Statistics and Applications {ISSN 2454-7395 (online)}
Volume 16, No. 2, 2018 (New Series) pp 115-131

Negative Adaptive Cluster Double Sampling

R.V. Latpate[1] and J.K. Kshirsagar[2]
[1]*Savitribai Phule Pune University, Pune*
[2]*Department of Statistics, NAC&S College, Ahmednagar*

Received January 13, 2018; Revised November 06, 2018; Accepted December 16, 2018

Abstract

Negative adaptive cluster double sampling (ACDS) is the combination of double sampling and negative adaptive cluster sampling (NACS). It is an efficient method for estimation of the parameters of a rare and clustered population in the presence of an auxiliary variable. If auxiliary variable is highly positively correlated with study variable, then auxiliary information is also rare. In such situations, the maximum utilization of auxiliary information at design stage and estimation stage is not possible. Hence, we consider the auxiliary variable which is highly negatively correlated with the study variable and the population related to study variable is highly clustered and patchy. According to the nature of such kind of population, auxiliary information is abundant and we can exploit it at design stage and estimation stage. The adaptation of units is based on the auxiliary information instead of study variable. Regression and ratio estimators based on S*ä*rndal's estimators are proposed. Also, the product type ratio HT estimator is proposed. The estimators of the variances of these estimators are also obtained. Sample survey in Tamhini Ghat, Maharashtra, India is conducted using the proposed methodology. The proposed design is found to be cost effective as compared to the existing designs.

Key words: Adaptive cluster sampling (ACS), ACDS, regression estimator, ratio estimator.

1. Introduction

In general the traditional sampling methods such as simple random sampling (SRS), stratified sampling, systematic sampling etc. are used to draw a sample from the population. The estimate of the population mean/ total is obtained by using the sample so drawn. But if the population under study is rare and patchy with respect to the variable of interest then these traditional methods may lead to poor estimates. The pattern of clustering and patchiness is observed with many types of animals, vegetation types and contagious diseases. In such situations investigators are tempted to go away from the predefined sampling plan and add nearby or associated units in the sample.

Thompson (1990) introduced adaptive cluster sampling (ACS) design. In this design, whenever the observed value of a selected unit satisfies a condition of interest, additional units

Corresponding Author: Raosaheb Latpate
E-mail: rvl@unipune.ac.in

are added to the sample from the neighbourhood of that unit. ACS suffers from the drawback of losing the control on the final sample size. Salehi and Seber (1997) presented two stage sampling design in which primary units are selected by using a conventional design and secondary units within the selected primary units are selected by using ACS. The subsamples are not allowed to cross the borders of the primary units. This limitation imposes an artificial constraint on the adaptive response to encountered patterns and consequently tends to reduce the efficiency of the design. Latpate and Kshirsagar (2018 c) evaluated the expected sample size for the adaptive cluster sampling.

Lee(1998) developed a two phase version of ACS in which the first phase sample is an ACS sample based on an auxiliary variable and the second phase sample is selected from the first phase using probability proportional to size(PPS) with replacement sampling design. This design allows the sampler to control the number of measurements of the study variable but not that of the auxiliary variable. In this design the travelling costs are increased because the second phase sample is selected after the first phase sample is completed. The second phase sampling design does not allow the sampler to allocate the subsample near the places of interest. The proposed unbiased estimators of the population mean do not take the advantage of the relationship between the study variable and the auxiliary variable.

Martin Medina and Steven Thompson (2004) proposed adaptive cluster double sampling (ACDS). It is a method based on combining the idea of the double sampling and ACS. This method requires the availability of an inexpensive and easy to measure auxiliary variable. While developing this method, they have completely ignored the type of relationship between the variable of interest and the auxiliary variable. This method can be used when the complete information about the auxiliary variable is not available. In that case, in the first phase, using some condition of interest related to auxiliary variable, a sample of clusters in the population is identified. Then treating those clusters as primary stage units, the secondary stage units are selected from them. On the basis of a simulation study, Martin Medina and Steven Thompson have shown that ACDS is more efficient than ACS. ACDS controls the number of measurements of the study variable. It allows the sampler to allocate the second phase subsample near the places of interest. It allows starting the second phase sampling before the first phase sampling is completed. This method uses the auxiliary variable at the estimation stage.

Gattone et al. (2016) proposed adaptive cluster sampling for negatively correlated variables. They did not utilize the auxiliary information at estimation stage. But, they use the auxiliary information at estimation stage.

Latpate and Kshirsagar (2018 b) proposed two-stage negative adaptive cluster sampling design. It is a combination of two-stage sampling and negative adaptive cluster sampling (NACS) designs. In this design, they consider an auxiliary variable which is highly negatively correlated with the variable of interest and auxiliary information is completely known. In the first stage of this design, an initial random sample is drawn by using the auxiliary information. Further using Thompson's(1990) adaptive procedure, networks in the population are discovered. These networks serve as the primary stage units (PSU's). In the second stage, random samples of unequal sizes are drawn from the PSU's to get the secondary stage units (SSU's). The values of the auxiliary variable and the variable of interest are recorded for these SSU's.

In section 2 of this article, Latpate and Kshirsagar (2018 a) presented negative adaptive cluster sampling (NACS). This method assumes a high negative correlation between the variable of interest and the auxiliary variable. It also assumes the availability of complete information on the auxiliary variable. In this method, first we discover a sample of adaptive clusters in the population by using Thompson (1990) procedure on auxiliary information. Further, the values of the variable of interest are observed for the elements in these discovered clusters. With the help of a simulation study, they have shown that NACS is more efficient than ACDS.

In section 3 of this article, we have proposed method for estimating the mean/total of the variable of interest. This method is a two phase variant of the NACS obtained by combining the idea of the double sampling and NACS. We have named this method as negative adaptive cluster double sampling (NACDS). Here, we assume that the auxiliary information is easily available and is less expensive. It can be seen as a variant of ACDS. When the two variables have a positive correlation, the selection of units for observing the variable of interest based on the condition related to the auxiliary variable is justified. But, the nature of underlined population is rare and patchy. The auxiliary information is also rare and patchy. Hence, we could not exploit the auxiliary information at design stage. To exploit the auxiliary information at design stage; auxiliary information must be abundantly available, easy to measure and less costly. In such cases, we advocate our new method NACDS. This design covers the merits of double sampling and NACS. In NACDS, we have knowledge about the negative relationship between the variable of interest and the auxiliary variable. According to this relationship, the condition of adaptation related to auxiliary variable is reversed.

In section 4, we have proposed regression and ratio estimators given by S*ä*rndal and Swensson (1987), of population total of the variable of interest, using NACDS. Also, the product type ratio HT estimator is proposed for NACDS. The estimates of the variances of these estimators are also given.

Section 5 discusses a sample survey based on this method. Results and discussions are incorporated in section 6. Lastly, the concluding remarks are added in section 7.

2. Negative Adaptive Cluster Sampling

In ACS the units in the initial sample are identified whether they satisfy the desired condition C with respect to the variable of interest or not. Further the networks are expanded around the units in the initial sample that satisfy the condition C. Here, we propose different adaptive procedure. The variables are negatively correlated and the adaptive procedure involves the auxiliary variable instead of the variable of interest. We get the clusters of units during the adaptation. Hence this method is called the negative adaptive cluster sampling.

In ACDS, by using adaptation technique the first phase units are decided by using an auxiliary variable. Then by using some traditional method such as SRSWOR, the second phase units are selected. In NACS, the adaptation is used to discover the networks in the population with reference to the auxiliary variable. Further the networks corresponding to the variable of interest are identified. There is no second phase in NACS. That is how NACS is different than ACDS. So in general, NACS is not ACDS. But NACS can be looked upon as a particular case of

ACDS where the entire networks identified in the first phase, corresponding to the variable of interest are considered as the second phase units.

Secondly, ACDS does not bother about the type of relationship between the auxiliary variable and the variable of interest. In contrast to this NACS requires a negative relationship between the auxiliary variable and the variable of interest. The networks corresponding to the auxiliary variable and the variable of interest are discovered by using exactly the opposite conditions on the two variables. The use of auxiliary variable is justified by ACDS in the first phase sampling. In NACS, we assume that the population information of auxiliary variable is known. The networks are formed by using ACS with auxiliary information. The corresponding Y is observed only for those units which satisfy the condition C_X which based on auxiliary information. Here, the population is rare and clustered and we observe Y only for the units that satisfy the condition C_X . So, there is substantial reduction in sample size with respect to Y. This reduced sample size is called as the effective sample size.
Consider a population of N units which can be observed and measured with respect to variables X and Y which are negatively correlated. Suppose the population is rare with respect to the variable of interest (Y); equivalently we can say that it is highly abundant with respect to the auxiliary variable X. Taking observations on X is easy and inexpensive.The procedure of NACS is as follows:

Form a grid of population containing N grid points of equal size and shape. Draw an initial sample of size n grid points from this grid using simple random sampling without replacement (SRSWOR) or simple random sampling with replacement (SRSWR) method.

Check whether each of the selected units satisfies the condition C_X or does not satisfy the condition C_X. Add the unit to the left, right above and below to each unit included in the initial sample that satisfies the condition C_X. These units are called neighbours of that unit. If any of these neighbours satisfy the condition C_X, add their neighbors also to the sample. Continue this way till the neighbours that do not satisfy the condition C_X are found. The set of neighbour units satisfying the condition C_X along with the corresponding unit selected in the initial sample that satisfies the condition C_X constitutes a network. Thus in this design, the networks are formed around the units selected in the initial sample that satisfy C_X. Note that a unit selected in the initial sample which does not satisfy the condition C_X forms a network of size one.

Suppose K distinct clusters are formed with respect to X population. A cluster includes the units in a network and the corresponding edge units. Edge units do not satisfy the condition C_X. If all edge units in a cluster are dropped we get a network. From the K clusters, we get the K networks.

Observe the values of the variable of interest corresponding to all the units in these K networks. Further using the proposed estimators, the population total of Y can be estimated. Estimates of the standard error of these estimators can be obtained. If we drop the auxiliary information to get modified Hansen-Hurwitz and Horvitz-Thompson estimators then NACS reduces to ACS.

3. Negative Adaptive Cluster Double Sampling

Medina and Thompson (2004) introduced ACDS. It considers the auxiliary variable which is easy to measure and inexpensive. But it does not take into account the type of the relationship between the two variables. If all information on auxiliary variable is available and the two variables are negatively correlated then we propose the new method, negative adaptive cluster double sampling (NACDS). It is a combination of NACS and double sampling. It is a cost effective method. To exploit the auxiliary information at design and estimation stage; auxiliary information must be available, easy to measure and less costly. In such cases, we advocate our new method NACDS. This design covers the merits of double sampling and NACS. According to the relationship between the two variables, the condition of adaptation related to auxiliary variable is reversed. This design is generalization of NACS when the population parameters of auxiliary information is unknown.

Sampling Design and Notations

Let $U = \{u_{1,}u_{2,} \dots, u_N\}$ be a finite population of N units. Let Y and X be the interest and auxiliary variable respectively. They are known to be highly negatively correlated. Let X_i and $Y_i, i = 1,2 \dots, N$ be the values of X and Y respectively associated with the unit u_i. It is assumed that the information on auxiliary variable can be obtained from all the units selected in the sample. The goal is to estimate the population total of Y given by $\tau_Y = \sum_{i=1}^{N} Y_i$.
An initial sample of size n units is drawn from the population by using SRSWOR. We denote this initial sample drawn as S_0. From S_0, obtain an adaptive cluster sample S_1 by using the following procedure:

Denote the condition of interest with respect to X values by C_X. According to the negative correlation the condition is reversed for adaptation. Now following the procedure given by Thompson (1990), we add the neighbours of the units in S_0 that satisfy the condition C_X. The units to the right, left, above and below a unit are called as the neighbours of that unit. If any of these neighbours satisfy C_X then their neighbours are also added to the sample. This is continued till the neighbours not satisfying C_X are obtained. The units added to the sample S_0 adaptively which satisfy the condition C_X constitute a network. The units added to the sample S_0 adaptively which do not satisfy the condition C_X are called as the edge units. The set of units in a network along with its edge units is called as a cluster. The set of units included in all such clusters is called as an adaptive cluster sample. We denote it by S_1. Thus; indirectly we are assuming that the condition C_X for the additional sampling and a set of neighbouring units for each $u_i \in U$ have been defined.

Let K denote the number of distinct clusters formed by S_0.

Mark the corresponding K clusters in the Y population and drop down the edge units to get K networks. This completes the first phase of the design.

From each of these selected networks draw a sample by using SRSWOR. The sizes of these samples may be different. Suppose m_i denotes the number of units selected from the i th selected network. Collection of all these units selected be denoted by S_2. This completes the

second phase of sampling design. Now, note the values of X and Y for all the units included in S_2. This data is used to estimate the population parameter. In this design, the X value associated with every unit in the adaptive cluster sample S_1 has to be measured. Hence, the procedure does not control the number of observations on the auxiliary variable, but only the number of observations on the survey variable.

The first phase sample $S_1(S_1 \subset U)$ of size n_{s_1} is drawn by a design denoted by $P(.)$ such that $P(S_1)$ is the probability of choosing S_1. The inclusion probabilities are defined as follows:

Π_k : Probability that unit k is included in $S_1 = \sum_{k \in S_1} P(S_1)$
Π_{kl} : Probability that unit k and l are included in $S_1 = \sum_{k,l \in S_1} P(S_1)$
with $\Pi_{kk} = \Pi_k$

Let $\Delta_{kl} = \Pi_{kl} - \Pi_k \Pi_l$

We assume that $\Pi_k > 0$ for all k and $\Pi_{kl} > 0$ for all k$\neq l$.

Given S_1, the second phase sample S_2 $(S_2 \subset S_1)$ of size n_{s_2} is drawn according to a sampling design $P(.|S_1)$ such that $P(S_2|S_1)$ is the conditional probability of choosing S_2.
The inclusion probabilities given S_1 are defined by:
$\Pi_{k|S_1}$: Probability that k th unit is included in S_2 given $S_1 = \sum_{k \in S_2} P(S_2|S_1)$
$\Pi_{kl|S_1}$: Probability that k th and lth units are included in S_2 given $S_1 = \sum_{k,l \in S_2} P(S_2|S_1)$
with $\Pi_{kk|S_1} = \Pi_{k|S_1}$.

Let $\Delta_{kl|S_1} = \Pi_{kl|S_1} - \Pi_{k|S_1}\Pi_{l|S_1}$

We assume that for any $S_1, \Pi_{k|S_1} > 0$ for all k $\in S_1$ and $\Pi_{kl|S_1} > 0$ for all $k \neq l \in S_1$.
For all $k, l \in S_1$ and any S_1

$\Pi_k^* = \Pi_k \Pi_{k|S_1}$, $\Pi_{kl}^* = \Pi_{kl}\Pi_{kl|S_1}$ with $\Pi_{kk}^* = \Pi_k^*$
Let $\Delta_{kl}^* = \Pi_{kl}^* - \Pi_k^* \Pi_l^*$
$\breve{y}_k = \frac{y_k}{\Pi_k}$, $\breve{y}_k^* = \frac{\breve{y}_k}{\Pi_k^*}$ and $\breve{\Delta}_{kl} = \frac{\Delta_{kl}}{\Pi_{kl}}$

4. Proposed Estimators

Definition 1 (Robinson 1983): A predictor $(\hat{\tau}_y)_m$ is asymptotically unbiased if

$$\lim_{m \to \infty} (E((\hat{\tau}_y)_m | Y_m) - (\tau_y)_m) = 0$$

with ξ- probability one.

Definition 2 (Robinson 1983): A predictor $(\hat{\tau}_y)_m$ is consistent if, for all $\varepsilon > 0$

$$\lim_{m \to \infty} \mathrm{P}(|(\hat{\tau}_y)_m - (\tau_y)_m| > \varepsilon | Y_m) = 0$$

with ξ- probability one.

i) Regression Estimator

Using the ideas from Särndal and Swensson (1987) we propose the regression type estimator of the population total of Y. We assume that the relationship between $Y$and X can be modelled through a regression model ξ such that

$E_\xi(Y_j) = X_j'\beta$ and $Var_\xi(Y_j) = \sigma_j^2$, $j = 1,2,\dots m_i$; $i = 1,2,\dots,n$.

Where Y_j's are independent and $\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$.

The model ξ is just a tool to express the relationship between Y and X in the finite population.

If all the N points (Y_i, X_i) are available then the weighted least square estimator of β and the associated residuals are:

$\beta = \left(\sum_U \frac{X_j X_j'}{\sigma_k^2}\right)^{-1}\left(\sum_U \frac{X_j Y_j}{\sigma_k^2}\right)$ and $E_j = Y_j - X_j'\beta$

The estimators of β and E_j are given by

$$b = \left(\sum_{S_2} \frac{x_k x_k'}{\sigma_k^2 \pi_k^*}\right)^{-1}\left(\sum_{S_2} \frac{x_k y_k}{\sigma_k^2 \pi_k^*}\right) \text{ and } e_j = y_j - \hat{y}_j = y_j - x_j' b,\ j \in S_2$$

Where $\pi_k^* = \pi_k \pi_{k|S_1}$

The first phase regression estimator of T_Y is given by

$$\hat{\mathrm{T}}_{Y_{1R}} = \left[\sum_{S_1} \frac{\hat{y}_k}{\pi_k} + \sum_{S_2} \frac{(y_k - \hat{y}_k)}{\pi_k^*}\right]$$

Approximate variance of the estimator $\hat{\mathrm{T}}_{Y_{1R}}$ is given by

$$AV(\hat{\mathrm{T}}_{Y_{1R}}) = \left[\sum\sum_U \Delta_{kl} \breve{y}_k\ \breve{y}_l + E_1\left(\sum\sum_{S_1} \Delta_{kl|S_1} \breve{\breve{E}}_k\ \breve{\breve{E}}_l\right)\right]$$

where $\breve{\breve{E}}_k = \frac{E_k}{\pi_k^*}$

Unbiased estimator of $V(\hat{\mathrm{T}}_{Y_{1R}})$ is given by

$$\hat{V}(\hat{\mathrm{T}}_{Y_{1R}}) = \left[\sum\sum_{S_2} \frac{\breve{\Delta}_{kl} \breve{y}_k\ \breve{y}_l}{\pi_{kl|S_1}} + \left(\sum\sum_{S_2} \frac{\Delta_{kl|S_2} g_{k_{S_2}} g_{l_{S_2}} \breve{\breve{e}}_{k_{S_2}} \breve{\breve{e}}_{l_{S_2}}}{\pi_{kl|S_1}}\right)\right]$$

where $\breve{\breve{e}}_{k_{S_2}} = \frac{e_k}{\pi_k^*}$

$g_{k_{S_2}} = 1 + \left(\sum_{S_1} \frac{x_k}{\pi_k} - \sum_{S_2} \frac{x_k}{\pi_k^*}\right)' \left(\sum_{S_2} \frac{x_k x_k'}{\sigma_k^2 \pi_k^*}\right)^{-1} \frac{x_k}{\sigma_k^2}$.

ii) **Ratio Estimator**

If we assume the above regression model without intercept then the ratio estimator of the population total of the variable Y can be defined as

$$\hat{\mathbb{T}}_{YRE} = \left(\sum_{s_1} \frac{\mathbb{T}_{x_k}}{\pi_k} \right) \hat{B}$$

Where $\hat{B} = \left(\sum_{S_2} \frac{x_k x_k'}{\sigma_k^2 \pi_k^*} \right)^{-1} \left(\sum_{S_2} \frac{x_k y_k}{\sigma_k^2 \pi_k^*} \right)$

$\mathbb{T}_{x_k}$: Total of X observations in the k^{th} network selected in the sample by using design s_1
$\mathbb{T}_{x_k} = \sum_{s_1} x_k$
π_{I_k} =Inclusion probability of the k^{th} network in the sample selected by using design s_I .
The prediction equation is
$\hat{y}_{k=} \hat{B}\, x_k, \qquad k \in S_1$

The residuals are

$$e_{k_{S_2}} = y_k - \hat{B}\, x_k$$

Its g weights are
$g_{k_{S_2}} = \frac{\sum_{S_1} \breve{x}_k}{\sum_{S_2} \breve{\breve{x}}_k}$ where $\breve{x}_k = \frac{x_k}{\pi_k}$ and $\breve{\breve{x}}_k = \frac{x_k}{\pi_k^*}$

The approximate variance of the above estimator of the population total of Y can be written as

$$V\left(\hat{\mathbb{T}}_{YRE}\right) = \sum\sum_{U} \Delta_{kl} \breve{y}_k\, \breve{y}_l + E_1 \left(\sum\sum_{S_2} \Delta_{kl|S_2} \breve{E}_k\, \breve{E}_l \right)$$

An unbiased estimator of this variance is given by

$$\hat{V}\left(\hat{\mathbb{T}}_{YRE}\right) = \left[\sum\sum_{S_2} \frac{\breve{\Delta}_{kl} \breve{y}_k\, \breve{y}_l}{\pi_{kl|S_1}} + \left(\sum\sum_{S_2} \frac{\Delta_{kl|S_2} g_{k_{S_2}} g_{l_{S_2}} \breve{\breve{e}}_{k_{S_2}} \breve{\breve{e}}_{l_{S_2}}}{\pi_{kl|S_1}} \right) \right]$$

iii) **Product Type Ratio HT Estimator:**

Since the two variables are negatively correlated, it is of interest to define a product estimator of the population total based on the HT estimators of the population totals of the two variables by using Murthy's (1964) product type estimator.

The product estimator of the population total $\boldsymbol{\tau}_y$ is defined as:

$$\hat{\boldsymbol{\tau}}_{PR} = (\hat{\boldsymbol{\tau}}_{IIy})_{HT} \left(\frac{(\hat{\tau}_{IIx})_{HT}}{(\hat{\tau}_{Ix})_{HT}} \right)$$

We assume the large sample approximation to obtain the MSE.

Let $e_1 = \frac{(\hat{\tau}_{IIy})_{HT} - \tau_y}{\tau_y}$

$$e_2 = \frac{(\hat{\tau}_{IIx})_{HT} - \tau_x}{\tau_x}$$

$$e_3 = \frac{(\hat{\tau}_{Ix})_{HT} - \tau_x}{\tau_x}$$

Where, $(\hat{\tau}_{IIy})_{HT}$ is HT estimator of the study variable at second stage.

$$(\hat{\tau}_{IIy})_{HT} = \sum_{S_2} \frac{y_k}{\pi_{i/k}\pi_k}$$

$(\hat{\tau}_{IIx})_{HT}$ is HT estimator of the auxiliary variable at second stage.

$$(\hat{\tau}_{IIx})_{HT} = \sum_{S_2} \frac{x_k}{\pi_{i/k}\pi_k}$$

$(\hat{\tau}_{Ix})_{HT}$ is HT estimator of the auxiliary variable at first stage.

$$(\hat{\tau}_{Ix})_{HT} = \sum_{S_1} \frac{x_k}{\pi_k}$$

We get, $E(e_1) = E(e_2) = E(e_3) = 0$, and

$$E(e_1^2) = \frac{1}{2}\frac{1}{\tau_y^2}\left[\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + E_I\left(\sum\sum_{i\neq j\in S_1}\left(\pi_{\frac{i}{S_I}}\pi_{\frac{j}{S_I}} - \pi_{\frac{ij}{S_I}}\right)\left(\frac{y_i}{\pi_i^*} - \frac{y_j}{\pi_j^*}\right)^2\right)\right]$$

$$E(e_2^2) = \frac{1}{2}\frac{1}{\tau_x^2}\left[\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 + E_I\left(\sum\sum_{i\neq j\in S_1}(\pi_{i/S_1}\pi_{j/S_1} - \pi_{ij/S_1})\left(\frac{x_i}{\pi_i^*} - \frac{x_j}{\pi_j^*}\right)^2\right)\right]$$

$$E(e_3^2) = \frac{1}{2}\frac{1}{\tau_x^2}\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$$

$$E(e_1e_2) = \frac{1}{2\tau_x\tau_y}\left[\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + E_I\left(\sum\sum_{i\neq j\in S_1}(\pi_{i/S_I}\pi_{j/S_I} - \pi_{ij/S_I})\left(\frac{x_i}{\pi_i^*} - \frac{y_j}{\pi_j^*}\right)^2\right)\right]$$

$$E(e_2e_3) = \frac{1}{2}\frac{1}{\tau_x^2}\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$$

$$E(e_1e_3) = \frac{1}{2\tau_x\tau_y}\left[\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right]$$

Hence,

$$\widehat{\tau}_{PR} = \tau_y(1+e_1)(1+e_2)(1+e_3)^{-1}$$

By neglecting the terms involving e_x with power three and above we get,

$$\widehat{\tau}_{PR} \cong \tau_y(1+e_1+e_2-e_3+e_1e_2-e_1e_3-e_2e_3+e_3^2)$$

$$Bias\,(\widehat{\tau}_{PR}) = E(\widehat{\tau}_{PR}) - \tau_y$$

$$= \tau_y\left[E_I\left(\sum\sum_{i\neq j\in S_1}(\pi_{i/S_I}\pi_{j/S_1}-\pi_{ij/S_1})\left(\frac{x_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2\right)-\frac{1}{2}\frac{1}{\tau_x^2}\sum\sum_{i\neq j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2\right]$$

The mean square error of $\widehat{\tau}_{PR}$ is given as

$$MSE(\widehat{\tau}_{PR}) = E(\widehat{\tau}_{RADE}-\tau_y)^2$$

$$= \tau_y^2 E(e_1^2+e_2^2+e_3^2+2e_1e_2-2e_1e_3-2e_2e_3)$$

$$MSE(\widehat{\tau}_{PR}) = \frac{1}{2}\left[\sum\sum_{i\neq j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+E_I\left(\sum\sum_{i\neq j\in S_1}(\pi_{i/S_1}\pi_{j/S_1}-\pi_{ij/S_1})\left(\frac{y_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2\right)\right]+\frac{\tau_y}{\tau_x}\left[\frac{\tau_y}{2\tau_x}+1\right]\left[E_I\left(\sum\sum_{i\neq j\in S_1}(\pi_{i/S_1}\pi_{j/S_1}-\pi_{ij/S_1})\left(\frac{x_i}{\pi_i^*}-\frac{x_j}{\pi_j^*}\right)^2\right)\right]$$

5. Sample Survey

A sample survey was conducted by using NACDS. The area of 400 acres in the Tamhini Ghat, Maharashtra, India was divided into 400 plots each of size 1 acre. A random sample of 12 plots was drawn from this area by using SRSWOR. The percentage of silica content of the soil (X) was measured on these selected plots. Silica is abundant in the soil from Tamhini Ghat to Mumbai. But, there are intermediate patches of laterite where the occurrence of evergreen plants is more. We considered the condition $C_X = \{X \leq 20\} = \{\%\, silica \leq 20\}$. Further the plots in the sample satisfying C_X were located. Then the clusters were formed around these plots by using the procedure given by Thompson (1990). Each plot with $\{X > 20\}$ and selected in the initial sample formed a cluster of size one. Here the clusters were formed by using auxiliary information and the domain knowledge of silica content and evergreen plants. The two variables, percentage of silica content(X) and number of evergreen plants (Y) are negatively correlated. After forming such clusters in the X population, the edge units of clusters of size more than one were dropped to get networks. The networks are formed by using percent of silica content. The corresponding networks of number of evergreen plants are located. Figures 1 and 2 given in the Appendix illustrate this methodology.

These plots formed the first phase sample S_1. Let K denote the number of distinct networks represented in this sample. A random sample of m_i (say), $(i = 1,2,\ldots,K)$ units was drawn from the i[th] network among these K networks by using SRSWOR. The collection of all so selected units formed the second phase sample S_2. In our study there were 12 networks formed in S_1. We took $m_1 = m_2 = m_3 = 2, m_4 = 3, m_5 = 4,, m_6 = 2, m_7 = 2, m_8 = 2, m_9 = m_{10} = 4, m_{11} = 3$ and $m_{12} = 0$. This set of units formed S_2.Values of the variables X and Y corresponding to the plots included in the second phase sample were recorded together to form a bivariate data. Using this data, the total number of evergreen plants in that area was estimated by using the proposed estimators.

6. Results and Discussion

For the computational efficiency in estimation, r number of repetitions were performed; where r varied as 5000, 10000, 20000 and 100000. We considered the initial sample sizes as $5, 10, 15, 20$ and 25 for each repetition.
The estimated population total over r repetitions is given by

$$\hat{\tau}_Y = \frac{\sum_{i=1}^{r} \hat{\tau}_{Y_i}}{r}$$

Where $\hat{\tau}_{Y_i}$ denotes the estimated value of an estimator of the population total of the variable Y for the i[th] repetition.

The estimated mean square error of the estimator of population total of the variable Y is given by

$$\widehat{MSE}(\hat{\tau}_Y) = \frac{\sum_{i=1}^{r}(\hat{\tau}_{Y_i} - \tau_Y)^2}{r}$$

The results are shown in Table 1.
The final adaptive sample size in NACS is denoted by n_s and it is given as:

$$n_s = n + number\ of\ units\ added\ by\ adaptation\ procedure$$

After dropping down the edge units the left over sample size is called as the effective sample size(n_e).

$$Effective\ sample\ size\ (n_e) = \sum_{j=1}^{n_s} \delta_{C_x}(j)$$

Where

$$\delta_{C_x}(j) = \begin{cases} 1 & ; \quad if\ U_j \text{ satisfies the condition } C_x \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

Effective sample size at the second phase is denoted by n_e^* and it is given as:
$n_e^* = \sum_{i=1}^{K} \delta_i m_i$. Where,

$$\delta_i = \begin{cases} 1 & ; \text{ if the ith cluster is selected in the sample in the first phase} \\ 0 & ; \ o\text{therwise.} \end{cases} \quad (i = 1,2, \dots, K)$$

The total expected sample size for the NACDS is the sum of expected sample size at first stage and second stage.

Expected sample size for NACDS$= n_s + \sum_{i=1}^{K} \delta_i m_i$

For the computational efficiency in estimation, r number of repetitions were performed; where r varied as 5000, 10000, 20000 and 100000. We considered the initial sample sizes as $5, 10, 15, 20$ and 25 for each repetition and calculated the expected sample sizes of the above three types. The results are given in Table 2.

It was observed that the expected final sample size and the expected effective sample size increase as the initial sample size increases. But a remarkable reduction in the expected effective sample size at the second phase of the design was observed.
In general the relation $E(n_s) > E(n_e) > E(n_e^*)$ was observed.

If the costs of sampling per unit with respect to X and Y variables are $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ respectively then the total cost of sampling in ACS design $= (\boldsymbol{C_1} + \boldsymbol{C_2})\, \boldsymbol{E}(n_s)$.
The total cost of sampling in NACS design = $\boldsymbol{C_1 E}(n_s) + \boldsymbol{C_2}\mathrm{E}(\mathrm{n_e})$

On the other hand, the total cost of sampling in NACDS = $\boldsymbol{C_1 E}(n_s) + \boldsymbol{C_2 E}(\sum_{i=1}^{K} \boldsymbol{\delta_i m_i})$.

It can be easily seen that $\sum_{i=1}^{k} \boldsymbol{m_i} \leq n_e$. Hence, cost of sampling in NACDS is usually lesser than that in NACS.

We have calculated the expected costs of sampling in ACS, NACS and NACDS and the results are given in Table 3.

It was found that, in general, Expected Sampling Cost in ACS > Expected Sampling Cost in NACS > Expected Sampling Cost in NACDS. Thus the new design is cost effective as compared to NACS and ACS. Also, NACS estimator is more precise as compared to NACDS. The regression estimator is more precise as compared to ratio estimator and product type ratio HT estimator for NACDS. The product type ratio HT estimator is more efficient as compared to ratio estimator. Similar to ACDS, in our method also we can start selecting the second phase units before completing the selection of the first phase units. Due to this, the survey does not require to visit the same plots again. It saves the travelling cost, time required to collect the information. At the first phase, we consider only auxiliary information. Hence, there is considerable cost saving as compared to ACS.

In NACS, it is required to take observations on interest variable related to all units included in the networks. In NACDS, we take a subsample from each of the networks discovered at the first phase. Usually, taking observations on the interest variable is costly hence NACDS is definitely cost effective. The methodology of Adaptive Cluster Double Sampling is introduced by Medina and Thompson (2004). But, they present only the simulation study. Also, they haven't presented the real world example and implementation.
Remark: The estimators used in NACDS and ACDS are the same. Hence, in terms of precision, both the methods are equally efficient.

7. Conclusions

The domain knowledge of population is utilized to conduct the survey of rare population. If the study variable is rare and we have abundant auxiliary information then there is negative correlation. The auxiliary information is utilized at the design and estimation stages in NACDS. The regression estimator is more precise as compared to ratio estimator and product type ratio HT estimator for NACDS. The product type ratio HT estimator is more efficient as compared to ratio estimator. ACDS and NACDS are equally efficient. The expected effective sample size for NACDS is very small as compared to ACS and NACS. Hence, NACDS is cost effective. There is only condition for adaptation on auxiliary variable.

If the type of the correlation between the auxiliary variable and study variable is known to be negative, it can be used in environmental, forestry, ecological, social science and medical surveys.

Acknowledgements: The authors would like to thank Prof. Ashis Sengupta, Department of Applied Statistics, ISI, Kolkata for his valuable guidance.

References

Brown, J.A. (1994). The application of adaptive cluster sampling in ecological studies. *Statistics in Ecology and Environmental Monitoring*, **2**, 86-97.

Cassel, C.M., S*ä*rndal, C.E. and Wretman, J.H. (1977). *Foundation of Inference in Survey Sampling*. Wiley, New York.

Felix, M-H, Medina and Thompson, S. K. (2004). Adaptive cluster double sampling. *Biometrika***, 91**, 877-891.

Gattone, S.A., Mohamed, E., Dryver, A.L. and Munich, R.T. (2016). Adaptive cluster sampling for negatively correlated data. *Environmetrics*, **27**, E103-E113.

Latpate, R.V. and Kshirsagar, J. (2018a). Negative adaptive cluster sampling. *Model Assisted Statistics and Applications*. DOI 10.3233/MAS-180452.

Latpate, R.V. and Kshirsagar, J. (2018b). Two stage negative adaptive cluster sampling. *Communications in Mathematics and Statistics*. (Springer) DOI: 10.1007/s40304-018-0151-z. https://link.springer.com/content/pdf/10.1007%2Fs40304-018-0151-z.pdf

Latpate, R.V. and Kshirsagar, J. (2018c). Sample size considerations in the adaptive cluster sampling. *Bulletin of Marathwada Mathematical Society.* 19(1), 32-41.

Lee, K. (1998). Two phase adaptive cluster sampling with unequal probabilities selection. *Journal of the Korean Statistical Society*, **27**, 265-78.

Murthy, M.N. (1964). Product method of estimation. *Sankhya: The Indian Journal of Statistics, A***26**(**1**), 69-74.

Robinson, P.M. and S*a*rndal, C.E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya B***45**, 240-248.

Salehi, M.M. and Seber, G.A.F. (1997). Two stage adaptive cluster sampling. *Biometrics*, **53**, 959-70.

S*ä*rndal, C.E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non response. *International Statistical Review*. **55(3)**, 279-294.

S*ä*rndal, C.E, Swensson, B. and Wretman, J. (1992). Model assisted survey sampling. *Springer-Verlag Inc., New York* ,

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, **85(412)**, 1050-1058.

Thompson, S. K. (2002). *Sampling.* Second Edition, *A Wiley-Interscience Publication.*

Appendix

Figure 1. Silica (SiO_2) % on the different plots in the region

24	25	86	60	52	35	65	50	60	1	22	23	83	48	30	56	43	52	1	4
40	30	30	75	18	19 *	55	30	4	14	38	27	27	14	14	49	23	6	6	10 *
45	48	56	23	15	17	53	30	13	12 *	43	45	53	11	12 *	47	23	7	8	7
47	47	23	25	80	60	45	45	35	70	45	44	20	76	55	39	38	27	61	34
48	50	25	35	57	68	40	23	80	40	46	47	22	53	63	34	26	72	31	37
49	43	36	65	58	58	90	45	90	30	47	40	33	54	53	84	38	82	21	30
45	35	56	85	19	30	18	18	40	50	43	32	53	25	25	42	41	32	41	22
48	53	65	55	13	16 *	15	18	30	60	46	50	62	29	17	18	51	22	51	40
70	30	17	18	15	48	44	44	35	50	68	27	24	29	43	19 *	12	28	41	27
30	30	18	17	15	43	36	50	80	36	28	27	25	24	22	14	43	72	27	27
29	31	93	68	61	45	66	52	63	25	27	29	93	42	32	59	47	57	27	42
45	36	37	83	27	29	56	32	37	48	43	36	37	88	26	52	27	41	43	58
50	54	63	31	20	27	54	32	76	77	16	18	63	85	24	50	27	7	14	20
52	53	30	15	18	20	46	47	38	75	18 *	21	20	70	57	42	42	12	11 *	9
53	57	32	19	18 *	70	41	25	83	45	51	20	32	47	65	37	20	10	18	15
54	50	43	73	16	78	91	47	93	24	26	42	43	59	55 *	87	42	87	48	38
50	42	63	93	67	68	29	28	20	19	12	24	63	33	27	35	35	37	34	30
53	60	72	63	28	40	26	20	19 *	13	11	23	70	35	29	65	85	5	8	10
75	37	24	26	22	26	45	46	12	10	9	28	24	45	45	22	26	14	9 *	5
35	37	25	25	24	58	37	52	18	16	14	48	25	52	24	77	47	77	9	11

* in a square indicates selection in initial sample.

Figure 2. Number of evergreen plants observed on the plots in the population

									405%									350	305
			N1	35%	20*		**N2**	306	130			**N3**	125	130		**N4**	225%	206%	120*
				100%	65			107	108*				170%	155*%			175	167	188
				15		40	40												
		N5		120%	65*	95	30%							73%	67				
		75%	32	91										**N6**	62*	98			
		36	35	81%										53	87				
				18						90%	78						211	93	45
		N7	83	100%	55				**N8**	43*%						**N9**	97%	105*	175%
			68	71*													144%	68	83%
				75%										**N12**	0*				
								52	47	95%									
							41	63	93	78							290%	175	125%
						N11		91	115%	200%						**N10**	95%	127*	268%
								23	35	68								160	152

N1 to N12 denote the network numbers. % in a plot indicates the selection at phase two.

Table 1: Estimated values of population total of interest variable and its SE by NACS design(Regression Estimator) and by NACDS(Regression Estimator, Ratio Estimator and Product type Ratio HT Estimator)

		Regression Estimator for NACS		**Regression Estimator, Ratio Estimator and Product type Ratio HT Estimator for NACDS**					
Number of samples	**Initial Sample Size(n)**	$\hat{\tau}_y$	$\widehat{SE}(\hat{\tau}_y)$	$(\hat{t}_y)_{Reg}$	$\widehat{SE}(\hat{t}_y)_{Reg}$	$(\hat{t}_y)_{Ratio}$	$\widehat{SE}(\hat{t}_y)_{Ratio}$	$\hat{\tau}_{PR}$	$\widehat{MSE}(\hat{\tau}_{PR})$
5000	5	9092.12	9813.04	9139.53	9295.75	9555.85	10461.68	9105.24	9636.09
	10	9073.51	6291.11	9095.31	6493.03	9206.55	6778.47	8937.40	6526.64
	15	9067.73	4982.09	9073.74	5154.54	9230.12	5373.81	9021.54	5174.64
	20	9057.88	4200.33	9057.41	4378.46	9106.03	4595.43	9044.97	4387.01
	25	9041.63	3588.37	9043.17	3809.32	9144.12	3992.87	9012.65	3832.03
10000	5	9091.96	9448.93	9117.21	9239.01	9458.85	10383.6	9049.83	9559.09
	10	9081.00	6252.73	8971.90	6440.05	9223.40	6738.29	9078.18	6454.95
	15	9055.30	4937.97	8959.31	5147.22	9114.49	5328.22	9079.26	5162.03
	20	9028.43	4142.43	9098.10	4357.54	9173.93	4577.09	068.40	4359.69
	25	9020.88	3621.87	9062.58	3808.90	9110.27	3964.64	043.36	3819.96

20000	5	9114.37	9619.88	9095.66	9226.77	9421.81	10328.04	8957.53	9475.32
	10	9090.99	6226.61	9086.71	6418.04	9224.23	6724.55	8971.37	6469.45
	15	9035.02	4981.76	9071.33	5134.35	9097.89	5313.16	9034.15	5143.68
	20	9022.01	4158.73	9056.96	4346.72	9047.11	4529.84	9039.21	4360.50
	25	9039.70	3627.15	9042.25	3788.86	9107.88	3947.24	9043.41	3812.29
100000	5	9196.61	9470.26	9069.59	9203.23	9431.05	10295.4	9007.55	9435.50
	10	9138.80	6097.93	9067.15	6350.09	9222.82	6676.93	9026.92	6457.02
	15	9077.27	4816.89	9052.93	5101.60	9115.33	5266.15	9030.47	5138.30
	20	9056.61	4079.37	9049.11	4332.31	9118.94	4496.88	9032.90	4338.40
	25	9047.59	3566.08	9032.29	3748.15	9136.57	3902.81	9042.67	3801.48

Table 2: Estimated values of Final Sample Sizes, Effective Sample Sizes and Effective Sample Sizes for Second Phase of NACDS.

Number of samples	**Initial Sample Size(n)**	**Expected Final Sample Size (n_S)**	**Expected Effective Sample Size (n_e)**	**Expected Effective Sample Size for Second Phase**(ne^*)
5000	5	12.99	8.84	3.07
	10	24.32	16.19	5.65
	15	35.33	23.26	8.12
	20	45.31	29.28	10.26
	25	55.14	35.18	12.31
10000	5	12.81	8.65	3.01
	10	24.38	16.27	5.68
	15	35.51	23.48	8.20
	20	45.66	29.67	10.37
	25	54.99	35.04	12.28
20000	5	12.97	8.82	3.06
	10	24.43	16.31	5.69
	15	35.53	23.50	8.21
	20	45.55	29.57	10.34
	25	54.99	35.04	12.29
100000	5	12.84	8.67	3.02
	10	24.57	16.47	5.74
	15	35.35	23.29	8.14
	20	45.45	29.45	10.31
	25	54.99	35.05	12.29

Table 3: Expected Sampling costs in ACS, NACS and NACDS

Number of samples	Initial Sample Size(n)	Expected Sampling Cost in ACS	Expected Sampling Cost in NACS	Expected Sampling Cost in NACDS
5000	5	324.75	259.97	115.72
	10	608.00	477.71	214.21
	15	883.25	687.49	308.99
	20	1132.75	867.93	392.43
	25	1378.50	1044.92	473.17
10000	5	320.25	254.68	113.68
	10	609.50	479.89	215.14
	15	887.75	693.53	311.53
	20	1141.50	878.73	396.23
	25	1374.75	1040.97	471.97
20000	5	324.25	259.41	115.41
	10	610.75	481.04	215.54
	15	888.25	694.09	311.84
	20	1138.75	875.90	395.15
	25	1374.75	1040.97	472.22
100000	5	321..00	255.27	114.02
	10	614.25	485.46	217.21
	15	883.75	688.30	309.55
	20	1136.25	872.60	394.10
	25	1374.75	1041.22	472.22