Thailand Statistician
January 2020; 18(1): 43-54
http://statassoc.or.th
Contributed paper

Application of Midzuno Scheme in Adaptive Cluster Sampling

Prayad Sangngam* and Wipawan Laoarun
Department of Statistics, Faculty of Science, Silpakorn University, Nakorn Pathom, Thailand.
*Corresponding author; e-mail: sangngam_p@su.ac.th

Received: 9 October 2018
Revised: 25 February 2019
Accepted: 5 March 2019

Abstract

This paper proposes an adaptive cluster sampling using unequal probability without replacement for selecting an initial sample. Midzuno scheme was applied for selecting an initial sample in adaptive cluster sampling. Two unbiased estimators of the population total are proposed. The variances of the proposed estimators and their unbiased estimators were also derived. A small population was also used to show the unbiased property of the estimators under the proposed sampling design. The simulation study was used to compare the efficiency of the proposed sampling design to the original adaptive cluster sampling. The auxiliary variable is created to construct the initial probability. The coefficient of correlation between the study variable and auxiliary variable consists of 0.3, 0.5, 0.7 and 0.9. The results showed that the proposed sampling design was more efficient than the original adaptive cluster sampling. In particular, when the correlation coefficient between the auxiliary and the study variables increases, the proposed sampling scheme was more efficient. In addition, the units in the initial sample are easy to draw and the proposed estimates are easy to compute.

Keywords: Adaptive cluster sampling, unbiased estimator, Horvitz-Thompson estimator, rare population.

1. Introduction

It is difficult to find the most efficient sampling design for a rare and clustered population. In order to collect samples from this population, Thompson (1990) proposed adaptive cluster sampling designs and demonstrated that an adaptive cluster sampling strategy can be more efficient than a simple random sampling strategy. In applications, the design can be used to estimate the number of rare plants or animals in a given area or to draw a hidden human population. For example, Smith et al. (2003) applied adaptive cluster sampling to survey freshwater mussels. For the simplest form of adaptive cluster sampling, an initial sample of units is drawn by simple random sampling. Whenever the value of study variable from a sampled unit in the initial sample satisfies a specified condition, its neighboring units are added to be sampled. If the values of study variable from the added neighboring units satisfy the condition, then their neighborhoods of these units are also added to be sample. This procedure is continued until none of units satisfy the condition. Many researchers studied about adaptive cluster sampling. Thompson (1991a) considered an adaptive cluster sampling in which the initial sample is selected by stratified sampling. Thompson (1991b) proposed an adaptive cluster sampling when the initial sample is drawn by systematic sampling.

When we use the simple random sampling to draw an initial sample, the initial sample might not contain units of interest. Therefore, the estimates do not use the information form units of interest. However if the size measure is available which is positively correlated with the study variable, it is advantage to select units with unequal probability. Roesch (1993) presented an adaptive cluster sampling with initial unequal probability design with replacement. Smith et al. (1995) compared the efficiency of four sampling designs using simulation study: simple random sampling, unequal probability sampling, adaptive cluster sampling with initial simple random sample and adaptive cluster sampling with initial unequal probability sample with replacement. Sangngam (2013) considered adaptive cluster sampling with initial unequal probability inverse sample with replacement. In theorem, for a given sample size, sampling with replacement is usually less efficient than sampling without replacement.

There are many sampling procedures in which units are drawn with unequal probability without replacement. One was introduced by Midzuno (1952). In this procedure, the first unit in a sample is drawn by unequal probability sampling and the remaining units in the sample will be drawn by simple random sampling. In this procedure, the units are easy to be drawn because only the first unit is selected with unequal probabilities. In addition, the initial probabilities of all units in the sample are used to construct the unbiased estimators of population total.

This paper applies Midzuno scheme to select an initial sample in adaptive cluster sampling. Unbiased estimators of the population total are derived. The variances of the unbiased estimators and their unbiased estimators are also derived. A small population is used to demonstrate the computation of the estimates and to study the properties of the estimates. The simulation study is used to compare the efficiency of the proposed sampling strategies to the original adaptive cluster sampling. These results can be used to suggest the researchers to select the suitable sampling design for rare and clustered populations.

2. Proposed Sampling Design

Suppose that a finite population consists of N distinct units with label $1,2,\ldots,N.$ Associated with the N units are the values of a study; $y_1, y_2, \ldots, y_N.$ Let $x_1, x_2, \ldots, x_N$ be the size measures of the units and assume that the measure of sizes are known before selection. Let $z_i = x_i / X_0$ be an initial selection probability of the i^{th} unit where $X_0 = \sum_{i=1}^{N} x_i.$ The parameter to be estimated is the population total $\tau = \sum_{i=1}^{N} y_i.$

For every unit i in the population, the neighborhood of a unit is defined as a collection of units which includes the unit $i.$ These neighborhoods do not depend on the study values, $y_i.$ The neighborhoods are symmetric; if unit i is in the neighborhood of unit $j,$ then unit j is also in the neighborhood of unit $i.$ The condition for selecting neighborhood units is given by $C = \{y : y \geq c\}$ where c is a given constant. The unit i satisfies the condition if the study value y_i is greater than or equal to the constant $c.$

The proposed sampling procedure consists of an initial sample of size n to be selected by Midzuno scheme and other units to be drawn by adaptive sampling. The sampling procedure can be implemented using the following method. To draw the initial sample of size $n,$ the first unit is drawn by using the initial selection probability, and other units in the initial sample are drawn by simple random sampling without replacement. Whenever each study value of a unit in the initial sample satisfies the condition C, its neighborhood units are added to be sampled and observed. For any units in the added neighborhood, if they satisfy the condition C, their neighborhoods are also included to the sample and observed. The procedure continues until none of units satisfy the condition. The final sample of size n_1 consists of the initial sample and all adaptively units. This sampling scheme combines the concept of adaptive cluster sampling and unequal probability sampling without replacement.

The collection of all units that are observed from an initial unit i is called *cluster*. Within a cluster, a subcollection of units is called a *network*, with the property that if any units within a network are selected, every other unit in the network is also included. The units that are adaptively sampled that did not satisfy the condition are called *edge units*. By this way, if any unit in the k^{th} network is selected in the initial sample, all units in that network will be included in the final sample. Any unit not satisfying the condition is called *network of size* 1. From definition of network, the population can be divided into K mutually exclusive networks.

Let s_0 be the set of units under the initial sample. With the Midzuno scheme, the probability of getting the initial sample (Sampath 2005, pp.73-74) is

$$P(s_0) = \frac{1}{\binom{N-1}{n-1}} \sum_{i \in s_0} z_i .$$

Let s_1 be the final sample. Under the proposed sampling scheme, the probability of getting the final sample is

$$P(s_1) = \frac{1}{\binom{N-1}{n-1}} \sum_{s_0 \approx s_1} \sum_{i \in s_0} z_i ,$$

where $\sum_{s_0 \approx s_1}$ refers to the summation of all initial samples leading to the final sample s_1.

If the initial probability of selecting the unit i equals to $1/N$ for every unit in the population, the proposed sampling design will become the design of Thompson (1990).

3. Proposed Estimators

In this section, we would like to derive the estimators of population total and the variances of these estimators. Finally, the unbiased property of the population total estimators and the variances of these estimators can be also illustrated.

For any sampling designs, if the probability that unit i will be drawn into the sample is known for every unit in the population, the Horvitz-Thompson estimator is an unbiased estimator of the population total. With the proposed sampling design, the unit i will be included in the sample if either some units in its network are selected to be the sample or any unit of a network of which unit i is an edge unit is drawn to be the sample. Unfortunately, under the proposed design, these inclusion probabilities might be unknown for some units in the final sample.

The first unbiased estimator of the population total is derived by applying the Horvitz-Thompson estimator. The new study value of a population unit i is the mean of study values in a network which includes the i^{th} unit. The observations not satisfying the condition will not be used in the estimator except when they are included in the initial sample. Let ψ_k be the set of units comprising the k^{th} network and m_k be the number of units in the network k. The total and the average of study values in the k^{th} network is represented by $y_k^* = \sum_{j\in\psi_k} y_j$ and $\overline{y}_k^* = \frac{1}{m_k}\sum_{j\in\psi_k} y_j$, respectively. The population total can be written as $\tau = \sum_{i=1}^{N}\overline{y}_i^* = \sum_{i=1}^{N} y_i$. In order to obtain an unbiased estimator, the study value y_i will be replaced by the new study value given by $\overline{y}_i^*$. Under Midzuno scheme in Sampath (2005, pp.74-76), the probability that the i^{th} unit will be selected to be an initial sample is

$$\pi_i = 1-(1-z_i)\frac{\binom{N-2}{n-1}}{\binom{N-1}{n-1}}.$$

In addition, the probability that both unit i and unit j are selected in the initial sample is given by

$$\pi_{ij} = 1-(1-z_i)\frac{\binom{N-2}{n-1}}{\binom{N-1}{n-1}}-(1-z_j)\frac{\binom{N-2}{n-1}}{\binom{N-1}{n-1}}+(1-z_i-z_j)\frac{\binom{N-3}{n-1}}{\binom{N-1}{n-1}}.$$

Theorem 1 *Under the initial sample s_0 of size n in the proposed sampling design, an unbiased estimator of the population total is*

$$\hat{\tau}_1 = \sum_{i=1}^{n}\frac{\overline{y}_i^*}{\pi_i}. \tag{1}$$

The variance of $\hat{\tau}_1$ is given by

$$V(\hat{\tau}_1) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\overline{y}_i^*\overline{y}_j^*. \tag{2}$$

An unbiased estimator of this variance is

$$\hat{V}(\hat{\tau}_1) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{\overline{y}_i^*\overline{y}_j^*}{\pi_{ij}}. \tag{3}$$

Proof: Let $\overline{y}_i^*$ be a new study value of the i^{th} unit for $i = 1,2,\ldots,N$. We knew that $\tau = \sum_{i=1}^{N}\overline{y}_i^*$.

Define the indicator function $I_i = \begin{cases} 1 \text{ ; the } i^{th} \text{ unit is included in the initial sample} \\ 0 \text{ ; otherwise.} \end{cases}$

The estimator $\hat{\tau}_1$ can be written as $\hat{\tau}_1 = \sum_{i=1}^{N}\frac{\overline{y}_i^*}{\pi_i}I_i$. Since $E[I_i] = \pi_i$, using Horvitz-Thompson approach, we can prove that

$$E\left[\hat{\tau}_1\right]=\sum_{i=1}^{N}\frac{\overline{y}_i^*}{\pi_i}E\left[I_i\right]=\sum_{i=1}^{N}\overline{y}_i^*=\tau.$$

That is $\hat{\tau}_1=\sum_{i=1}^{n}\frac{\overline{y}_i^*}{\pi_i}$, is an unbiased estimator of the population total, τ.

Since $E\left[I_iI_j\right]=\pi_{ij}$, we have

$$\begin{aligned}V(\hat{\tau}_1)&=\sum_{i=1}^{N}V\left[\frac{\overline{y}_i^*}{\pi_i}I_i\right]+\sum_{i=1}^{N}\sum_{j\neq i}^{N}Cov\left[\frac{\overline{y}_i^*}{\pi_i}I_i,\frac{\overline{y}_j^*}{\pi_j}I_j\right]\\&=\sum_{i=1}^{N}\frac{\overline{y}_i^{*2}}{\pi_i^2}\pi_i(1-\pi_i)+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\overline{y}_i^*}{\pi_i}\frac{\overline{y}_j^*}{\pi_j}Cov\left[I_i,I_j\right]\\&=\sum_{i=1}^{N}\frac{\overline{y}_i^{*2}}{\pi_i^2}\pi_i(1-\pi_i)+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\overline{y}_i^*}{\pi_i}\frac{\overline{y}_j^*}{\pi_j}(\pi_{ij}-\pi_i\pi_j).\end{aligned}$$

Hence, $V(\hat{\tau}_1)=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\overline{y}_i^*\overline{y}_j^*.$

Since $E\left[I_i\right]=E\left[I_i^2\right]=\pi_i$ and $E\left[I_iI_j\right]=\pi_{ij}$, an unbiased estimator of $V(\hat{\tau}_1)$ is given by

$$\begin{aligned}\hat{V}(\hat{\tau}_1)&=\sum_{i=1}^{N}\frac{\overline{y}_i^{*2}}{\pi_i^2}\pi_i(1-\pi_i)I_i+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\overline{y}_i^*}{\pi_i}\frac{\overline{y}_j^*}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)I_iI_j\\&=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{\overline{y}_i^*\overline{y}_j^*}{\pi_{ij}}I_iI_j.\end{aligned}$$

Hence, $\hat{V}(\hat{\tau}_1)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{\overline{y}_i^*\overline{y}_j^*}{\pi_{ij}}.$

Note that the initial selection probabilities of units are used in the estimator only when these units were selected in the initial sample, although all study values in a network are used to construct the estimate.

The second unbiased estimator will be also derived by modifying the Horvitz-Thompson estimator. The new study value is the total of study values in the networks. Any network size one will not be used in this estimator except when it is selected to be the initial sample. This estimator uses new inclusion probabilities. To obtain the inclusion probabilities of a network to be use in the estimator, it is convenient to deal with networks.

Let $z_k^*=\sum_{j\in\psi_k}z_j$ be the total of initial probabilities in the k^{th} network. Under the notations of networks, the population total can be written as $\tau=\sum_{k=1}^{K}y_k^*=\sum_{k=1}^{K}\sum_{j\in\psi_k}y_j=\sum_{i=1}^{N}y_i$. The network k will be used in the estimator when any unit in its network was selected to be the initial sample. The probability that a network will be drawn into an initial sample is given by Lemma 1.

Lemma 1 *Under the proposed sampling design, the probability that at least one unit in a network* k *is included in an initial sample is given by,*

$$\pi_k^* = \frac{\binom{N-1}{n-1} - \binom{N-m_k-1}{n-1}\left(1-z_k^*\right)}{\binom{N-1}{n-1}}.$$

Proof: The k^{th} network is included in an initial sample when any unit in its network is selected in the initial sample. Let φ_k denote the event that any unit in the network k is selected in the initial sample and φ_k' the event that the initial sample does not contain any unit in the k^{th} network. The event φ_k' occurs when all units in the network are not selected in the first draw and the remaining $(n-1)$ draws. Let $\varphi_{\text{k},(1)}'$ denote the event that all units in the k^{th} network are not selected in the first draw. We found that $P(\varphi_{\text{k},(1)}') = 1 - z_k^*$. Let $\varphi_{\text{k},(\text{n}-1)}'$ be the event that all units in the network k were not selected in the remaining $(n-1)$ draws. We can find the conditional probability,

$$P(\varphi_{\text{k},(n-1)}' \mid \varphi_{\text{k},(1)}') = \frac{\binom{N-m_k-1}{n-1}}{\binom{N-1}{n-1}}.$$

From the definition of π_k^*, we get that

$$\begin{aligned}
\pi_k^* &= P(\varphi_k) = 1 - P(\varphi_k') \\
&= 1 - P(\varphi_{\text{k},(1)}' \cap \varphi_{\text{k},(n-1)}') \\
&= 1 - P(\varphi_{\text{k},(1)}')\, P(\varphi_{\text{k},(n-1)}' \mid \varphi_{\text{k},(1)}') \\
&= 1 - (1-z_k^*)\frac{\binom{N-m_k-1}{n-1}}{\binom{N-1}{n-1}}.
\end{aligned}$$

Therefore, $$\pi_k^* = \frac{\binom{N-1}{n-1} - \binom{N-m_k-1}{n-1}(1-z_k^*)}{\binom{N-1}{n-1}}.$$

Lemma 2 *Under the considered sampling design, the probability that the initial sample contains at least one unit in each of networks k and h is*

$$\pi_{kh}^* = 1 - (1-z_k^*)\frac{\binom{N-m_k-1}{n-1}}{\binom{N-1}{n-1}} - (1-z_h^*)\frac{\binom{N-m_h-1}{n-1}}{\binom{N-1}{n-1}} + (1-z_k^*-z_h^*)\frac{\binom{N-m_k-m_k-1}{n-1}}{\binom{N-1}{n-1}}.$$

Proof: From the definition of π_{kh}^*, it can be shown that

$$\begin{aligned}\pi_{kh}^{*} &= P(\varphi_k \cap \varphi_h) \\ &= 1 - P\left[(\varphi_k \cap \varphi_h)' \right] \\ &= 1 - \left[P(\varphi_k') + P(\varphi_h') - P(\varphi_k' \cap \varphi_h') \right].\end{aligned} \tag{4}$$

Consider the following probabilities. The probabilities of $P(\varphi_k')$ and $P(\varphi_h')$ can be found in Lemma 1. The $P(\varphi_k' \cap \varphi_h')$ in (4) equals to

$$\begin{aligned}P(\varphi_k' \cap \varphi_h') &= P\left[(\varphi_{\mathrm{k},(1)}' \cap \varphi_{\mathrm{h},(1)}') \cap (\varphi_{k,(n-1)}' \cap \varphi_{\mathrm{h},(n-1)}') \right] \\ &= P(\varphi_{\mathrm{k},(1)}' \cap \varphi_{h,(1)}')\, P\left[(\varphi_{\mathrm{k},(n-1)}' \cap \varphi_{\mathrm{h},(n-1)}') \,|\, (\varphi_{\mathrm{k},(1)}' \cap \varphi_{\mathrm{h},(1)}') \right].\end{aligned}$$

We can derive that $P(\varphi_{\mathrm{k},(1)}' \cap \varphi_{\mathrm{h},(1)}') = 1 - z_k^* - z_h^*$. In addition, the conditional probability can be derived by,

$$P\left[(\varphi_{\mathrm{k},(n-1)}' \cap \varphi_{\mathrm{h},(n-1)}') \,|\, (\varphi_{\mathrm{k},(1)}' \cap \varphi_{h,(1)}') \right] = \frac{\binom{N - m_k - m_h - 1}{n-1}}{\binom{N-1}{n-1}}.$$

Substituting these probabilities in (4), we will get the probability that the initial sample contains at least one unit in each of networks k and h.

Theorem 2 *Under the proposed sampling design, let* v *be the number of distinct networks within the initial sample. An unbiased estimator of population total is*

$$\hat{\tau}_2 = \sum_{k=1}^{v} \frac{y_k^*}{\pi_k^*}. \tag{5}$$

The variance of the estimator $\hat{\tau}_2$ *can be written as*

$$V(\hat{\tau}_2) = \sum_{k=1}^{K} \sum_{h=1}^{K} \left(\frac{\pi_{kh}^* - \pi_k^* \pi_h^*}{\pi_k^* \pi_h^*} \right) y_k^* y_h^*, \tag{6}$$

where $\pi_{kk}^* = \pi_k^*$. *An unbiased estimator of the variance of* $\hat{\tau}_2$ *is*

$$\hat{V}(\hat{\tau}_2) = \sum_{k=1}^{v} \sum_{h=1}^{v} \left(\frac{\pi_{kh}^* - \pi_k^* \pi_h^*}{\pi_k^* \pi_h^*} \right) \frac{y_k^* y_h^*}{\pi_{kh}^*}. \tag{7}$$

Proof: We define the new study variable and indicator function for a network. Let y_k^* be a study value of the k^{th} network for $k = 1, 2, \ldots, K$. We knew that $\tau = \sum_{k=1}^{K} y_k^*$.

Let I_k be the indicator function defined as

$$I_k = \begin{cases} 1 \text{ ; some units in network } k \text{ are included in the initial sample} \\ 0 \text{ ; otherwise.} \end{cases}$$

The estimator $\hat{\tau}_2$ can be written as $\hat{\tau}_2 = \sum_{k=1}^{K} \frac{y_k^*}{\pi_k^*} I_k$. The derivations of (5), (6) and (7) can be derived as the proof of Theorem 1.

If the initial probability of the unit i equal to $1/N$ for every unit in the population, the two proposed estimators are reduced to the estimators as in Thompson (1990).

4. A Small Population Example

In this section, a small population is used to demonstrate the computation of the estimates and to show the unbiased property of the estimators. The proposed sampling strategy is also compared with the sampling strategy introduced by Thompson (1990).

In Thompson strategy, the initial sample will be selected by simple random sampling without replacement. The probability of getting an initial sample is given by

$$P(s_0) = \frac{1}{\binom{N}{n}}.$$

The unbiased estimator $\hat{\tau}_1$ of the population total reduces to the modified Hansen-Hurwitz (Thompson 1990),

$$\hat{\tau}_{HH} = \frac{N}{n}\sum_{i=1}^{n} \overline{y}_i^*.$$

In addition, the estimate $\hat{\tau}_2$ becomes to be the modified Horvitz-Thompson estimator (Thompson 1990),

$$\hat{\tau}_{HT} = \sum_{k=1}^{V} \frac{\binom{N}{n}}{\binom{N}{n} - \binom{N-m_k}{n}} y_k^*.$$

Assume that the population consists of five units, the study values, y_i 's; $y_i \in \{50, 100, 0, 5, 10\}$ corresponding to the initial probabilities, $\mathrm{z_i}$'s; $z_i \in \{0.30, 0.40, 0.05, 0.10, 0.15\}$. The neighborhood of each unit includes all adjacent units (of which there are either one or two). The condition is defined by $C = \{y : y \geq 20\}$. The initial sample size is given by $n = 2$. The parameter to be estimated is $\tau = 165$. The probabilities of getting the initial sample and the estimates under the two sampling strategies are represented in Table 1.

The following description is one example that we use to clearly illustrate. For example, the observations 100, 5; 50, 0 means that the initial sample consists of 100 and 5, and the adaptive observations are 50 and 0. From these observations with the proposed strategy, the probability of getting the initial sample $s_0 = \{100, 5\}$ is $P(s_0) = 0.5/4 = 0.125$ and the proposed estimates are computed by

$$\hat{\tau}_1 = \frac{4}{4-(1-0.4)(3)}(75) + \frac{4}{4-(1-0.1)(3)}(5) \;=\; 151.75 \text{ and}$$

$$\hat{\tau}_2 = \frac{4}{4-(1-0.7)(2)}(150) + \frac{4}{4-(1-0.1)(3)}(5) \;=\; 191.86.$$

For Thompson's strategy, the probability of getting the initial sample is $P(s_0) = 1/10 = 0.100$ and the estimates are

$$\hat{\tau}_{HH} = \frac{5}{2}(75+5) \;=\; 200 \text{ and } \hat{\tau}_{HT} = \frac{10}{10-3}(150) + \frac{10}{10-6}(5) \;=\; 226.79.$$

Table 1 All possible final samples, probabilities of getting the initial sample and the estimates under the two sampling strategies

Observations	Proposed strategy			Thompson's strategy		
	$P(s_0)$	$\hat{\tau}_1$	$\hat{\tau}_2$	$P(s_0)$	$\hat{\tau}_{HH}$	$\hat{\tau}_{HT}$
50, 100;0	0.175	294.26	176.47	0.100	375.00	214.29
50, 0;100	0.088	157.89	176.47	0.100	187.50	214.29
50, 5; 100, 0	0.100	173.28	191.86	0.100	200.00	226.79
50, 10;100, 0	0.113	185.48	204.06	0.100	212.50	239.29
100, 0; 50	0.113	136.36	176.47	0.100	187.50	214.29
100, 5; 50, 0	0.125	151.75	191.86	0.100	200.00	226.79
100, 10; 50, 0, 5	0.138	163.95	204.06	0.100	212.50	239.29
0, 5	0.038	15.38	15.38	0.100	12.50	12.50
0, 10	0.050	27.59	27.59	0.100	25.00	25.00
5, 10	0.063	42.97	42.97	0.100	37.50	37.50
Mean		165.00	165.00		165.00	165.00
Variance		5,810.92	3,307.22		11,118.75	8,507.14

We found that the final sample size varies from sample to sample. The two sampling strategies give the unbiased estimators of the population total. In addition, the variances of the estimators under the proposed sampling strategy are less than that of the estimators under Thompson's strategy.

5. Simulation Study

The simulation study is used to compare the efficiency of the proposed sampling strategy (PSS) to the original adaptive cluster sampling (OACS) given by Thompson (1990). Figure 1 consists of a real data, the numbers of ring-necks ducks in a given area (see Smith et al. 1995). The neighborhood of each unit includes four adjacent units. The number of ring-necked ducks in a rectangular will be used as the study variable (y). The population consists of $N = 200$ units. Auxiliary variable (x) correlated to the study variable are created with the 4 setting coefficients of correlation (ρ) : 0.3, 0.5, 0.7 and 0.9. The condition is defined by $C = \{y : y > 0\}$. For PSS, the initial sample is selected by probability proportional to auxiliary variable. For OACS, the initial sample will be drawn with equal probability without replacement.

The simulation consists of 50,000 samples according to the initial sample sizes for each n = 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. The formulas that are used to estimate the expectation and variances of estimators are

$$\tilde{E}(\hat{\tau}^*) = \frac{1}{50,000}\sum_{j=1}^{50,000}\hat{\tau}_j^* \quad \text{and} \quad \tilde{V}(\hat{\tau}^*) = \frac{1}{50,000-1}\sum_{j=1}^{50,000}\left[\hat{\tau}_j^* - \tilde{E}(\hat{\tau})\right]^2, \quad \text{respectively,}$$

where the $\hat{\tau}_j^*$ is the value of the estimator for the sample j for each sampling strategy, and the $\tilde{E}(\hat{\tau}^*)$ is the average of the estimates for the estimator for each sampling strategy. The estimate of relative bias is defined as $RB = \left[\tilde{E}(\hat{\tau}^*) - \tau\right]/\tau.$ The estimate of standard error is defined by the squared root of the variance.

0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	675	0	0	0
0	0	0	0	100	100	75	0	0	0	0	0	0	0	0	0	0	0	0	0
0	4,000	13,500	0	0	154	120	200	0	0	0	0	0	0	0	0	0	0	55	0
0	0	0	0	0	80	585	430	0	4	0	0	0	0	0	35	0	0	0	0
0	0	0	0	0	0	0	0	40	0	0	0	0	2	0	0	0	1,615	0	0
0	0	0	0	0	0	0	0	57	0	0	0	0	2	0	0	0	200	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1,141	13
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	107	22

Figure 1 The numbers of ring-necks ducks in a given area

Table 2 The averages of estimates under two sampling strategies

n	OACS		PSS							
			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	$\hat{\tau}_{HH}$	$\hat{\tau}_{HT}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$
5	22,707.2	22,757.6	23,103.7	23,149.4	23,208.4	23,285.0	23,032.7	23,084.1	23,145.4	23,232.9
10	23,379.9	23,370.6	23,431.4	23,389.9	23,474.5	23,418.9	23,455.1	23,351.8	23,315.6	23,202.1
15	23,465.8	23,529.6	23,445.6	23,480.8	23,447.1	23,473.9	23,528.0	23,557.7	23,642.1	23,676.1
20	23,439.1	23,441.6	23,459.4	23,453.9	23,434.2	23,436.9	23,440.9	23,435.3	23,462.2	23,478.4
25	23,543.3	23,430.2	23,460.2	23,359.6	23,486.5	23,395.5	23,471.6	23,391.7	23,467.9	23,418.2
30	23,463.8	23,476.2	23,475.7	23,471.4	23,456.6	23,454.5	23,438.4	23,439.7	23,424.1	23,408.6
35	23,258.4	23,269.8	23,265.2	23,291.9	23,221.0	23,258.7	23,203.5	23,245.8	23,178.8	23,214.8
40	23,294.2	23,251.1	23,299.6	23,269.8	23,289.3	23,264.9	23,275.9	23,241.8	23,303.1	23,273.2
45	23,363.8	23,287.1	23,349.1	23,302.0	23,335.7	23,292.0	23,329.8	23,278.1	23,344.4	23,295.0
50	23,325.7	23,318.1	23,366.4	23,342.2	23,350.3	23,322.8	23,354.2	23,328.2	23,338.7	23,304.9

In Table 2, the averages of estimates of all estimators are very close to the population total $(\tau = 23,333)$. The estimates of relative bias of all estimators in Table 3 are also close to zero. These results confirm that these estimators are unbiased estimators of the population total.

Table 4 shows that the proposed sampling design outperforms the original adaptive cluster sampling design. For a given initial sample size, the standard error of proposed estimator $(\hat{\tau}_1)$ is small than that of the Modified Hansen-Hurwitz estimator $(\hat{\tau}_{HH})$ of Thompson (1990) and the standard error of the estimator $\hat{\tau}_2$ is also less than that of the estimator $\hat{\tau}_{HT}$.

Table 3 The estimates of relative bias of estimators under two sampling strategies

n	OACS		PSS							
			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	$\hat{\tau}_{HH}$	$\hat{\tau}_{HT}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$
5	-0.027	-0.025	-0.010	-0.008	-0.005	-0.002	-0.013	-0.011	-0.008	-0.004
10	0.002	0.002	0.004	0.002	0.006	0.004	0.005	0.001	-0.001	-0.006
15	0.006	0.008	0.005	0.006	0.005	0.006	0.008	0.010	0.013	0.015
20	0.005	0.005	0.005	0.005	0.004	0.004	0.005	0.004	0.006	0.006
25	0.009	0.004	0.005	0.001	0.007	0.003	0.006	0.003	0.006	0.004
30	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.004	0.003
35	-0.003	-0.003	-0.003	-0.002	-0.005	-0.003	-0.006	-0.004	-0.007	-0.005
40	-0.002	-0.004	-0.001	-0.003	-0.002	-0.003	-0.002	-0.004	-0.001	-0.003
45	0.001	-0.002	0.001	-0.001	0.000	-0.002	0.000	-0.002	0.000	-0.002
50	0.000	-0.001	0.001	0.000	0.001	0.000	0.001	0.000	0.000	-0.001

Table 4 The estimates of standard errors under two sampling strategies

n	OACS		PSS							
			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	$\hat{\tau}_{HH}$	$\hat{\tau}_{HT}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$
5	75,974.6	75,925.3	73,088.6	71,977.6	69,046.5	67,317.0	63,156.3	60,425.6	54,225.4	50,321.2
10	54,150.1	53,471.7	52,691.1	51,718.0	51,006.2	49,762.8	48,418.2	46,706.3	43,527.5	41,154.0
15	43,574.5	43,013.0	42,750.5	41,986.7	41,719.8	40,828.3	40,180.4	39,117.8	37,185.4	35,689.7
20	37,306.7	36,375.7	36,741.5	35,741.4	36,005.2	34,975.4	34,937.6	33,775.5	32,697.1	31,301.8
25	33,230.1	31,847.1	32,732.5	31,362.8	32,227.4	30,842.1	31,382.3	29,957.7	29,635.3	28,102.4
30	29,587.2	28,445.1	29,304.6	28,104.0	28,899.0	27,678.0	28,275.7	26,994.5	26,971.7	25,514.7
35	26,896.5	25,636.4	26,620.2	25,376.6	26,287.6	25,036.6	25,789.8	24,497.7	24,736.7	23,314.5
40	24,876.4	23,412.9	24,657.9	23,205.0	24,391.3	22,932.8	24,001.2	22,486.6	23,113.6	21,512.8
45	23,175.4	21,553.2	22,961.1	21,366.9	22,745.4	21,141.8	22,419.1	20,777.4	21,665.5	19,949.7
50	21,494.6	19,907.9	21,387.9	19,756.0	21,219.1	19,571.9	20,928.1	19,258.9	20,303.4	18,559.8

When the initial sample size is fixed, the standard error of modified Horvitz-Thomson estimator $(\hat{\tau}_{HT})$ is smaller than that of Hansen-Hurwitz estimator $(\hat{\tau}_{HH})$. For given ρ and n, the standard error of $\hat{\tau}_2$ is also smaller than that of $\hat{\tau}_1$. The results of Thompson (1990) correspond with these results. With fixed the initial sample size n, the standard errors of the proposed estimators decrease when the coefficients of correlation increase.

6. Conclusions

This paper presented an adaptive cluster sampling with unequal probability sample without replacement. In the proposed sampling scheme, the initial sample is easy to drawn since only the first unit is selected with unequal probabilities but the others are drawn with equal probabilities. First unbiased estimate is created from the initial probabilities of all units in the initial sample and the second one is derived from the initial probabilities of all networks that are intersected of the initial sample. In addition, the both proposed estimates are easy to compute. The simulation study showed that the proposed sampling strategy was more efficient than the original adaptive cluster sampling strategy. When the correlation coefficient between the auxiliary and the study variables increases, the estimate standard errors of the proposed estimators decrease. However, for the proposed sampling design, the number of distinct network (v) and the final sample size (n_1) are random variables. It can be seen that the proposed sampling design is suitable for sampling rare and clustered populations. In addition, when there are high correlation between the initial probability and the study variable, this sampling design will have the high efficiency.

Acknowledgements

This work was supported by Silpakorn University Research and Development Institute. The authors are grateful for these supports. We also thank the anonymous reviewers and editor for their helpful comments that led to an improved paper.

References

Midzuno H. On the sampling system with probability proportionate to sum of the sizes. Ann. I Stat Math. 1952; 3: 99-107.

Roesch FA. Adaptive cluster sampling for forest inventories. Forest Sci. 1993; 39(4): 655-669.

Sampath S. Sampling theory and methods. Alpha Science International. Harrow, U.K; 2005.

Sangngam P. Unequal probability inverse adaptive cluster sampling. Chiang Mai J Sci. 2013; 40(4): 736-742.

Smith DR, Conroy MJ, Brakhage DH. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics.1995; 51(2): 777-788.

Smith DR, Villella RF, Lemarie´ DP. Application of adaptive cluster sampling to low-density populations of freshwater mussels. Environ Ecol Stat. 2003; 10(1): 7-15.

Thompson SK. Adaptive cluster sampling. J Am Stat Assoc. 1990; 85(412): 1050-1059.

Thompson SK. Stratified adaptive cluster sampling. Biometrika. 1991a; 78(2): 389-397.

Thompson SK. Adaptive cluster sampling: designs with primary and secondary units. Biometrics.1991b; 47(3): 1103-1115.