Regression Estimator for Adaptive Cluster Sample

Chang-Kyoon Son

Abstract Background/Objectives: Adaptive cluster sampling (ACS) is known as a sampling design for rare and clustered objects. We suggest the regression estimator to improve the efficiency of ACS estimator.

Methods/Statistical analysis: We estimate the population total of Pedicularisishidoyana Koidz. &Ohwi in the Gyeongju National Park by using the regression estimation for adaptive cluster sampling. We can consider an auxiliary variable which has strong correlation in the estimation procedure. To do this we simulate auxiliary variable has sample correlation r=0.86. The efficiency of the proposed estimator is evaluated by comparing the relative efficiency and 95% confidence limit of estimator with the typical adaptive cluster estimator.

Findings: In this study, we found that the regression estimator

I. INTRODUCTION

Adaptive cluster sampling design is known as a sampling method for rare clustered population. This method is often useful after locating a unit that meets a specified criterion to continue sampling in that region. In spatial sampling, adaptive cluster sampling can be provided unbiased efficiency for estimating the abundance of rare, clustered populations[1]. For sampling hidden human populations such as drug abuse, an illegal private loan or a prevalence of rare diseases, social links play the same role as geographic proximity in spatial sampling and adaptive cluster sampling becomes a type of link-tracing design in a graph or social network [2,3].The adaptive sampling performs when the variable of interest for a unit in the sample satisfies a pre-specified condition, neighboring or connected units are added to the sample and observed. This procedure continues until no more units are found that meet the criterion.

The adaptive sampling has an advantage that it is more flexible than conventional survey sampling, because the sampling design depends on the variable of interest neighboring or connected units are added to the sample and observed in the survey field[6,7,9]. Also, it has an advantage of obtaining enough observations compare to a conventional sampling method. Recently, it has been suggested a new estimator using stratified adaptive cluster sampling [4]. They have shown that the ratio estimator for adaptive cluster sampling is more efficient than the typical adaptive sampling.

In this study, we propose the regression estimation method using an auxiliary variable related to the variables of interest so that we investigate rare domestic indigenous plants, *Pedicularisishidoyana Koidz. & Ohwi*, which is native plants to Gyeongju National Park in Korea, and we consider the growth environment asthe auxiliary variable such as the humidity, temperature and location of a rare native plant. Finally we compared the efficiency of the proposed estimator with the existing estimator by adaptive cluster sampling.

II. MATERIALS AND METHOD

2.1. Review of the adaptive cluster sampling

Adaptive sampling method which the initial sampling unit by a given sampling design, then observes the sampled unit, and if the specific condition is satisfied, the method of additionally observing the unit and the neighboring object is included in the sample. If an initial sample unit is drawn by a given probability design and the sampling unit satisfies a given condition, then adaptive cluster sampling is a method of additionally including neighboring units of the object in a sample. For example, rare minerals and oil exploration do not show signs of minerals or oil in many areas, but if you discover the mineral in a particular area, or if you have found oil, it is natural to try to find or dig out further. Also, it is similar to investigate a disease in a specific area. It is a further investigation of the person around the infected person or the individual who has a certain relationship. In this respect, adaptive cluster sampling is a particularly suitable method for examining populations that are rare and forming communities in a particular area, and if the same sample size is used, this method is more accurate than conventional sampling methods [1,2,5,6].

In adaptive cluster sampling, the neighborhood A_i is the set of observations to be added to the sample when the observation value i satisfies a certain condition. The neighbors of the observed values are defined as the upper, lower, left, and right sides of the initial observation value as Figure 1[1,2,3].

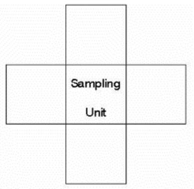


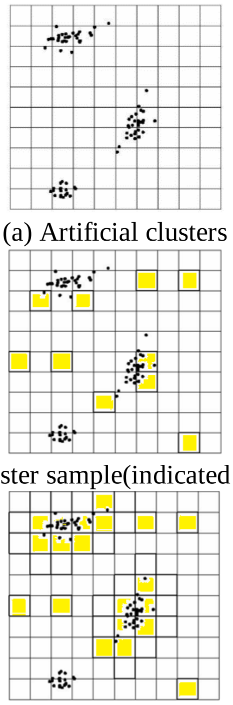
Figure 1. Neighborhood Units

On the other hand, if the interest condition is a set $C$of values in the range of the interest variable, then if

Revised Manuscript Received on May 22, 2019.

[1]**Chang-Kyoon Son,** Department of Applied Statistics, Dongguk University,38066, DongdaeRo123,Gyeongju, Gyeongbuk, Korea
ckson85@dongguk.ac.kr

y is in C, the observation value i is said to meet the condition and all observations in that neighborhood are included in the sample. In general, C is a set of values greater than or equal to some constant. If $C = \{y : y \geq 1\}$, that is, if the observation contains at least one unit, the observation satisfies the condition[11,12]. A cluster is a set of all observations contained in the sample when the observation value i is initially selected. The cluster consists of only the observation value i if the observation value i does not satisfy the condition. When the observation value i satisfies the condition, the cluster includes not only the neighbor of the observation value i and the observation value j but also the neighboring neighbor of the observation value i satisfying the condition as Figure 2. A network is a subset of a cluster, consisting of all the units of a cluster that satisfy a condition. The network is characterized in that all units of the network are included when the unit of the network is selected in the initial sample. An edge unit is a unit in the neighborhood of an observation that does not meet the condition but meets the condition. Observation values can be one or more cluster or network edge units[1,2,3,5].



(a) Artificial clusters

(b) Initial cluster sample(indicated yellow cells)

(c) Adaptive cluster sample

Figure 2. Adaptive cluster sample

The adaptive cluster sampling is designed in such a way that it is influenced by the initial set design of the unit selected by the probability sampling and when the random variable of the selected unit satisfies the given condition, it is added among the neighbors of the unit [1,2,3].

The basic assumptions for deriving the adaptive cluster estimator are following;

i)Neighboring units are predetermined before sampling and are not determined at the site.

ii)If the observation value y the selected unit in the sample satisfies a specific condition for adaptive sampling,$y > C$,the neighboring unit of the unit is included in the sample.

iii)If other units satisfy the condition, they also include their neighbor unit in the sample.

Until there are units that do not satisfy this condition C(border of the edge unit), continues to sample, ultimately identifying the units included in the sample and defining the initial sample as.

2.2. Estimator for an adaptive cluster sampling

For the adaptive cluster estimation, let N be the number of sampling units in the population, let y_i be the value of the ith observation unit, A_i be the network of the ith sampling unit, m_i be the number of sampling units in A_i, a_i be the total number of sampling units in the network that are edge units, and n_1 is the number of units finally included in the sample. Suppose that an initial sample $n_1$unit is taken with SRSWOR, then the inclusion probability for the ith unit is

$$\pi_i = 1 - \binom{N - m_i - a_i}{n_1} / \binom{N}{n_1}, \tag{2.1}$$

where$m_i$is the number of units in the network A_i which known for the sampled unit, but a_i that is unknown.

If we knew the inclusion probability $\pi_i$for all sample units, we defined the Horvitz-Thompson(HT) estimator for the population total as following,

$$\hat{\tau}_{HT} = \sum_{i=1}^{\upsilon} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{y_i I_i}{\pi_i}, \tag{2.2}$$

where$y_1, y_2, \cdots, y_N$ be the y-values of the υ distinct units in the final sample, and I_i takes the value 1 if unit i is included in the sample and 0 otherwise.

An alternative, the partial inclusion probability of ith unit does not satisfy a given condition is defined by

$$\pi_i^{'} = 1 - \binom{N - m_i}{n_1} / \binom{N}{n_1}. \tag{2.3}$$

The unbiased estimator of population total is

$$\hat{\tau} = \sum_{i=1}^{N} \frac{y_i I_i^{'}}{\pi_i^{'}}, \tag{2.4}$$

where $I_i^{'}$ is indicator variable with 1, if the initial sample intersect to the networkA_i, 0 otherwise.

In order to expression network unit instead of sampling unit, the total number of the network is L, A_k is set of including unit in the kth network, and $x_k$is the number of sampling unit in the k th network, then the inclusion probability of sampling unit in the kth network is

$$\alpha_k = 1 - \binom{N - x_k}{n_1} / \binom{N}{n_1}. \tag{2.5}$$

The unbiased estimator for population total $\tau$is

$$\hat{\tau}_{HT} = \sum_{k=1}^{K} \frac{y_k J_k}{\alpha_k} = \sum_{k=1}^{N} \frac{y_k^*}{\alpha_k}, \tag{2.6}$$

wherey_k^* is the total of units y_k in the kth network.

The variance estimator of the sample total and its variance are given by respectively.

$$V(\hat{\tau}_{HT}) = \sum_{j=1}^{N} \sum_{k=1}^{N} \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_{jk}}\right) y_j^* y_k^*, \tag{2.7}$$

$$\hat{V}(\hat{\tau}_{HT}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_{jk}}\right) \frac{y_j^*}{\alpha_j} \frac{y_k^*}{\alpha_k}, \tag{2.8}$$

where the joint intersection probability of network k and$j$,α_{jk} is

$$\alpha_{jk} = 1 - \left[\binom{N - x_j}{n_1} + \binom{N - x_k}{n_1} - \binom{N - x_j - x_k}{n_1}\right] / \binom{N}{n_1} \tag{2.9}$$

III. RESULTS AND DISCUSSION

3.1. Regression Estimator for the cluster sampling

The regression estimator is an estimate that is often considered along with the ratio estimator when there is auxiliary information on the population level. Unlike the ratio estimator, the regression estimator can be defined when there are multiple auxiliary variables, and the efficiency of the regression estimator is determined by the linear relationship between the auxiliary variables and the variables of interest. The regression estimator defined under the cluster sampling design is often considered as an efficient estimator than the unbiased HT estimator[8,10].From these various model assumptions, the generalized regression estimator can be derived as follows. The population elements $U = \{1, 2, \cdots, N\}$is divided by the N_I cluster as $U_1, U_2, \cdots U_{N_I}$.

$$U = \bigcup_{i=1}^{N_I} U_i \ ; \ N = \sum_{i=1}^{N_I} N_i$$

In addition, suppose that the supplementary information of cluster unit is available as PSU level auxiliary variable for the primary sampling unit.

$$u_i = (u_{1i}, u_{2i}, \cdots, u_{Ji})'.$$

Also, we defined the 1st and 2nd inclusion probabilities of the cluster under the sampling design $p_I(\cdot)$as

$$\pi_{Ii} = \textstyle\sum_{s_I \ni i} p(s_i)\,, \pi_{Iij} = \sum_{s_I \ni i\&j} p_I(s_i). \qquad (3.1)$$

For the level of an auxiliary information, we assume that auxiliary variable u_i can be observed for the sampled cluster $i \in s$ and $\sum_U u_i$is true value and it is available in the estimation procedure. Under these assumptions for the PSU and population level of auxiliary information, we consider Wher the population model for (y_i, u_i) of the numbers of N_I.

$$\begin{cases} E_\xi(y_i) = u_i{}'\beta_I \\ \\ V_\xi(y_i) = \sigma_{Ii}^2 \end{cases} \qquad (3.2)$$

where y_i is independent under this model assumption and if the N_I number in the population are available, then the population regression coefficient B_I is given

$$B_I = (\textstyle\sum_U u_i u_i{}'/\sigma_{Ii}^2)^{-1}(\sum_U u_i{}' y_i \ /\sigma_{Ii}^2)\,. \qquad (3.3)$$

Under the model (3.2), the regression coefficient estimator based on the sample is

$$\hat{B}_I = (\textstyle\sum_{s_I} u_i u_i{}'/\sigma_{Ii}^2\pi_{Ii})^{-1}(\sum_{s_I} u_i y_i^* \ /\sigma_{Ii}^2\pi_{Ii})\,. \qquad (3.4)$$

where $y_i^* = \sum_{s_I} y_k / \pi_{k|i}$.

Then the regression estimator for the population total is

$$\hat{\tau}_{yR} = \textstyle\sum_{s_I} \frac{y_i^*}{\pi_i} + \left(\sum_U u_i - \sum_{s_I} \frac{u_i}{\pi_i}\right)' \hat{B}_I, \qquad (3.5)$$

where $\hat{B}$is given in (3.4).

Also, the variance estimator for the regression estimator (3.5) is

$$\mathrm{V}(\hat{\tau}_{yR}) = \textstyle\sum_{i=1}^N \sum_{j=1}^N (\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \frac{D_{Ii}}{\pi_{Ii}} \frac{D_{Ij}}{\pi_{Ij}}, \qquad (3.6)$$

where $D_{Ii} = y_i - u_i{}'B_I$is the population difference, and its estimator is

$$\hat{\mathrm{V}}(\hat{\tau}_{yR}) = \textstyle\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}\right) \frac{g_{Ii} d_{Ii}}{\pi_{Ii}} \frac{g_{Ij} d_{Ij}}{\pi_{Ij}}, \qquad (3.7)$$

where $d_{Ii} = y_i^* - u_i{}'\hat{B}_I$is the sample difference and g_{Ii} is g-weight for th PSU which given by

$$g_{Ii} = 1 + \left(\textstyle\sum_U u_i - \sum_{s_I} \frac{u_i}{\pi_{Ii}}\right)' (\sum_{s_I} u_i u_i{}'/\sigma_{Ii}^2\pi_{Ii})^{-1} u_i/\sigma_{Ii}^2\,. \quad (3.8)$$

3.2. Regression Estimator for the adaptive cluster sampling

In order to derive the regression estimator of the adaptive cluster sample, we assume that the population distribution isthe same as model (3.2),

$$E_\xi(y_i) = u_i{}'\beta,\ V_\xi(y_i) = \sigma_i^2.$$

From (2.4) initial sample size is $n_1$and the intersection probability of ith network α_i with the number of sampling unit in th network is given by

$$\alpha_i = 1 - \binom{N - x_i}{n_1} / \binom{N}{n_1},$$

and the joint probability of ith and jth network is

$$\alpha_{ij} = 1 - \left[\binom{N - x_i}{n_1} + \binom{N - x_j}{n_1} - \binom{N - x_i - x_j}{n_1}\right] / \binom{N}{n_1}.$$

Then the proposed regression estimator of the population total which replaced the inclusion probability $\pi_{Ii}$with the intersection probability $\alpha_i$for adaptive cluster sampling is

$$\hat{\tau}_{yAR} = \textstyle\sum_s \frac{y_i^*}{\alpha_i} + \left(\sum_U u_i - \sum_s \frac{u_i}{\alpha_i}\right)' \hat{B}, \qquad (3.9)$$

where $\hat{B} = (\sum_s u_i u_i{}'/\sigma_i^2\alpha_i)^{-1}(\sum_s u_i y_i^* \ /\sigma_i^2\alpha_i)$.

Also, we can derive the variance estimator for is

$$\mathrm{V}(\hat{\tau}_{yAR}) = \textstyle\sum_{i=1}^N \sum_{j=1}^N (\alpha_{ij} - \alpha_i\alpha_j) \frac{D_i}{\alpha_i} \frac{D_j}{\alpha_j}, \qquad (3.10)$$

and its estimator is

$$\hat{\mathrm{V}}(\hat{\tau}_{yAR}) = \textstyle\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\alpha_{ij} - \alpha_i\alpha_j}{\alpha_{ij}}\right) \frac{g_i d_i}{\alpha_i} \frac{g_j d_j}{\alpha_j}, \qquad (3.11)$$

where $d_i = y_i^* - u_i{}'\hat{B}$ and

$$g_i = 1 + \left(\textstyle\sum_U u_i - \sum_s \frac{u_i}{\alpha_i}\right)' (\sum_s u_i u_i{}'/\sigma_i^2\,\alpha_i)^{-1} u_i/\sigma_i^2.$$

3.3. Typical estimation for an adaptive cluster sampling

In order to apply the adaptive cluster sampling, first of all, the rare plants which are naturally grown in Gyeongju National Park were consulted by landscape experts and we decided to survey for *Pedicularisishidoyana Koidz. & Ohwi*. According to the expert device, it appears that the plants grow mainly valleys with 107~494 meter above sea level and 5~20 degree of inclination. Figure3 shows (a) the picture of *Pedicularisishidoyana Koidz. & Ohwi*, and (b) and (c) are the survey plots in Gyeongju National Park.

(a) Pedicularisishidoyana Koidz. & Ohwi

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

(b) First survey plot

(c) Second survey plot

Figure 3. Picture for *Pedicularisishidoyana Koidz. & Ohwi* in Gyeongju National Park

The grid of the survey area was originally set to 2 × 2 km, but it has been redesigned to 200 × 200(m) for this study. As shown in Figure 4, the distribution of *Pedicularisishidoyana Koidz. & Ohwi* have shown in two survey plots.

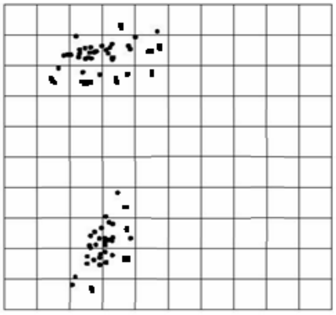
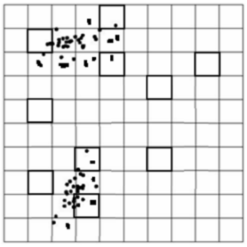
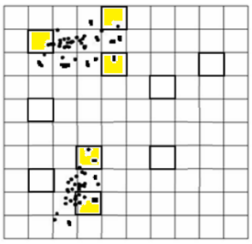


Figure 4. Distribution of Survey Plot #39th in Gyeongju National Park

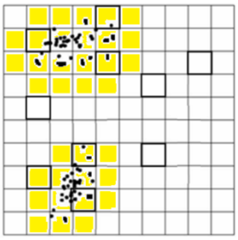
Figure. 5 shows an initial sample of $n_1 = 10$by simple random sampling from $N = 100$ and then an adaptive cluster sampling step that sequentially enumerates the neighboring individuals with the surveyed individuals from the selected initial population sample.



(a) Initial samples



(b) Sample clusters



(c) Final adaptive cluster sample(where yellow color means the network)

Figure 5. Adaptive cluster sampling of Survey Plot #39th in Gyeongju National Park

From Figure 5, in order to estimate for adaptive cluster sample we have been observed the initial sample $n_1 = 10$,and $m_1 = 21$, $y_1^* = 43$ in first network, $m_2 = 14$, $y_2^* = 38$ in second network, and otherwise $m_i = 0$, $y_1^* = 0$. Also, we can calculate the initial intersection probabilities are, for the first network

$$\alpha_1 = 1 - \left[\frac{\binom{100-21}{10}}{\binom{100}{10}}\right] = 0.917,$$

and for the second network

$$\alpha_2 = 1 - \left[\frac{\binom{100-14}{10}}{\binom{100}{10}}\right] = 0.796,$$

and for $i = 3, 4, \cdots, 11$

$$\alpha_i = 1 - \left[\frac{\binom{100-1}{10}}{\binom{100}{10}}\right] = 0.1.$$

In addittion to, we can define the joint intersection probabilities as $\alpha_{12} = \alpha_1 + \alpha_2 - (1 - p_{12}) = 0.723$, with $p_{12} = 0.01034$. Using thses values, we can estimate the population total of *Pedicularisishidoyana Koidz. & Ohwi* in Gyeongju National Park as

$$\hat{\tau}_y = \left(\frac{43}{0.917} + \frac{38}{0.796} + 0 + \cdots + 0\right) = 94.6,$$

$$\hat{V}(\hat{\tau}_y) = \left[\frac{43^2}{0.917}\left(\frac{1}{0.917} - 1\right) + \frac{38^2}{0.796}\left(\frac{1}{0.796} - 1\right) + \frac{2(43)(38)}{0.723}\left(\frac{0.723}{(0.917)(0.796)} - 1\right)\right] = 604.5$$

.

Finally we can obtain the 95% confidence interval for the populatopn total from adaptive cluster sampling to the *Pedicularisishidoyana Koidz. & Ohwi* in Gyeongju National Park is 94.6 ± 49.2.

3.4. Regression estimation for an adaptive cluster sampling

In this subsection, we consider the regression estimator for an adaptive cluster sampling using an auxiliary information which has strong correlation with the variable of interest. To do this, we will use data which observed in the subsection 4.1 as the cluster population is composed of size $N = 100$ and the initial sample is selected with size of $n_1 = 10$, and then we have been observed the survey object using the adaptive cluster sampling. In this step we generate that an auxiliary variable which are correlated with the variable of interest, ($r = 0.86$), are observed in the survey plot as shown in Table 1.Assume that the population total of auxiliary variable in this study area is known as $\tau_x = 250$. Table 2 shows the proposed regression estimation is more efficient than the typical estimation.

Table 1:Observed values of the variable of interest and an auxiliary variable for two networks

Networks	Variables	1	2	3	4	5	6	7	8	9	10	11	Total
1st Network	y_i	1	3	1	2	7	4	1	12	6	3	3	43
	u_i	10	10	10	10	16	15	10	18	20	10	11	140
2nd Network	y_i	4	11	9	7	4	3	-	-	-	-	-	38
	u_i	10	20	20	15	15	10	-	-	-	-	-	90

Table 2: Comparison between the original adaptive estimate and the proposed estimate

	Population Total Estimate	Variance Estimate	95%lowerlimit	95%upperlimit	RE
Original estimate($\hat{\tau}_{HT}$)	94.6	604.5	-3.74	192.95	6.86
Regression Estimate($\hat{\tau}_{yAR}$)	89.2	88.1	51.56	126.74	

IV. CONCLUSION

This paper proposed the regression estimator of an adaptive cluster sampling for the investigation of rare and clustered objects. The advantage of the regression estimator is that it can improve the degree of the estimator by using the available ancillary information in the estimation process. From this point of view, this paper applied the adaptive cluster sampling method to estimate of the population total for the domestic native plant, *Pedicularisishidoyana Koidz. & Ohwi*, and compared the efficiency by applying the regression estimator to improve the efficiency of the estimator. In section 4, we find the proposed estimator is more efficient than the original adaptive estimator in the relative efficiency (:RE) of about 6.86 times. In addition, for the 95% confidence interval of the estimates, it is logically appropriated intervals were estimated for the proposed estimation method.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B03034902).

REFERENCES

1. Thompson SK, Seber GAF. Adaptive Sampling. John Wiley & Sons, Inc., New York ; 1996, p.265.
2. Thompson SK. Sampling. John Wiley & Sons, Inc., New York; 1992, p. 339.
3. Thompson, SK. Adaptive cluster sampling. Journal of the American Statistical Association.1990 Dec;85(412): 1050-59.
4. Dyver L, Thompson SK. Improved unbiased estimators in Adaptive cluster sampling, Journal of Royal Statistical Society, Series B. 2005;67(1):157-66.
5. Philip T, John JB. A review of adaptive cluster sampling. Environmental and Ecological Statistics. 2005; 12:55-94.
6. Philppi T. Adaptive cluster sampling for estimation of abundance within local populations of low-abundance plants. Ecology. 2005; 86(5); 1091-1100.
7. GoldBerg NA, Heine JN, Brown JA. The application of adaptive cluster sampling for rare subtidal macroalgae. Mar Biol. 151, 1343-48. DIO:10.1007/s00227-006-0571-2.
8. Cochran WG. Sampling Techniques, 3rd ed. New York: Wiley; 1977.
9. Felix Medina MH. Analytical expressions for Rao-Blackwell estimators in adaptive cluster sampling. Journal of Statistical Planning and inference. 2000 Mar; 84(1):221-36.
10. Sarndal, CE, Swensson, B, Wretman, J. Model Assisted Survey Sampling.Springer, New York;1992, p.265.
11. Smith DR, Conroy MJ,Brakhage DH. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl.Biometrics.1995 Jun;51(2):777-88.
12. Brown JA. Designing an efficient adaptive cluster sample. Environmental and Ecological Statistics.2003; 10: 95–105.