Metodološki zvezki, Vol. 16, No. 1, 2019, 43–60

Two Stage Adaptive Cluster Sampling based on Ordered Statistics

Girish Chandra[1] Neeraj Tiwari[2] Raman Nautiyal[3]

Abstract

The estimation problem on sparsely distributed populations using adaptive cluster sampling (ACS) is discussed. In the first phase of ACS, two stage sampling is used in which primary and secondary sampling units are selected using simple random sampling without replacement. The idea of Thompson (1996) is introduced in order to choose an appropriate fixed value of pre-specified condition, which might represent the number of rare species, before conducting the survey by the use of order statistics. Different estimators of the population mean under the two possible schemes (open and closed boundaries of primary sampling units) are studied and the Rao-Blackwell theorem for improving these estimators is also used. Numerical illustrations, one on real life data and the other based on simulation study, are discussed for these two schemes. This design may be quite useful in environmental, forestry and other areas of research dealing with rare, endangered or threatened species.

1 Introduction

Thompson (1990) introduced Adaptive Cluster Sampling (ACS), as an efficient sampling procedure for estimating totals/means of rare and clustered populations based upon the observation that when rare species are found in nature, the presence of such species is likely to reveal in neighbouring sites also. Under this procedure, for example, to estimate the total number of rare plant species in a forest, the forest could be partitioned into even-sized units (quadrats). Select some quadrats by an appropriate sampling scheme, say by simple random sampling (SRS), and count the number of rare plants, say y, therein. Whenever a quadrat satisfies a previously specified condition C, say at least one plant is recorded, i.e., $C = y\colon y \geq 1$, neighbouring quadrats are added to the sample. If, at least one plant is again found in one of the added quadrats, then all the neighbourhoods of that quadrat are added to the sample, and so on. Commonly, the condition C consists of a fixed or pre-specified value but in many studies it is difficult to pick this value before conducting the survey.

[1] Division of Forest Statistics, Indian Council of Forestry Research and Education, Dehradun, India; gchandra23@yahoo.com

[2] Department of Statistics, Kumaun University, Almora, India; kumarn_amo@yahoo.com

[3] Division of Forest Statistics, Indian Council of Forestry Research and Education, Dehradun, India; nautiyalr@icfre.org

Inappropriate selection of C may result in under or over sampling, leading to the possibility of imprecise estimation of the population parameters. In order to overcome this problem, Thompson (1996) proposed the idea of using order statistics to choose C. Christman and Lan (2001) suggested that C may be chosen based upon the proportion of rare species found in the initial sample. Another related problem is of deciding the final sample size. Brown (1994) used sequential sampling in which the sample is selected sequentially until the final sample size attains its pre-specified value. Another important design to control the final sample size was suggested by Salehi and Seber (1997). It is based on primary and secondary units in which the subsamples are not allowed to cross the boundary of primary units even though the unit satisfying the condition C were found beyond the boundary of primary units.

The procedure for selecting the initial sample plays an important role in increasing the precision of the estimates of mean and variance. Most of the researchers used SRS. Other designs are systematic sampling (Acharya et al., 2000), stratified sampling (Thompson, 1991), inverse sampling (Christman and Lan, 2001), double sampling (Félix-Medina and Thompson, 2004), cluster sampling with or without replacement of clusters (Dryver, 1999; Salehi and Seber, 1997) and ranked set sampling (Chandra et al., 2011). For survey situations in which the population consists of primary sampling units (PSUs) and each PSU consists of secondary sampling units (SSUs), the two stage sampling scheme proposed by Mahalanobis (1944) may be appropriate.

This paper deals with survey situations in which two-stage sampling methods for selecting the initial sample is found to be appropriate and it is difficult to pick a requisite value of C before the survey. SRS without replacement (SRSWOR) is used to select PSUs and SSUs. The idea of Thompson (1996) is used in which C is chosen relative to the observed sample values based on the sample order statistics. For example, in forest surveys, the number of rare plant species is measured at each quadrat in an initial sample of 50 quadrats. Additional neighbourhood quadrats are then added to the sample of the top 10 quadrats, i.e., those quadrats with the 10 largest order statistics in terms of counts of rare plants. If any of the added quadrats also have large values, still more sites may be added to the sample and so on.

The proposed design along with the notations used is described in Section 2. Section 3 deals with the various estimators of the population mean. Improvement of the estimators using Rao-Blackwell theorem is discussed in Section 4. In Section 5, the utility of the proposed design is demonstrated with the help of examples. The conclusions of the present study are discussed in Section 6.

2 The Design used and Notations

In what follows, the population is partitioned into PSUs (layout at Figure 1) so as to maintain as much as possible the homogeneity between and heterogeneity within these units, with respect to the y-values. Operational convenience may also be a criterion for PSU construction. Notations and structure of the neighbourhood used in this paper are given in Table 1 and Figure 2, respectively. The neighbourhoods do not depend on the y-values and are symmetric in relation.

The proposed design for estimating population mean μ or equivalently population

	PSU 1	PSU 2	$\cdots$	PSU i	$\cdots$	PSU M
SSU 1	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{M1}
SSU 2	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{M2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
SSU j	y_{1j}	y_{2j}	$\cdots$	y_{ij}	$\cdots$	y_{Mj}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
SSU M_i	y_{1M_1}	y_{2M_2}	$\cdots$	y_{iM_i}	$\cdots$	y_{MM_M}

Figure 1: Population layout showing y-values of SSUs for two stage sampling

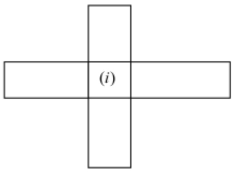

Figure 2: Neighborhood of i-th SSU for the proposed design

total Y is explained as under:

1. Select an initial sample of m PSUs using SRSWOR.

2. Select m_i SSUs from the selected i-th PSU ($i = 1, 2, \ldots, m$) using SRSWOR.

3. Add neighbourhoods of those selected SSUs satisfying $C = \{y\colon y \geq y_{(r)}\}$, where $y_{(r)}$ is the r-th sample order statistics, such that $y_{(1)} \leq y_{(2)} \leq \cdots y_{(r)} \cdots \leq y_{(n)}$. If ties occur between two order statistics, the first value in the serial number of PSUs and SSUs therein would be considered as the lower order statistic.

4. If any of added SSUs satisfies C, their neighbourhoods are also added. This process is continued until a cluster that has a boundary comprising of SSUs that do not satisfy C (also called edge SSUs) is obtained.

The final sample then consists of n (not necessarily distinct) clusters generated by each SSU selected in the initial sampling stage.

A network A_i for i-th SSU is defined to be the cluster generated by i but excluding edge SSUs. If i is the only SSU in a cluster satisfying C, then A_i consists of just i-th SSU, i.e., network of size one. If initially selected i-th SSU does not satisfy C, then A_i is a network of size one, as its selection does not lead to the inclusion of any other SSUs.

N population units are partitioned into disjoint networks, such that selection in the initial sample of any SSU in a network will result in inclusion in the final sample of all units in that network. As the final sample depends on $y_{(r)}$, it follows that $y_{(r)}$ or equivalently C depend on the initially selected sample of SSUs. This implies that the network structure for the population induced by C is not fixed. Suppose, k-th network of the population is denoted by A_k ($k = 1, 2, \ldots, K_{ir}$) having b_k units based upon $C = \{y\colon y \geq y_{(r)}\}$.

Table 1: Notations used for the design

M	Total number of PSUs in the population	$\bar{Y}_i^* = \frac{1}{M_i}\sum_{j=1}^{M_j} y_{ij}$	Mean per SSUs in i-th PSU
M_i	Size of i-th PSU	$\bar{Y}_{ib}^* = \frac{1}{M}\sum_{i=1}^{M} \bar{Y}_i^*$	Mean between the PSU means
m	Sample size of PSUs	$y_i^* = \sum_{j=1}^{m_i} y_{ij}$	Total y-values of the initial sampled SSUs of i-th PSU
m_i	Initial sample size from i-th PSU	$\bar{y}_i^* = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$	Initial sample mean per SSU of i-th PSU
$N = \sum_{i=1}^{M} M_i$	Total number of SSUs in the population	$Y = \sum_{i=1}^{M} Y_i^*$	Total y values of SSUs of the population
$n = \sum_{i=1}^{m} m_i$	Initial sample size of SSUs	$\mu = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{M} M_i\bar{Y}_i^*$	Population mean per SSU
y_{ij}	y-value of the j-th SSU in i-th PSU	$s' = {}^{N}C_n$	Number of possible initial samples
$Y_i^* = \sum_{j=1}^{M_i} y_{ij}$	Total y-values of the i-th PSU		

3 Estimators of Population Mean

3.1 Estimators without use of Adaptive Scheme

Two estimators of μ which do not make use of the observations added adaptively to the sample have been considered. The first estimator is the mean of the initial sample means per SSU i.e.

$$t_0 = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i^*$$

This estimator is biased as

$$\mathrm{E}(t_0) = \mathrm{E}_1\,\mathrm{E}_2(t_0) = \mathrm{E}_1\left(\frac{1}{m}\sum_{i=1}^{m} \mathrm{E}_2\,\bar{y}_i^*\right),$$

where E_2 denotes the averaging over all possible units of a fixed PSU and E_1 is the averaging over all selected PSUs. As all SSUs of a fixed PSU have equal probability of selection, then

$$\mathrm{E}(t_0) = \mathrm{E}_1\left(\frac{1}{m}\sum_{i=1}^{m} \bar{Y}_i^*\right) = \bar{Y}_{ib}^* \neq \mu;$$

t_0 is unbiased only when all M_i's are equal.

The variance of the above estimator, is given by

$$\mathrm{Var}(t_0) = \mathrm{Var}_1\left(\mathrm{E}_2(t_0)\right) + \mathrm{E}_1\left(\mathrm{Var}_2(t_0)\right),$$

where Var_2 and Var_1 represent the variances over all selected SSUs of a given PSU and the variance over all selected PSUs, respectively.

Since $\mathrm{E}_2(t_0) = \frac{1}{m}\sum_{i=1}^{m} \bar{Y}_i^*$ and $\mathrm{Var}_1\left(\frac{1}{m}\sum_{i=1}^{m} \bar{Y}_i^*\right)$ is the variance of the sample mean per PSU for one stage simple random sample of m SSUs, hence by analogy to SRSWOR

$$\mathrm{Var}_1\left(\frac{1}{m}\sum_{i=1}^{m} \bar{Y}_i^*\right) = \frac{M-m}{Mm} S_1^2,$$

where

$$S_1^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\bar{Y}_i^* - \bar{Y}_{ib}^*)^2.$$

Furthermore, as all contributions from cross-product term vanish, therefore,

$$\mathrm{Var}_2(t_0) = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}_2(\bar{y}_i^*) = \frac{1}{m^2}\sum_{i=1}^{m} \frac{M_i - m_i}{M_i m_i} S_{2i}^2.$$

Here,

$$S_{2i}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i^*)^2$$

is the variance among SSUs of i-th PSU. Now,

$$\mathrm{E}_1(\mathrm{Var}_2(t_0)) = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{E}_1\left(\frac{M_i - m_i}{M_i m_i} S_{2i}^2\right) = \frac{1}{mM}\sum_{i=1}^{M}\left(\frac{M_i - m_i}{M_i m_i} S_{2i}^2\right)$$

therefore

$$\mathrm{Var}(t_0) = \frac{M-m}{Mm} S_1^2 + \frac{1}{mM}\sum_{i=1}^{M}\left(\frac{M_i - m_i}{M_i m_i} S_{2i}^2\right).$$

To find an unbiased estimator of $\mathrm{Var}(t_0)$, theorem 11.2 of Cochran (1977, p. 301) is used. It is given by

$$\hat{\mathrm{Var}}(t_0) = \frac{M-m}{Mm} s_1^2 + \sum_{i=1}^{m}\left(\frac{M_i - m_i}{M_i m_i} s_{2i}^2\right),$$

where

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{y}_i^* - \bar{y}_{ib}^*)^2, \quad \bar{y}_{ib}^* = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i^*, \quad s_{2i}^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i^*)^2$$

Another, sample mean estimator can be obtained for the two stage sampling scheme as

$$t_1 = \frac{1}{m}\sum_{i=1}^{m} t_{1i},$$

where

$$t_{1i} = \frac{MM_i}{N}\bar{y}_i^*.$$

This estimator is unbiased as

$$\mathrm{E}(t_{1i}) = \mathrm{E}_1\,\mathrm{E}_2\left(\frac{MM_i}{N}\bar{y}_i^*\right) = \frac{M}{N}\,\mathrm{E}_1(Y_i^*) = \frac{1}{N}\sum_{i=1}^{M} Y_i^* = \mu$$

Furthermore,

$$\begin{aligned}\mathrm{Var}(t_1) &= \mathrm{Var}_1(\mathrm{E}_2(t_1)) + \mathrm{E}_1(\mathrm{Var}_2(t_1))\\ &= \mathrm{Var}_1\left(\frac{M}{Nm}\sum_{i=1}^{m} Y_i^*\right) + \frac{M^2}{N^2m^2}\,\mathrm{E}_1\left(\sum_{i=1}^{m} M_i^2\,\mathrm{Var}_2(\bar{y}_i^*)\right)\\ &= \frac{M^2}{N^2}\,\mathrm{Var}_1\left(\frac{1}{m}\sum_{i=1}^{m} Y_i^*\right) + \frac{M^2}{N^2m^2}\,\mathrm{E}_1\left(\sum_{i=1}^{m} M_i\frac{M_i - m_i}{m_i}S_{2i}^2\right).\end{aligned}$$

Or

$$\mathrm{Var}(t_1) = \frac{M(M-m)}{N^2m}\frac{1}{M-1}\sum_{i=1}^{M}\left(Y_i^* - \frac{Y}{M}\right)^2 + \frac{M}{N^2m}\sum_{i=1}^{M}\frac{M_i(M_i - m_i)}{m_i}S_{2i}^2$$

3.2 Estimators under Open Boundary

An open boundary means, the boundaries of PSUs are ignored while including the neighbourhoods of those SSUs whose y-values exceed $y_{(r)}$, for some r. Thus the cluster generated by a SSU may contain the SSUs from two or more PSUs. Suppose the final sample is the unordered set $s = \{s_1, s_2\}$, where s_1 is the set of n unordered labels from the initial sample, and s_2 is the set of distinct unordered labels from the remainder of the sample s. It is clear that all SSUs from s are distinct as the initial sampling procedure is SRSWOR.

3.2.1 Modified Type of Horvitz-Thompson (HT) Estimator

Using the idea introduced in Thompson (1990), a modified type of HT estimator of μ in terms of networks can be written as

$$t_{2(OB)} = \frac{1}{s'}\sum_{i=1}^{s'} t_{2(OB)i},$$

where

$$s' = {}^{N}C_n$$

and

$$t_{2(OB)i} = \frac{1}{N}\sum_{k=1}^{K_{ir}}\frac{y_k^{**}J_k}{\pi_k}.$$

π_k = Partial inclusion prabability that the SSU belonging to A_k is included in s = Probability that s_1 intersect A_k

$$= 1 - \left(\frac{\binom{N-b_k}{n}}{\binom{N}{n}} \right)$$

Indicator variable

$$J_k = \begin{cases} 1 & \text{if } s_1 \text{ intersects } A_k \\ 0 & \text{otherwise} \end{cases}$$

and y_k^{**} is the sum of the y-values for A_k.

Practically, it is not possible to calculate $t_{2(OB)i}$ for each possible initial sample. Therefore, the estimator $t_{2(OB)}$ cannot be unbiased based upon a particular initial sample. It is unbiased provided all K_{ir}'s are equal, however, $t_{2(OB)} = t_{2(OB)i}|i$-th initial sample may be considered as an estimator of μ. Now

$$\mathrm{E}\left(t_{2(OB)} = t_{2(OB)i}|i\text{-th initial sample}\right) = \mu \quad \text{as} \quad \mathrm{E}(J_k) = \pi_k.$$

To calculate the variance of $t_{2(OB)} = t_{2(OB)i}|i$-th initial sample, we applied the idea of Thompson (1990) and get

$$\mathrm{Var}\left(t_{2(OB)} = t_{2(OB)i}|i\text{-th initial sample}\right) = \frac{1}{N^2}\left(\sum_{j=1}^{K_{ir}}\sum_{k=1}^{K_{ir}} y_j^{**}y_k^{**}\left(\frac{\pi_{jk}-\pi_j\pi_k}{\pi_j\pi_k}\right)\right),$$

where $\pi_{jk} = P(J_j = 1, J_k = 1)$, the partial inclusion probability that both A_j and A_k intersect s_1

$$\begin{aligned} &= P(J_j = 1) + P(J_k = 1) - P(J_j = 1 \text{ or } J_k = 1) \\ &= \pi_j + \pi_k - (1 - P(J_j \neq 1, J_k \neq 1)) \\ &= 1 - \frac{\left(\binom{N-b_j}{n} + \binom{N-b_k}{n} - \binom{N-b_j-b_k}{n}\right)}{\binom{N}{n}} \end{aligned}$$

with $\pi_{jj} = \pi_j$ as $P(J_j \neq 1, J_k \neq 1)$ = $P(j$-th and k-th network do not intersect$)$ $=$ $\binom{N-b_j-b_k}{n}/\binom{N}{n}$.

An unbiased estimator of the variance of $t_{2(OB)} = t_{2(OB)i}|i$-th initial sample is

$$\hat{\mathrm{Var}}\left(t_{2(OB)} = t_{2(OB)i}|i\text{-th initial sample}\right) = \frac{1}{N^2}\left(\sum_{j=1}^{K_{ir}}\sum_{k=1}^{K_{ir}} y_j^{**}y_k^{**}\left(\frac{\pi_{jk}-\pi_j\pi_k}{\pi_{jk}\pi_j\pi_k}\right)\right)$$

provided that none of the joint probabilities are zero.

Just as the HT estimator has lower variance when the y-values are approximately proportional to the inclusion probabilities, the estimator $t_{2(OB)}$ should have low variance when the network totals y_k^{**}'s are proportional to π_k.

3.2.2 Modified Type of Hansen-Hurwitz (HH) Estimator

Another type of estimator, as suggested by Thompson (1990), which is a modified HH type estimator, was used under this design. As we know the selection probability may not be known for every SSU in the sample. An unbiased estimator can be formed by modifying the HH estimator by making use of observations having y-values even less than or equal to $y_{(r)}$. This estimator depends upon the total of n networks (which may not be distinct) generated by each SSU of the initial sample. The modified HH type of estimator in terms of SSUs of s_1 can be written as

$$t_{3(OB)} = \frac{1}{n}\sum_{k=1}^{n}\bar{y}_k^{**} = \frac{1}{n}\sum_{k=1}^{n}\sum_{j\in A_k}\frac{y_{(j)}}{b_j}.$$

Here it should be noted that the number of networks is taken to be n instead the number of distinct networks K_{ir} as taken in $t_{2(OB)}$. The network size, however, may vary from sample (initial) to sample.

As this estimator does not depend on the network structure of the population induced by C for additional sampling as well as the number of networks (which are not fixed in this design) in the population, it does not require the computations of estimators, their mean square errors (MSEs) etc. on each possible initial samples for unbiasedness, as in the case of $t_{2(OB)}$. This estimator is unbiased using Chen, Bai and Sinha (2004, theorem 6.1, p. 165) and due to the fact that the initial sampling is SRSWOR.

The variance of $t_{3(OB)}$ is

$$\text{Var}(t_{3(OB)}) = \frac{N-1}{Nn}S_0^2$$

where

$$S_o^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_i^{**} - \mu\right)^2.$$

An unbiased estimator of $\text{Var}(t_{3(OB)})$ is

$$\hat{\text{Var}}\left(t_{3(OB)}\right) = \frac{N-n}{Nn(n-1)}\sum_{k=1}^{n}\left(\bar{y}_k^{**} - t_{3(OB)}\right)^2$$

3.3 Estimators under Closed Boundary

In the closed boundary case we do not allow the additional SSUs to cross the boundaries of the PSUs during the final sample selection. Hence, the networks do not always consist of the SSUs from two or more PSUs. The order statistics based on the y-values of the SSUs for each PSU are independent and made under similar guidelines as for the open boundary case. Without any loss of generality, we assume that the first m PSUs are selected and first m_i SSUs are selected from the selected PSUs.

3.3.1 Modified Type of HT Estimator

The modified HT type of estimator under this case is

$$t_{2(CB)} = \frac{M}{N}\sum_{i=1}^{m}\frac{\eta_i}{m},$$

where η_i is sum of y-values of networks intersected by initial sample of i-th PSU divided by the corresponding intersection probabilities.

If we denote by K_i, y^{**}_{ik}, and π_{ik}, the number of distinct networks in the i-th PSU, sum of y-values associated with network k, and the probability that the initial sample of i-th PSU intersect network k, respectively, then,

$$\eta_i = \sum_{k=1}^{K_i} y^{**}_{ik}\left(\frac{J_{ik}}{\pi_{ik}}\right)$$

where

$$J_{ik} = \begin{cases} 1 & \text{if } s_1 \text{ intersects network } k \text{ of } i\text{-th PSU} \\ 0 & \text{otherwise.} \end{cases}$$

We note that $\mathrm{E}(\eta_i) = Y^*_i$ and $E(J_{ik}) = \pi_{ik}$. Therefore $t_{2(CB)}$ is an unbiased estimator of μ. We have

$$\mathrm{Var}(t_{2(CB)}) = \frac{M^2}{N^2m^2}\sum_{i=1}^{m}\mathrm{Var}(\eta_i) = \frac{M^2}{N^2m^2}\sum_{i=1}^{m}V_i$$

where

$$V_i = \left(\sum_{j=1}^{K_i}\sum_{k=1}^{K_i} y^{**}_{ij}y^{**}_{ik}\left(\frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{\pi_{ijk}\pi_{ij}\pi_{ik}}\right)\right)$$

and

$$\begin{aligned}\pi_{ijk} &= P(J_{ij} = 1, J_{ik} = 1)\\ &= 1 - \frac{\binom{M_i-b_{ij}}{m_i} + \binom{M_i-b_{ik}}{m_i} - \binom{M_i-b_{ij}-b_{ik}}{m_i}}{\binom{M_i}{m_i}},\end{aligned}$$

the probability that the initial sample intersects the networks j and k both of the i-th PSU. Here, $\pi_{ijk} = \pi_{ik}$ and $V_i = 0$, if $K_i = 0$. The unbiased estimator of variance of $t_{2(CB)}$ is

$$\hat{\mathrm{Var}}(t_{2(CB)}) = \frac{M^2}{N^2m^2}\sum_{i=1}^{m}\left(\sum_{j=1}^{K'_i}\sum_{k=1}^{K'_i} y^{**}_{ij}y^{**}_{ik}\left(\frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{\pi_{ijk}\pi_{ij}\pi_{ik}}\right)\right)$$

where K'_i is the number of distinct networks intersected by initial sample in the i-th PSU and none of the joint probabilities π_{ijk} is zero.

3.3.2 Modified Type of HH Estimator

The modified HH estimator ($t_{3(CB)}$) can also be obtained by averaging

$$t_{3(CB)i} = \frac{1}{m_i}\sum_{k=1}^{m_i} \bar{y}_k^{**}.$$

That is

$$t_{3(CB)} = \frac{1}{m}\sum_{i=1}^{m} t_{3(CB)i}.$$

The variance of $t_{3(CB)i}$ is

$$\text{Var}\left(t_{3(CB)i}\right) = \frac{M_i - m_i}{M_i m_i} S_{3i}^2$$

where

$$S_{3i}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \bar{Y}_i^{**}\right)^2.$$

As $t_{3(CB)i}$ are independent, we have

$$\text{Var}\left(t_{3(CB)}\right) = \frac{1}{m}\sum_{i=1}^{m}\text{Var}\left(t_{3(CB)i}.\right)$$

An unbiased estimator of Var $\left(t_{3(CB)i}\right)$ is

$$\hat{\text{Var}}\left(t_{3(CB)i}\right) = \frac{M_i - m_i}{M_i m_i}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i^{**}\right)^2.$$

Hence, unbiased estimator of Var $\left(t_{3(CB)}\right)$ is

$$\hat{\text{Var}}\left(t_{3(CB)}\right) = \frac{1}{m^2}\sum_{i=1}^{m}\hat{\text{Var}}\left(t_{3(CB)i}\right).$$

4 Improvement of the Estimators using Rao-Blackwell Method

Unbiased estimators t_1, $t_{2(OB)}$, $t_{3(OB)}$, $t_{2(CB)}$, and $t_{3(CB)}$ are not functions of the minimal sufficient statistic, say D. They may be improved by using the Rao-Blackwell theorem which involves taking conditional expectations given D. Here, we can use $D = \{(k, y_k)\colon k \in s\}$, the unordered set of distinct, labelled observations, as suggested by Basu (1969) for a finite population.

Starting with any unbiased estimator $t = t_1, t_{2(OB)}, t_{3(OB)}, t_{2(CB)}, t_{3(CB)}$, we take $t_{RB} = \text{E}(t|D)$. Let n' denote the number of distinct units in the final sample s. As the initial sample s_1 is selected without replacement there is a total of $G = \binom{n'}{n}$ possible combinations of n distinct units from the n' in the sample. Suppose that these combinations are labelled in an arbitrary way by $g = 1, 2, \ldots, G$. Let t_g denote the value of t when s_1 consists of combination g and let $\hat{\text{Var}}_g(t)$ denote the value of the unbiased estimator $\hat{\text{Var}}(t)$, when computed using the g-th combination.

An initial sample that gives rise through the design to a given value D of the minimal sufficient statistic is called compatible with D. Let the g-th indicator variable (I_g) take the value 1 if the g-th combination can give rise to D (i.e., compatible with D), and 0 otherwise. The number of compatible combinations is

$$\xi = \sum_{g=1}^{G} I_g.$$

The estimator t may be improved using the Rao-Blackwell theorem and is the average of the values of t obtained over all those initial samples that are compatible with D. This improved estimator t_{RB} is

$$t_{RB} = \text{E}\left(t|D\right) = \frac{1}{\xi}\sum_{g=1}^{G} t_g I_g$$

and its variance is given by

$$\text{Var}(t_{RB}) = \text{Var}(t) - \text{E}\left(\text{Var}\left(t|D\right)\right).$$

An unbiased estimator of the variance of t_{RB} due to Thompson (1990) is given by

$$\hat{\text{Var}}\left(t_{RB}\right) = \frac{1}{\xi}\sum_{g=1}^{G}\left(\hat{\text{Var}}(t_g) - (t_g - t_{RB})^2\right) I_g.$$

From the above, the steps to improve t using Rao-Blackwell theorem can be summarized as:

1. List all possible combinations of initial samples of the same size taken from final sample s which are compatible with D. Let Ψ denote the set of such initial samples.

2. Calculate t for all the samples generated from above initial sample Ψ.

3. The values of the Rao-Blackwell version of any t are obtained by averaging the value of the corresponding estimator over the samples generated under step (2) which give t_{RB}.

For $D = \{(k, y_k) \colon k \in s\}$ and initial sample s_1, the improved estimator can be obtained by averaging t for all those initial samples which give rise to exactly the same final sample s. For large samples the calculation of t_{RB} are difficult due to large number of such initial samples. In Section 5.2, we consider a simulation study in which the t_{RB}'s are calculated.

0	$\underline{\mathbf{2}}$	2	2	$\mathbf{0}$	1	0	$\mathbf{0}$	0	2
1	$\mathbf{1}$	3	1	$\underline{\mathbf{0}}$	0	0	$\mathbf{0}$	0	0
0	$\underline{\mathbf{0}}$	1	1	$\underline{\mathbf{0}}$	0	2	$\underline{\mathbf{0}}$	1	3
1	$\mathbf{1}$	0	0	$\mathbf{0}$	0	0	$\underline{\mathbf{0}}$	0	0
0	$\underline{\mathbf{1}}$	1	0	$\underline{\mathbf{0}}$	0	2	$\mathbf{0}$	2	1
3	$\mathbf{0}$	0	0	$\mathbf{0}$	0	0	$\mathbf{1}$	0	0
0	$\mathbf{0}$	0	0	$\mathbf{0}$	0	0	$\underline{\mathbf{1}}$	0	0
2	$\underline{\mathbf{0}}$	0	1	$\underline{\mathbf{2}}$	3	1	$\mathbf{0}$	2	2
1	$\underline{\mathbf{0}}$	0	1	$\mathbf{1}$	0	0	$\underline{\mathbf{0}}$	0	0
0	$\underline{\mathbf{0}}$	0	3	$\underline{\mathbf{3}}$	0	0	$\underline{\mathbf{0}}$	1	3

Figure 3: Occurrence of *R. Edgeworthii* in eastern Himalaya with 10 PSUs and 10 SSUs in each PSU

$\mathbf{0}$	$\underline{2}$*	$\mathbf{2}$*	$\mathbf{2}$*	$\mathbf{0}$	1	0	0	0	0
1	$\mathbf{1}$	$\mathbf{3}$*	$\mathbf{1}$	$\underline{\mathbf{0}}$	0	0	0	0	0
0	$\underline{\mathbf{0}}$	$\mathbf{1}$	1	$\underline{\mathbf{0}}$	0	0	$\underline{\mathbf{0}}$	1	0
1	1	0	0	0	0	0	$\underline{\mathbf{0}}$	0	0
0	$\underline{\mathbf{1}}$	1	0	$\underline{\mathbf{0}}$	0	2	0	0	1
0	0	0	0	0	0	0	1	0	0
0	0	0	0	$\mathbf{0}$	$\mathbf{0}$	0	$\underline{\mathbf{1}}$	0	0
0	$\underline{\mathbf{0}}$	0	$\mathbf{1}$	$\underline{2}$*	$\mathbf{3}$*	$\mathbf{1}$	0	0	0
1	$\underline{\mathbf{0}}$	0	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{0}$	0	$\underline{\mathbf{0}}$	0	0
0	0	$\mathbf{0}$	$\mathbf{3}$**	$\underline{3}$**	$\mathbf{0}$	0	$\underline{\mathbf{0}}$	1	0

Figure 4: Final sample selection based on $y_{(n)}$ and $y_{(n-1)}$ under open boundary

5 Numerical Comparisons

5.1 Based upon Empirical Study

An example is illustrated in Figure 3, giving the occurrence data (Menon et al., 2012) of *Rhododendron* species (*R. Edgeworthii*), in which the aim was to estimate its occurrence in Indian Eastern Himalaya. The study area was divided into 10 × 10 square quadrats (SSUs). Ten columns represent PSUs and the y-value of i-th SSU represents the counts of this species in each cell. Two values of C, $y_{(n)}$ (largest order statistics) and $y_{(n-1)}$, are taken for the purpose of demonstration and computations. Three PSUs ("bold" outline) and five SSUs from selected PSUs (with "underlined" outline) were selected as per the procedure. It is considered that the boundaries of PSUs are open for the selection of SSUs under adaptive scheme of the design. The final sample (networks plus edge SSUs) based on $y_{(n)} = y_{(15)} = 3$ (shown by "**") and $y_{(n-1)} = y_{(14)} = 2$ (shown by "*" and "**") are shown in Figure 4. SSUs having black background represent the edge SSUs. There are total 1 network based on $y_{(15)}$ and total 3 networks based on $y_{(14)}$. In the closed boundary case, the final sample based on $y_{(n)}$ and $y_{(n-1)}$ is shown in Figure 5 with the same representations of the networks based upon $y_{(n)}$ and $y_{(n-1)}$ and edge SSUs, as shown in open boundary case.

0	2*	2	2	0	1	0	0	0	0
1	**1**	3	1	**0**	0	0	0	0	0
0	**0**	1	1	**0**	0	0	**0**	1	0
1	1	0	0	0	0	0	**0**	0	0
0	**1**	1	0	**0**	0	2	0	0	1
0	0	0	0	0	0	0	1	0	0
0	0	0	0	**0**	0	0	**1**	0	0
0	**0**	0	1	2*	3	1	0	0	0
1	**0**	0	1	**1**	0	0	**0**	0	0
0	0	0	3	3**	0	0	**0**	1	0

Figure 5: Final sample based on $y_{(n)}$ and $y_{(n-1)}$ under closed boundary

In our notations, we have $N = 100$, $M = M_j = 10$ ($j = 1, 2, \ldots, 10$), $m = 3$, $m_i = 5$ ($i = 1, 2, 3$) and use of neighbourhood leads to Figure 2, but for the case of closed boundaries it consists only of north and south SSUs. In this example we have carried out additional sampling in the vicinity of the largest ($y_{(15)}$) and the second largest ($y_{(14)}$) order statistics of the initial sample. The population mean and variance are 0.630 and 0.882 respectively. Calculations for open and closed boundaries using Figure 4 and Figure 5 are given in the Table 2.

Table 2: Result of performance

		Open boundaries		
Condition	$t_{2(OB)}$	$\mathrm{Var}(t_{2(OB)})$	$t_{3(OB)}$	$\mathrm{Var}(t_{2(OB)})$
$C = \{y \colon y \geq y_{(n)} = 3\}$	0.615	0.055	0.600	0.055
$C = \{y \colon y \geq y_{(n-1)} = 2\}$	0.714	0.093	0.650	0.065
		*Closed boundarie*s		
Condition	$t_{2(CB)}$	$\mathrm{Var}(t_{2(CB)})$	$t_{3(CB)}$	$\mathrm{Var}(t_{2(CB)})$
$C = \{y \colon y \geq y_{(n)} = 3\}$	0.605	0.030	0.605	0.130
$C = \{y \colon y \geq y_{(n-1)} = 2\}$	0.605	0.030	0.605	0.130
		Initial estimators		
t_0	$\mathrm{Var}(t_0)$	t_1	$\mathrm{Var}(t_1)$	
0.600	0.046	0.600	0.046	

From Table 2, it is seen that both the initial estimators producing the equal mean estimates due to the equal sizes of each PSU. The HT type estimators are producing more than or equal yield than that of the HH type estimators. The main reason might be that in the calculation of HT type estimators, two overlapping networks are considered only once which does not reduce the values while averaging as in the HH type of estimators. In this example, it is seen that some of the networks generated by the initial sample are overlapped. The networks which are overlapped have smaller y-values than the non-overlapped networks. However, it may not true in general. Considering that there is not much difference in these two types of estimators.

5.2 Based upon Simulation

A simulation to see the performance of the improved estimators using Rao-Blackwell theorem was conducted. The three columns each of size three were generated using the R software from the Pareto distribution with the shape and scale parameters of 3 and 5, respectively. The R code for the simulation is given in the Appendix.

Table 3 shows the population consisting $N = 9$, $M = M_i = 3$ $(i = 1, 2, 3)$. Here, we used $m = 2$, $m_i = 1$ $(i = 1, 2)$ and the additional sampling carried out in the vicinity of C: $y = y_{(1)} \geq 6.0000$. With the proposed design, there are 27 possible initial samples (Table 4) with the y-values of SSU from selected first PSU (SSU 1) and SSU from selected second PSU (SSU 2). The population mean and variance are 9.6570 and 31.3200 respectively.

Table 3: 3×3 populations from the Pareto (3, 5) distribution

A1	A2	A3
5.0296	16.7234	6.6590
22.0306	10.8277	7.7327
5.5818	5.0781	7.2504

Table 4 provides all possible initial Samples (SSU 1, SSU 2) with the value of estimates, bias and MSE. The values of t_0 and t_1 for all the initial samples results to the same values and therefore only t_1 is mentioned. Consider our fourth initial sample with y-values (22.0306, 16.7234) and $y_{(1)} = 16.7234 \geq 6.0000$. Since both SSUs are satisfying C, the final sample under open boundary case consists of the whole population consisting of a network containing the SSUs (22.0306, 16.7234, 10.8277, 6.6590, 7.7327 and 7.2504). The computations of the estimators give $t_{2(OB)} = (71.2238/0.9259)/9 = 8.55$ in which network total is 71.2238 and $\pi_1 = \pi_2 = 0.9259$ and $t_{3(OB)} = {}^1\!/_2(71.2238/6 + 71.2238/6) = 11.87$. The Rao-Blackwell version of any of the estimators for this particular sample are obtained by averaging the value of the corresponding estimator over all samples except the 3-rd and the 9-th sample that are not compatible with D. For the case of closed boundary case $t_{2(CB)} = {}^1\!/_6(66.0918 + 41.3266) = 17.90$ in which $\eta_1 = 0.33$ and $\eta_2 = 0.67$ and $t_{3(CB)} = {}^1\!/_2\,(22.0306 + (16.7234 + 10.8277)/2) = 17.90$.

All the estimators are unbiased in this example. The implementation of Rao-Blackwellization in the open boundary case substantially reduces the MSEs of both the estimators of adaptive scheme. However, the implementation of Rao-Blackwellization does not affect the case of closed boundary as far as the MSE is concerned. It is predicted that MSE may reduce for the large population.

Table 4: All possible initial samples (SSU1, SSU 2) and values of different estimators

SN	SSU 1	SSU 2	t_1	$t_{2(OB)}$	$t_{2(CB)}$	$t_{3(OB)}$	$t_{3(CB)}$	$t_{2(OB)RB}$	$t_{2(CB)RB}$	$t_{3(OB)RB}$	$t_{3(CB)RB}$
1	5.0296	16.7234	10.88	11.06	8.45	9.40	9.40	10.01	10.01	9.40	9.40
2	5.0296	10.8277	7.93	11.06	8.45	9.40	9.40	10.01	10.01	9.40	9.40
3	5.0296	5.0781	5.05	5.05	5.05	5.05	5.05	5.05	5.05	5.05	5.05
4	22.0306	16.7234	19.38	8.55	11.87	17.90	17.90	10.01	10.01	17.90	17.90
5	22.0306	10.8277	16.43	8.55	11.87	17.90	17.90	10.01	10.01	17.90	17.90
6	22.0306	5.0781	13.55	11.09	8.47	13.55	13.55	10.01	10.01	13.55	13.55
7	5.5818	16.7234	11.15	11.34	8.73	9.68	9.68	10.01	10.01	9.68	9.68
8	5.5818	10.8277	8.20	11.34	8.73	9.68	9.68	10.01	10.01	9.68	9.68
9	5.5818	5.0781	5.33	5.33	5.33	5.33	5.33	5.33	5.33	5.33	5.33
10	5.0296	6.6590	5.84	11.06	8.45	6.12	6.12	10.01	10.01	6.12	6.12
11	5.0296	7.7327	6.38	11.06	8.45	6.12	6.12	10.01	10.01	6.12	6.12
12	5.0296	7.2504	6.14	11.06	8.45	6.12	6.12	10.01	10.01	6.12	6.12
13	22.0306	6.6590	14.34	8.55	11.87	14.62	14.62	10.01	10.01	14.62	14.62
14	22.0306	7.7327	14.88	8.55	11.87	14.62	14.62	10.01	10.01	14.62	14.62
15	22.0306	7.2504	14.64	8.55	11.87	14.62	14.62	10.01	10.01	14.62	14.62
16	5.5818	6.6590	6.12	11.34	8.73	6.40	6.40	10.01	10.01	6.40	6.40
17	5.5818	7.7327	6.66	11.34	8.73	6.40	6.40	10.01	10.01	6.40	6.40
18	5.5818	7.2504	6.42	11.34	8.73	6.40	6.40	10.01	10.01	6.40	6.40
19	16.7234	6.6590	11.69	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
20	16.7234	7.7327	12.23	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
21	16.7234	7.2504	11.99	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
22	10.8277	6.6590	8.74	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
23	10.8277	7.7327	9.28	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
24	10.8277	7.2504	9.04	8.55	11.87	10.49	10.49	10.01	10.01	10.49	10.49
25	5.0781	6.6590	5.87	11.09	8.47	6.15	6.15	10.01	10.01	6.15	6.15

continued . . .

…continued

SN	SSU 1	SSU 2	t_1	$t_{2(OB)}$	$t_{2(CB)}$	$t_{3(OB)}$	$t_{3(CB)}$	$t_{2(OB)RB}$	$t_{2(CB)RB}$	$t_{3(OB)RB}$	$t_{3(CB)RB}$
26	5.0781	7.7327	6.41	11.09	8.47	6.15	6.15	10.01	10.01	6.15	6.15
27	5.0781	7.2504	6.16	11.09	8.47	6.15	6.15	10.01	10.01	6.15	6.15
Mean			9.66	9.66	9.66	9.66	9.66	9.66	9.66	9.66	9.66
Bias			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MSE			14.91	3.17	4.11	13.92	13.92	1.60	1.60	13.92	13.92

6 Conclusions

The sampling design presented in this paper provides the distribution patterns with density of rare species for the PSUs under interest. The final sample size may be large or small in some cases; it can be adjusted through the 'condition of interest'. The design efficiency of open and closed boundary cases is not compared due to different sample sizes. Where cost is not an issue, estimators under open boundary case may be preferred over the closed boundary case. The calculation of variances of modified Horvitz-Thompson estimators may be more complicated than that of Hansen-Hurwitz estimators for large sample size. Hansen-Hurwitz type of estimators for such cases may be preferred. The Horvitz-Thompson estimator has smaller variance when y-values are approximately proportional to the inclusion probabilities. Similarly, Horvitz-Thompson estimator of open and closed boundary cases should have low variance when the network totals are proportional to the corresponding partial inclusion probability. Example 1 demonstrates that the estimators under adaptive designs are closer to the population mean in comparison to the classical estimators. In the simulation study, we see that the MSE of Rao-Blackwell version does not exceed that of the original estimator and the Horvitz-Thompson estimator gives smaller or equal MSE than the Hansen-Hurwitz estimators. Further, all the adaptive strategies are more efficient than classical sampling.

Acknowledgements

The authors are highly grateful to the two unknown reviewers for their constructive comments which led to substantial improvement in the paper.

References

[1] Acharya B., Bhattarai, G., Gier, A. and Stein, A. (2000): Systematic adaptive cluster sampling for the assessment of rare tree species in Nepal. *Forestry and Ecology Management*, **137**, 65–73.

[2] Basu, D. (1969): Role of the sufficiency and likelihood principle in sample survey theory. *Sankhya*, **31**(A), 441–454.

[3] Brown, J. A. (1994): The application of adaptive cluster sampling to ecological studies. In: D. J. Fletcher and B. F. J. Manly (Eds.): *Statistics in Ecology and Environmental Monitoring*, 86–97. Dunedin, New Zealand: University of Otago Press.

[4] Chandra, G., Tiwari, N. and Chandra, H. (2011): Adaptive cluster sampling based on ranked sets. *Metodološki zvezki*, **8**(1), 39–55.

[5] Chen, Z., Bai, Z. D. and Sinha, B.K. (2004): *Ranked Set Sampling: Theory and Applications*. New York, NY: Springer.

[6] Christman, M. C. and Lan, F. (2001): Inverse adaptive cluster sampling. *Biometrics*, **57**, 1096–1105.

[7] Cochran, W. G. (1977): *Sampling Techniques*. New York, NY: John Wiley.

[8] Dryver, A. L. (1999): Adaptive sampling designs and associated estimators. Dissertation, The Pennsylvania State University, USA.

[9] Félix-Medina, M. H. and Thompson S. K. (2004): Adaptive cluster double sampling. *Biometrica*, **91**, 877–891.

[10] Mahalanobis, P. C. (1944): On large scale sample surveys. *Philosophical Transactions of the Royal Society of London*, **B231**, 329–451.

[11] Menon, S., Khan, M. L., Paul, A. and Peterson, A. T. (2012): Rhododendron species in the Indian eastern himalayas: New approaches to understanding rare plant species distributions. *Journal of the American Rhododendron Society*, **1**, 78–84.

[12] Salehi, M. M. and Seber, G. A. F. (1997): Two stage adaptive cluster sampling. *Biometrics*, **53**, 959–970.

[13] Thompson, S. K. (1990): Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**, 1050–1059.

[14] Thompson, S. K. (1991): Stratified adaptive cluster sampling. *Biometrika*, **78**, 389–397.

[15] Thompson, S. K. (1996): Adaptive cluster sampling based on order statistics. *Environmetrics*, **7**(2), 123–133.