M S_b^2) = S^2$$

In that case cluster sampling and simple random sampling of elements are of equal efficiency.

The efficiency of cluster sampling is also expressed in terms of the intra - class correlation ' ρ ' between elements of a cluster and is given by

$$RE = \frac{M(N-1)}{NM-1} \cdot \frac{1}{[1 + \overline{M-1}\,\rho]} \qquad \ldots\ldots (1.3.2)$$

$$\doteq \frac{1}{[1 + \overline{M-1}\,\rho]} \qquad \ldots\ldots (1.3.3)$$

when N is large.

Usually ρ is positive because the units within clusters are likely to be more alike in character expression and hence cluster sampling is generally less efficient than sampling of elements by simple

random sampling.

From (1.3.3) it is clear that the efficiency of cluster sampling not only depends on ρ but also on the cluster size M. Several workers such as Fair-field Smith (1938), Mahalanobis (1940) and Jessen (1942) have carried out investigations to find out the functional relationship between the size of the cluster S_b^2 and $\bar{S}_w^2$.

Fair-field Smith obtained the relationship

$$S_b^2 = \frac{g^2}{M^g}$$

between the mean square between clusters S_b^2 and cluster size M, where g is a constant less than one.

But most economic characteristics relating to farm data follow a slightly different law. Mahalanobis (1940) and Jessen (1942) postulated that the mean square among elements within a cluster is a monotone increasing function of the size of the cluster which is given by

$$\bar{S}_w^2 = a M^b \quad (b > 0)$$

where a and b are constants to be evaluated from the data.

Hendricks (1944), however, pointed out that the law may not hold good for large sizes of clusters. This was supported by the findings of Asthana (1950).

As stated earlier, cluster sampling is generally less efficient than straight selection of elements by simple random sampling, if

variances of the estimators are only considered. But from the operational and cost considerations cluster sampling scores over simple random sampling. This is because of the fact that less cost would be involved in travel between elements in the form of clusters than when they are spread out. Thus the efficacy of cluster sampling has to be studied only with the cost involved.

An appropriate cost function for cluster sampling could be written as

$$C = C_1 n M + C_2 d$$

where C_1 is the cost of enumeration per element including the travel cost from one element to another within the cluster, C_2 that of travelling a unit distance between clusters and 'd' the total distance between clusters. Mahalanobis (1940) showed empirically that the expected value of the minimum distance between n points located at random is proportional to $(\sqrt{n} - \frac{1}{\sqrt{n}})$. Jessen (1942) showed that the approximation $\sqrt{n}$ works well in practical situations. Then the cost function takes the form

$$C = C_1 n M + C_2 \sqrt{n}$$

From the variance and cost function optimum cluster size and sample of clusters could be determined and the solutions are given by Sukhatme, P. V. (1950).

Many recent developments have taken place in cluster sampling

Ghosh (1963) studied the problem of post cluster sampling i.e. selection of clusters when the structure of clusters is not known in advance. Sethi (1965) discussed the problem of forming optimum clusters of two units when information on study variate is available for estimating the population total.

Mishra and Sukhatme (1972) explained some conditions under which cluster sampling in connection with ratio or regression method of estimation is more efficient than simple random sampling even if the intra-class correlation is positive. All these above studies pertain to the situation where clusters are distinct or non-over-lapping.

We come across many instances of the use of non-over-lapping clusters in sample survey practices, though much theoretical work has not been carried out in this direction. Grid sampling defined by Mahalanobis and others for collecting the statistics of area and yield of crops is a well known example of this type. He explained that bias introduced in estimation procedure in over-lapping clusters (grids) is not significant. For estimation livestock products Singh, D., Rajagopalan, M. and Maini, J.S. (1970) and Panse, Singh and Murthy (1964, 66) ; Singh, Murthy and Goel (1970) have mentioned some techniques for formation of clusters at various stages of sampling. Murthy (1967) made a reference about over-lapping clusters but did not discuss about the nature of bias.

Goel (1973) discussed in detail the criteria for forming clusters as encountered in survey practices. The criteria are that

the clusters should be as homogeneous as possible and the average travel cost between elements in a cluster should be small as compared to average cost of travel between clusters. With these in view he suggested two systems of cluster formation, namely, clustering before sampling (CBS) and clustering after sampling (CAS). CBS system consists in serially listing the elements in a certain order, such as their location etc., starting from a suitable point and then forming clusters of a given size M by grouping M contiguous elements each. This system leads to non-over lapping clusters. But this system is found to be generally less efficient than selecting of natural clusters or sampling of individual elements. CAS consists in first selecting a sample of elements at random from the population and then combining with each of these (M-1) neighbouring elements. This system leads to over- lapping clusters and the estimate of population mean is generally biased.

In the sample surveys conducted by Institute of Agricultural Research Statistics, CAS system is followed with certain modifications. A random sample of villages are selected at first, and the enumerator is asked to list out all the villages within a specified distance for each such village and select a random sample of (M - 1) villages from them and form a cluster of M villages together with the originally selected village. In this dissertation, this problem is attempted and an unbiased estimate of the population total is obtained together with the estimate of its variance, when clusters of size 2 are formed by this procedure.

CHAPTER - II

STATEMENT OF THE PROBLEM AND NOTATIONS

2.1. Ordered Cluster Sampling: In the last chapter the difficulty that arises in practice, in adopting cluster sampling when they are non-over-lapping has been indicated. In many of the surveys, the enumerator is given a list of sample villages which are termed as main villages and around each of which he is asked to list out all the villages within a 4 or 5 miles radius and asked to choose one or more villages at random to form a cluster of specified size with the main village that has already been allotted. This procedure is adopted mainly because no information could possibly be obtained about the distances of the villages from each other before the commencement of the survey work. In the absence of such information no satisfactory procedure of estimation could be developed. However, this procedure of cluster formation could be modified in the following way.

Select a main village at random and form the cluster according to the criterian specified above, then select the second main village and form a cluster with the same criterian and proceed successively for selection for all the specified sample of main villages thus selecting every cluster without replacement. This procedure could be termed as ordered cluster sampling. The main advantage of this method as can be seen in the later section is that the number of villages at a specified distance need not be known before hand for all the villages in the population. It would be enough to know the villages within a

specified distance only for the main villages selected and for the other village constituting the cluster.

2.2. Notations : Let the finite population consist of N distinct and identifiable units specified as $U_1, U_2, \ldots, U_N$. Let U_i be the unit selected at random by simple random sampling at the first draw. Let the distance specified be 'd' miles and there be M_i number of villages within the distance 'd' from U_i.

Units selected at first draw	No. of units within the distance 'd'
U_1	M_1
U_2	M_2
.	.
.	.
.	.
U_N	M_N

It is required to form clusters of equal size '2'. The next step is to select 1 unit at random from M_i villages. It is clear that at first draw the unit U_i has a probability of selection $\frac{1}{N}$. If M_i the number of units within a distance 'd' from U_i, the number of ways of selecting 1 unit out of M_i is $\frac{1}{M_i}$. The probability of simultaneous occurance of U_i and one more unit is given by

$$\frac{1}{N} \cdot \frac{1}{M_i}$$

and so probabilities of forming clusters of size '2' at the first draw from the N units of the population are given as below:

$$P_1 = \frac{1}{N} \cdot \frac{1}{M_1}$$

$$P_2 = \frac{1}{N} \cdot \frac{1}{M_2}$$

$$\vdots$$

$$P_i = \frac{1}{N} \cdot \frac{1}{M_i}$$

$$\vdots$$

$$P_N = \frac{1}{N} \cdot \frac{1}{M_N}$$

Let the i-th cluster consist of the units U_{i_1}, U_{i_2}. If the i-th unit of the population is selected at the first draw, it is associated with other 1 unit. One of these 2 units could have been drawn at the first draw. Thus the probability of selecting the specified cluster at the first draw is the sum of the probabilities of selecting these 2 units at the first draw. Let this probability be denoted by

$$P_i^{(1)} = \sum_{i=1}^{2} P_i$$

where summation is from 1 through 2 (not the first 2 probabilities)

Consider the possible clusters formed are $C_1^{(1)}, C_2^{(1)}, \ldots, C_{L_1}^{(1)}$ with the corresponding probabilities of formation as

$$P_1^{(1)}, P_2^{(1)}, \ldots, P_{L_1}^{(1)}$$

It is clear that $\sum_i^{L_1} P_i^{(1)} = 1$ (condition of probability)

The probabilities of units included in the clusters can be denoted as

$$v_1^{(1)}, v_2^{(1)}, \ldots, v_N^{(1)}$$

where

$$v_i^{(1)} = \sum_i P_i^{(1)} \qquad \ldots (2.2.1)$$

Summation is taken over all clusters where i unit occurs. The object is to estimate the population total $Y = \sum_i^N y_i$, where y_i be the i-th characteristic value under study for the unit U_i.

For the given first draw, we define

$$t_1 = \sum_i^2 \frac{y_i}{v_i^{(1)}}$$

Then we observe that

$$E(t_1) = E\left[\sum_i^2 \frac{y_i}{v_i^{(1)}}\right]$$

$$= \sum_j^{C_{L_1}^{(1)}} \left[\sum_i^2 \frac{y_i}{v_i^{(1)}}\right] \cdot P_j^{(1)}$$

$$= \sum_j^{C_{L_1}^{(1)}} \left[\frac{y_1}{v_1^{(1)}} + \frac{y_2}{v_2^{(1)}}\right] \cdot P_j^{(1)}$$

$$= \left[\frac{y_1}{\pi_1^{(1)}} + \frac{y_2}{\pi_2^{(1)}}\right]_1 \cdot P_1^{(1)} + \left[\frac{y_1}{\pi_1^{(1)}} + \frac{y_2}{\pi_2^{(1)}}\right]_2 \cdot P_2^{(1)} + \cdots$$

$$+ \cdots + \left[\frac{y_1}{\pi_1^{(1)}} + \frac{y_2}{\pi_2^{(1)}}\right]_{C_{L_1}^{(1)}} \cdot P_{C_{L_1}^{(1)}}^{(1)}$$

$$= \frac{y_1}{\pi_1^{(1)}} \cdot \sum_{i(1)} P_i^{(1)} + \frac{y_2}{\pi_2^{(1)}} \cdot \sum_{i(2)} P_i^{(1)} + \cdots + \frac{y_i}{\pi_i^{(1)}} \cdot \sum_{i(i)} P_i^{(1)}$$

$$+ \cdots + \frac{y_N}{\pi_N^{(1)}} \cdot \sum_{i(N)} P_i^{(1)}$$

where each Σ on R.H.S. is taken wherever 1, 2, 3, ... , N units occurs and so by (2.2.1) we get

$$E(t_1) = y_1 + y_2 + \cdots + y_N$$

(since each sum is cancelled away with the corresponding denominator). Thus t_1 is an unbiased estimate of Y.

Now for the second draw, we remove from the list of clusters, the clusters selected at the first draw. Now we have (N-2) units. We proceed to form clusters, probabilities and inclusion probabilities on the similar lines as above. The units are relabelled and we find the associates of key-village within a distance 'd' miles from it.

Let the clusters be

$$C_{1'}^{(2)},\ C_{2'}^{(2)},\ \ldots,\ C_{L_2'}^{(2)}$$

with their specified probabilities

$$P_{1'}^{(2)},\ P_{2'}^{(2)},\ \ldots,\ P_{L_2'}^{(2)}$$

so for the second draw, we have

$$\sum_{1'}^{L_2'} P_{1'}^{(2)} = 1$$

Inclusion probabilities are

$$\pi_{1'}^{(2)},\ \pi_{2'}^{(2)},\ \ldots,\ \pi_{(N-2)'}^{(2)} \quad \text{where}$$

$$\pi_{1'}^{(2)} = \sum_{1'} P_{1'}^{(2)}\ .$$

Summation is taken over those clusters where ($1' \neq 1$) occurs. Now select a cluster say $C_{1'}^{(2)}$, we define

$$t_2 = \sum_{1}^{2} y_1 + \sum_{1'}^{2} \frac{y_{1'}}{\pi_{1'}^{(2)}}$$

where $y_{1'}$ is the character value for $1'$-th unit ($1' \neq 1$) which could be any of the units in the (N - 2) units of the remaining population. Now

$$E(t_2) = \sum_{1}^{2} y_1 + E \sum_{1'}^{2} \frac{y_{1'}}{\pi_{1'}^{(2)}}$$

On the lines discussed above, it can be shown that $E \sum_{1'}^{2} \frac{y_{1'}}{\pi_{1'}^{(2)}}$ is

an unbiased estimate of the total of (N- 2) units = Y' (say), where

$$Y' = Y - \sum_i y_i = \text{Population total} - \text{Total units in the first cluster drawn.}$$

Therefore, the expected value of t_2 having drawn the first cluster is given by

$$E(t_2 / t_1) = \sum_i y_i + Y' = Y \text{ (Population total)}.$$

Hence the combined estimate of $\hat{Y}$ based on both the draws is given by

$$\hat{Y} = \frac{1}{2}(t_1 + t_2) = \frac{1}{2}(Y + Y) = Y$$

Which is an unbiased estimate of the population total. In general, the cluster selected at r-th draw, gives

$$t_r = \sum_{i}^{r-1} \bar{y}_i + \sum_i^2 \frac{y_i}{\pi_i^{(r)}}$$

Then it can be seen as before that $E(t_1) = Y$ and

$$E(t_r / t_1 , t_2 , \ldots , t_{r-1}) = Y$$

Hence

$$E(t_r) = Y \text{ (Population total)}$$

For n draws we get similar values combining all estimates of $t_1 , t_2 , \ldots , t_n$ is given by

$$\hat{Y} = \frac{1}{n} \sum_{r=1}^{n} t_r$$

which is an unbiased estimate of the population total. Again we observe for r less than k, then

$$E\left[t_r\, t_k\right] = E\left[t_r \left\{ E\left(t_k \,/\, \text{having drawn } k\text{-th cluster prior to } r\text{-th cluster} \right)\right\}\right]$$

$$= E\left[t_r\, Y\right] = Y^2$$

Therefore, we have

$$E\left[\frac{\sum_{r \neq k}^{n} t_r \cdot t_k}{n(n-1)}\right] = Y^2$$

$$\text{Est.}\left(Y^2\right) = \frac{1}{n(n-1)} \sum_{r \neq k}^{n} t_r \cdot t_k$$

$$V(\hat{Y}) = E\left(\hat{Y}^2\right) - Y^2$$

$$\therefore\ \hat{V}(Y) = \hat{Y}^2 - \text{Est. } Y^2$$

$$= \frac{1}{n^2}\left(\sum_{r}^{n} t_r\right)^2 - \frac{1}{n(n-1)} \sum_{r \neq k}^{n} t_r \cdot t_k$$

$$= \frac{1}{n^2}\left[\sum_{r}^{n} t_r^2 + \sum_{r \neq k}^{n} t_r \cdot t_k\right] - \frac{1}{n(n-1)} \sum_{r \neq k}^{n} t_r \cdot t_k$$

$$= \frac{1}{n^2} \sum_{r}^{n} t_r^2 + \left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right) \sum_{r \neq k}^{n} t_r \cdot t_k$$

$$= \frac{1}{n^2} \sum_{r}^{n} t_r^2 - \frac{1}{n^2(n-1)} \sum_{r \neq k}^{n} t_r \cdot t_k$$

$$= \frac{1}{n^2} \sum_{r}^{n} t_r^2 - \frac{1}{n^2(n-1)} \left[\left(\sum_{r}^{n} t_r \right)^2 - \sum_{r}^{n} t_r^2 \right]$$

$$= \left[\frac{1}{n^2} + \frac{1}{n^2(n-1)} \right] \sum_{r}^{n} t_r^2 - \frac{1}{n^2(n-1)} \left(\sum_{r}^{n} t_r \right)^2$$

$$= \frac{1}{n(n-1)} \sum_{r}^{n} t_r^2 - \frac{1}{n^2(n-1)} \cdot n^2 \hat{Y}^2 \quad \because \hat{Y} = \frac{1}{n} \sum_{r}^{n} t_r$$

$$= \frac{1}{n(n-1)} \left[\sum_{r}^{n} t_r^2 - n \hat{Y}^2 \right]$$

$$= \frac{1}{n(n-1)} \sum_{r=1}^{n} (t_r - \hat{Y})^2$$

CHAPTER - III

ILLUSTRATION

The methodology which is developed in Chapter -II is illustrated with the help of data on cattle population from Velvadam. Firka (Vijayawada Taluk), Krishna Delta Area, Andhra Pradesh, which consists of 17 villages. The map showing the location of these villages is given in Appendix - I. The names of villages with their code numbers are also given in the Appendix - II.

It is proposed to select a sample of two clusters of 2 villages each such that the distance between the villages in a cluster is not more than 4 miles. The list of associated villages satisfying this criterian for each of the 17 villages is given below and the probabilities attached with each of the 17 clusters of 2 villages each to be drawn at the first draw are as follows:

TABLE - 3.1

Sl. No.	Code No. of villages	Cattle population	Associated villages within a distance of 4 miles		Probabilities $[P_i]$
			Codes of villages	Number	
(1)	(2)	(3)	(4)	(5)	(6)
1	18	85	19,20,25,26,27,33, 34	7	0.0084
2	19	1184	18,20,21,22,25,26	6	0.0098
3	20	299	18,19,21,22,25,26,27	7	0.0084
4	21	51	19,20,22,25	4	0.0147
5	22	690	19,20,21,25	4	0.0147

contd...

table -3.1 (contd.)

(1)	(2)	(3)	(4)	(5)	(6)
6	23	2561	24	1	0.0588
7	24	506	23,25	2	0.0294
8	25	1070	18,19,20,21,22,24, 26,27	8	0.0074
9	26	85	18,19,20,25,27,28, 33,34	8	0.0074
10	27	417	18,20,25,26,28,29, 31,32,33,34	10	0.0059
11	28	320	26,27,29,30,31,32, 33	7	0.0084
12	29	138	27,28,30,31,32,33, 34	7	0.0084
13	30	000	28,29,31,32,33	5	0.0118
14	31	000	27,28,29,30,32,33	6	0.0098
15	32	506	27,28,29,30,31,33	6	0.0098
16	33	159	18,26,27,28,29,30, 31,32,34	9	0.0065
17	34	1166	18,26,27,29,33	5	0.0118

Total Cattle Population = 9237.

Since the probability of selecting a village of code number 18 is $\frac{1}{17}$. Its associates within a distance of four miles and seven in number given in the first row of fourth column (Table -3.1) and so the probability of selecting any one code number is $\frac{1}{{}^{7}C_{1}} = \frac{1}{7}$. Hence the probability of forming a cluster of two villages each is given by

$$P_1 = \frac{1}{17} \cdot \frac{1}{7} = \frac{1}{119} = 0.0084 \text{ (given in sixth column).}$$

Similarly the other probabilities are calculated which are given in sixth column for all the seventeen villages.

From the above points the following clusters of size two satisfying the distance criterion are listed below together with their probabilities of formation and inclusion probabilities.

TABLE - 3.2

Sl. No.	Initial village	Clusters of two villages with code numbers	Probabilities $[P_i^{(1)}]$	Inclusion Probabilities $[\pi_i^{(1)}]$
(1)	(2)	(3)	(4)	(5)
1	18	(18, 19)	0.0182	
2	18	(18, 20)	0.0168	
3	18	(18, 25)	0.0158	
4	18	(18, 26)	0.0158	0.1160
5	18	(18, 27)	0.0143	
6	18	(18, 35)	0.0149	
7	18	(18, 34)	0.0202	
8	19	(19, 18)	0.0182	
9	19	(19, 20)	0.0182	
10	19	(19, 21)	0.0245	0.1198
11	19	(19, 22)	0.0245	
12	19	(19, 25)	0.0172	
13	19	(19, 26)	0.0172	

contd...

table - 3.2 (contd.)

(1)	(2)	(3)	(4)	(5)
14	20	(20, 18)	0.0168	
15	20	(20, 19)	0.0182	
16	20	(20, 21)	0.0231	
17	20	(20, 22)	0.0231	0.1271
18	20	(20, 25)	0.0158	
19	20	(20, 26)	0.0158	
20	20	(20, 27)	0.0149	
21	21	(21, 19)	0.0245	
22	21	(21, 20)	0.0231	0.0991
23	21	(21, 22)	0.0294	
24	21	(21, 25)	0.0221	
25	22	(22, 19)	0.0245	
26	22	(22, 20)	0.0231	0.0991
27	22	(22, 21)	0.0294	
28	22	(22, 25)	0.0221	
29	23	(23, 24)	0.0882	0.0882
30	24	(24, 23)	0.0882	0.1250
31	24	(24, 25)	0.0368	

contd...

table -3.2 (contd.)

(1)	(2)	(3)	(4)	(5)
32	25	(25, 18)	0.0158	
33	25	(25, 19)	0.0172	
34	25	(25, 20)	0.0158	
35	25	(25, 21)	0.0221	0.1579
36	25	(25, 22)	0.0221	
37	25	(25, 24)	0.0368	
38	25	(25, 26)	0.0148	
39	25	(25, 27)	0.0133	
40	26	(26, 18)	0.0158	
41	26	(26, 19)	0.0172	
42	26	(26, 20)	0.0158	
43	26	(26, 25)	0.0148	0.1258
44	26	(26, 27)	0.0133	
45	26	(26, 28)	0.0158	
46	26	(26, 33)	0.0139	
47	26	(26, 34)	0.0192	
48	27	(27, 18)	0.0143	
49	27	(27, 20)	0.0143	
50	27	(27, 25)	0.0133	
51	27	(27, 26)	0.0133	
52	27	(27, 28)	0.0143	0.1453
53	27	(27, 29)	0.0143	

contd...

table - 3.2 (contd.)

(1)	(2)	(3)	(4)	(5)
54	27	(27, 31)	0.0157	
55	27	(27, 32)	0.0157	
56	27	(27, 33)	0.0124	
57	27	(27, 34)	0.0177	
58	28	(28, 26)	0.0158	
59	28	(28, 27)	0.0143	
60	28	(28, 29)	0.0168	
61	28	(28, 30)	0.0202	0.1184
62	28	(28, 31)	0.0182	
63	28	(28, 32)	0.0182	
64	28	(28, 33)	0.0149	
65	29	(29, 27)	0.0143	
66	29	(29, 28)	0.0168	
67	29	(29, 30)	0.0202	
68	29	(29, 31)	0.0182	0.1228
69	29	(29, 32)	0.0182	
70	29	(29, 33)	0.0149	
71	29	(29, 34)	0.0202	

contd...

table -3.2 (contd.)

(1)	(2)	(3)	(4)	(5)
72	30	(30, 28)	0.0202	
73	30	(30, 29)	0.0202	
74	30	(30, 31)	0.0216	0.1019
75	30	(30, 32)	0.0216	
76	30	(30, 33)	0.0183	
77	31	(31, 27)	0.0157	
78	31	(31, 28)	0.0182	
79	31	(31, 29)	0.0182	
80	31	(31, 30)	0.0216	0.1096
81	31	(31, 32)	0.0196	
82	31	(31, 33)	0.0163	
83	32	(32, 27)	0.0157	
84	32	(32, 28)	0.0182	
85	32	(32, 29)	0.0182	
86	32	(32, 30)	0.0216	0.1096
87	32	(32, 31)	0.0196	
88	32	(32, 33)	0.0163	
89	33	(33, 18)	0.0149	
90	33	(33, 26)	0.0139	
91	33	(33, 27)	0.0124	
92	33	(33, 28)	0.0149	
93	33	(33, 29)	0.0149	0.1402

contd...

table -3.2 (contd.)

(1)	(2)	(3)	(4)	(5)
94	33	(33, 30)	0.0183	
95	33	(33, 31)	0.0163	
96	33	(33, 32)	0.0163	
97	33	(33, 34)	0.0183	
98	34	(34, 18)	0.0202	
99	34	(34, 26)	0.0192	
100	34	(34, 27)	0.0177	0.0956
101	34	(34, 29)	0.0202	
102	34	(34, 35)	0.0183	

The first cluster consists of two villages with code numbers 18 and 19. They are associated with seven and six villages each and their probabilities of forming clusters are P_1 and P_2 as are given in sixth column of Table - 3.1. Hence

$$P_1^{(1)} = \sum_{i}^{2} P_i = 0.0084 + 0.0098 = 0.0182$$ (given in fourth column of Table - 3.2)

Similarly the other probabilities are calculated which are given in fourth column of Table -3.2. and

$$\text{Inclusion Probability} = \pi_1^{(1)} = \sum_{i=1}^{6} P_i^{(1)} = 0.0182 + 0.0158 + 0.0158$$

$$+ 0.0158 + 0.0143 + 0.0149 + 0.0202 = 0.1160 .$$

Σ = Summation over the probabilities given in

fourth column where eighteenth village occurs in the clusters. These inclusion probabilities are given in fifth column of Table - 32.

After the formation of clusters, we select a cluster (20, 27) at random at the first draw. We define

$$t_1 = \frac{Y_{20}}{\pi^{(1)}_{20}} + \frac{Y_{27}}{\pi^{(1)}_{27}} = \frac{299}{0.1271} + \frac{417}{0.1453} = 2352 + 2870 = 5222$$

Since sampling is being done without replacement, we remove this cluster and again form a new set of clusters. Before that we form the new list of probabilities for the remaining fifteen villages.

TABLE - 3.3

Sl. No.	Code No. of villages	Cattle population	Associated villages within a distance of 4 miles		Probabilities $[P_i,]$
			Codes of villages	Number	
(1)	(2)	(3)	(4)	(5)	(6)
1	18	85	19,25,26,33,34	5	0.0133
2	19	1184	18,21,22,25,26	5	0.0133
3	21	81	19,22,25	3	0.0222
4	22	690	19,21,25	3	0.0222
5	23	2561	24	1	0.0667
6	24	506	23,25	2	0.0333
7	25	1070	18,19,21,22,24,26	6	0.0111
8	26	85	18,19,25,28,33,34	6	0.0111
9	28	320	26,29,30,31,32,33	6	0.0111
10	29	138	28,30,31,32,33,34	6	0.0111

contd...

table - 3.3 (contd.)

(1)	(2)		(3)	(4)	(5)
11	30	000	28,29,31,32,33	5	0.0133
12	31	000	28,29,30,32,33	5	0.0133
13	32	806	28,29,30,31,33	5	0.0133
14	33	159	18,26,28,29,30,31,32,34	8	0.0083
15	34	1166	18,26,29,33	4	0.0167

Total Cattle Population = 8521

From above, like table - 3.2, we have got the following possible number of clusters of size 2 with their probabilities of selection and inclusion probabilities

TABLE - 3.4

Sr. No.	Initial village	Clusters of two villages with code numbers	Probabilities $[P_i^{(2)}]$	Inclusion Probabilities $[\pi_i^{(2)}]$
(1)	(2)	(3)	(4)	(5)
1	18	(18, 19)	0.0266	
2	18	(18, 28)	0.0244	
3	18	(18, 26)	0.0244	0.1270
4	18	(18, 33)	0.0216	
5	18	(18, 34)	0.0300	

contd/..

table - 3.4 (contd.)

(1)	(2)	(3)	(4)	(5)
6	19	(19, 18)	0.0266	
7	19	(19, 21)	0.0355	
8	19	(19, 22)	0.0355	0.1464
9	19	(19, 25)	0.0244	
10	19	(19, 26)	0.0244	
11	21	(21, 19)	0.0355	
12	21	(21, 22)	0.0444	0.1132
13	21	(21, 25)	0.0333	
14	22	(22, 19)	0.0355	
15	22	(22, 21)	0.0444	0.1192
16	22	(22, 25)	0.0333	
17	23	(23, 24)	0.0667	0.0667
18	24	(24, 23)	0.0667	0.1111
19	24	(24, 25)	0.0444	
20	25	(25, 18)	0.0244	
21	25	(25, 19)	0.0244	
22	25	(25, 21)	0.0333	0.1820
23	25	(25, 22)	0.0333	
24	25	(25, 24)	0.0444	
25	25	(25, 26)	0.0222	

contd...

table - 3.4 (contd.)

(1)	(2)	(3)	(4)	(5)
26	26	(26, 18)	0.0244	
27	26	(26, 19)	0.0244	
28	26	(26, 25)	0.0222	0.1404
29	26	(26, 28)	0.0222	
30	26	(26, 33)	0.0194	
31	26	(26, 34)	0.0278	
32	28	(28, 26)	0.0222	
33	28	(28, 29)	0.0222	
34	28	(28, 30)	0.0244	0.1370
35	28	(28, 31)	0.0244	
36	28	(28, 32)	0.0244	
37	28	(28, 33)	0.0194	
38	29	(29, 28)	0.0222	
39	29	(29, 30)	0.0244	
40	29	(29, 31)	0.0244	0.1426
41	29	(29, 32)	0.0244	
42	29	(29, 33)	0.0194	
43	29	(29, 34)	0.0278	
44	30	(30, 28)	0.0244	
45	30	(30, 29)	0.0244	
46	30	(30, 31)	0.0266	0.1236
47	30	(30, 32)	0.0266	
48	30	(30, 33)	0.0216	

table - 3.4 (contd.)

(1)	(2)	(3)	(4)	(5)
49	31	(31, 28)	0.0244	
50	31	(31, 29)	0.0244	
51	31	(31, 30)	0.0266	0.1236
52	31	(31, 32)	0.0266	
53	31	(31, 33)	0.0216	
54	32	(32, 28)	0.0244	
55	32	(32, 29)	0.0244	
56	32	(32, 30)	0.0266	0.1236
57	32	(32, 31)	0.0266	
58	32	(32, 33)	0.0216	
59	33	(33, 18)	0.0216	
60	33	(33, 26)	0.0194	
61	33	(33, 28)	0.0194	
62	33	(33, 29)	0.0194	0.1696
63	33	(33, 30)	0.0216	
64	33	(33, 31)	0.0216	
65	33	(33, 32)	0.0216	
66	33	(33, 34)	0.0250	
67	34	(34, 18)	0.0300	
68	34	(34, 26)	0.0278	0.1106
69	34	(34, 29)	0.0278	
70	34	(34, 33)	0.0250	

At the second draw, like the first one, we have selected a cluster at random namely (24, 25) with serial number nineteen. We define

$$t_2 = (Y_{20} + Y_{27}) + \frac{y'_{24}}{\pi'_{24}(2)} + \frac{y'_{25}}{\pi'_{25}(2)} = 299 + 417 + \frac{506}{0.1111} + \frac{1070}{0.1820}$$

$$= 716 + 4554 + 5879 = 11149 .$$

Hence the combined estimate based on both the draws is given by

$$\hat{\bar{Y}} = \frac{1}{2} (5222 + 11149) = 8186$$

$$\text{Est. } \lfloor V(\hat{\bar{Y}}) \rfloor = \frac{1}{n(n-1)} \sum_{i=1}^{n} (t_i - \hat{\bar{Y}})^2$$

$$= \frac{1}{2(2-1)} \lfloor (t_1 - \hat{\bar{Y}})^2 + (t_2 - \hat{\bar{Y}})^2 \rfloor$$

$$= \frac{1}{2} \lfloor t_1^2 + t_2^2 - 2\hat{\bar{Y}}(t_1 + t_2) + 2\hat{\bar{Y}}^2 \rfloor$$

$$= \frac{1}{2} \lfloor (5222)^2 + (11149)^2 - 2(16371)(8186) + 2(8186)^2 \rfloor$$

$$= \frac{1}{2} \lfloor (27269284) + (124300201) - (268026012) + (134021192) \rfloor$$

$$= \frac{1}{2} \lfloor 17564665 \rfloor = 8782333 .$$

SUMMARY AND CONCLUSIONS

In survey practice, choice of clusters of elements as the sampling units is unavoidable. This choice is dictated on account of low cost and operational convenience. The current theory of cluster sampling presumes that the frame of district clusters is available before sampling. For instance, the list of households as clusters of individuals, list of rows of trees in forests would be available before hand for adopting cluster sampling. Such cases do not present any difficulty in the development of theory of cluster sampling. However, in many large scale surveys, the clusters of households or clusters of villages are to be chosen as sampling units. There may be many ways of forming these clusters and element of the population is likely to enter in more than one cluster. Such clusters can be termed as over - lapping clusters. For instance, in many of the sample surveys conducted by Institute of Agricultural Research Statistics, the enumerator is given a list of villages around each of which he is asked to select a sample of villages within a specified distance to form clusters at the field level. In such cases also the cluster becomes over - lapping and the usual procedures of estimating mean or total according to conventional method are not appropriate and lead to considerable bias in the estimate which itself can not be estimated. Some work in this direction has been attempted by Goel (1973), but methods of estimation have not been suggested to build up an unbiased estimate of the total

in such situations. In this dissertation a sampling procedure is slightly modified to build up an unbiased estimate and also to obtain the estimate of the variance of the estimate. This procedure is termed as Ordered Cluster Sampling.

The procedure consists in selecting a main village and forming a cluster one by one without replacement. The main advantage of this method is that the distance between villages need not be known for all villages in the population, but only to know the distances between those villages which are selected at every draw. The procedure has been illustrated by taking an example for estimating cattle population in Velvadam Firka in Vijayawada Taluk, Krishna Delta Area in Andhra Pradesh. The estimation procedure is given in Chapter -II and as also the estimation of the variance on the similar lines with Des Raj's estimator with varying probabilities without replacement.

APPENDIX - I

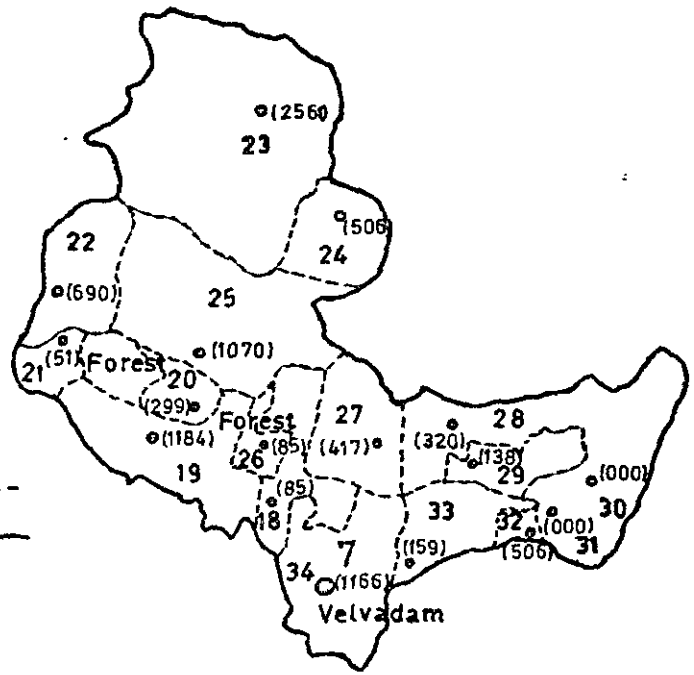
Map of Velvadam Firka showing the Location of Villages Krishna Delta Area (A.P.)

Scale 1 Inch = 4 Miles

KHAMMAM DISTRICT

NALGONDA DISTRICT

WEST GODAVARI DISTRICT



REFERENCE

Village Boundary -----
Firka Boundary ———
Firka Number 7
Village Number 21
Firka H.Q. O
Village Site o

GUNTUR DISTRICT

The figure in the bracket is the cattle population of the village.

APPENDIX - II

VELVADAM PIRKA (Vijayawada Taluk)

Sl. No,	Code number of village	Name of village	Cattle population
1	18	SAMJAPADU	85
2	19	CHANDRAGUDEM	1184
3	20*	DASULLAPALEM	299
4	21	MULAKALAPENTA	51
5	22	MORSUMILLI	690
6	23	CHINALAPADU	2561
7	24*	RUDRAVARAM	506
8	25*	PULLURU	1070
9	26	JANGALAPALLI	85
10	27*	THOLUKODU	417
11	28	NAGULURU	320
12	29	PATANAGULURU	138
13	30	PARVATHAPURAM	000
14	31	KANUMERLA	000
15	32	VEDURUBEEDEM	506
16	33	KIRTIRAYANAGUDEM	159
17	34	VELVADAM	1166

* Villages selected for two draws.

BIBLIOGRAPHY

1. Asthana, R.S. (1950), " The size of sub-sampling unit in area enumeration", Unpublished thesis for Diploma, I.C.A.R.

2. Cochran, W.G. (1963), " Sampling techniques", Second edition, John Wiley and Sons, Inc. New York, London.

3. Des Raj (1956), " Some estimators in sampling with varying probabilities without replacement", Jour. of Amer. Stat. Assoc. 51.

4. Fairfield Smith, H. (1938), " An empirical law describing heterogeneity in the yields of agricultural crops," Jour. of Agri. Science, 28. pp 1-23.

5. Ghosh, S.P. (1963), " Post cluster sampling" Ann. Math. Stat. Vol. 34, pp. 587-597.

6. Goel, B.B.P.S. (1973), " Efficiency of certain systems of cluster sampling and its application" Ph.D. thesis submitted to I.A.R.I., New Delhi.

7. Hendricks, W.A. (1944) " The relative efficiencies of groups of farms as sampling units," Jour. Amer. Stat. Assoc. Vol. 39, pp. 366-376.

8. Jessen, R.J. (1942) "Statistical investigation of a sample for obtaining farm facts." Iowa Agricultural Experimental Station Research Bulletin, No. 304.

9. Mahalanobis, P.C. (1940) " A sample survey of the acreage under Jute in Bengal," Sankhya, 4. pp. 511-530,

10. Mahalanobis, P.C. (1946) " Sample surveys of crop yield in India ": Sankhya, 7.

11. Mishro, G.K. and Sukhatme, B.V. (1972). " Efficiency of cluster sampling in conjunction with ratio and regression methods of estimation". Jour. of Ind. Soc. of Agri. Stat., Vol. 24, No. 2, pp. 81-90.

12. Murthy, M.N. (1967). " Sampling theory and methods". Statistics Publishing House, Calcutta,

13. Sethi, V.K. (1965). " On optimum pairing of units". Sankhya (B), Vol. 27, pp. 315-320.

14. Singh, D. (1956). "On efficiency of cluster sampling." Jour. of Ind. Soc. of Agri. Stat., 8.

15. Singh, D., Rajagopalan, M. and Maini, J. S. (1970). "Monograph on estimation of wool production." I. C. A. R. Research Series, New Delhi.

16. Singh, D., Murthy, V. V. R. and Goel, B. B. P. S. (1970). "Monograph on estimation of milk production." I. C. A. R. Research Series, New Delhi.

17. Sukhatme, P. V. and Sukhatme, B. V. (1970). "Sampling theory of surveys with applications." Asia Publishing House.