ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 39 (2019), 119-130

Logarithmic Product Type HT Estimator for Adaptive Cluster Sampling With Negatively Correlated Auxiliary Variable

Raosaheb Latpate*, J. K. Kshirsagar and Vijay Narkhede
[Received on November, 2018. Accepted on December, 2019]

ABSTRACT

Adaptive cluster sampling (ACS) was specially designed for the rare and clustered populations by Thompson (1990). It has been proved that the efficiency of the estimators used in ACS increases if the auxiliary information is used in estimation. Different estimators are so for designed to estimate the population parameters of the rare and clustered population in the presence of auxiliary information. This paper proposes a logarithmic product type Horvitz- Thompson (HT) estimator of the population total of the interest variable under ACS when the interest and the auxiliary variables are negatively correlated. The exact expression for the bias and mean square error has obtained by using second degree of approximation. A simulation study was carried out to show the performance of the proposed estimator. Also the performance of the proposed estimator is compared with Murthy's product HT estimator.

1. Introduction

Adaptive cluster sampling was specially designed for the rare and clustered populations by Thompson (1990). It has shown to be efficient as compared to the traditional sampling designs for such type of populations. Several authors have developed ACS designs under different settings. Thompson (1991b) used stratified sampling to select the initial sample in ACS. Salehi and Seber (1997) introduced two-stage ACS. Dryver et al. (2012) used systematic sampling to select the initial sample in ACS.

*Corresponding Author**: Raosaheb Latpate, Department of Statistics and Centre for Advance Studies, Savitribai phule Pune University, Pune, Email: rvl.@unipune.ac.in, J. K. Kshirsagar. Department of Statistics, NAC&S College, Ahmednagar, Vijay Narkhede, Department of Higher Education, Maharashtra State, Pune.

In ACS, the final sample size is variable. Considering it as a random variable Latpate et al. (2018b) have derived expression for the expected final sample size under ACS. Latpate and Kshirsagar (2020) have proposed two-stage inverse adaptive cluster sampling which is a combination of two-stage inverse sampling and ACS. Under ACS, usually the estimators use the information on the interest variable only. It has proved that the use of auxiliary information at the estimation stage improves the efficiency of the estimators of the population parameters when the interest variable and the auxiliary variable are highly correlated. The estimators that use the auxiliary information include the ratio, regression and product estimators. Felix and Thompson (2004) presented a multiphase variant of ACS. They called it as the adaptive cluster double sampling (ACDS). They proposed regression estimator of the population mean of the interest variable. On the basis of Monte- Carlo simulation study they showed that the regression estimator performs better than the HT estimator under ACDS.

Ratio estimators are suitable when the correlation between the auxiliary and the interest variable is highly positive. Chao (2004) for the first time analyzed the behavior of ACS when the auxiliary and the interest variable are positively correlated. He proposed a generalized ratio estimator based on the modified HT estimator under ACS. Another ratio estimator based on the modified Hansen-Hurvitz (HH) estimator under ACS was proposed by Dryver and Chao (2007). Chao et al. (2011) and Lin and Chao(2014) improved the ratio estimators under ACS using Rao-Blackwell theorem. Bahl and Tuteja (1991) proposed product type ratio exponential estimators.

Latpate and Kshirsagar (2019) developed new design for the rare and clustered population with respect to the interest variable where the correlation between the auxiliary and the interest variable is negative. They named it as negative adaptive cluster sampling (NACS). They have shown that this design is cost effective as compared to ACS. Latpate and Kshirsagar (2020) have also introduced another design, negative adaptive cluster double sampling (NACDS). In this design they have considered high negative correlation between the auxiliary and the interest variable. Latpate and Kshirsagar (2019) have introduced a variant of NACS called two-stage NACS. Under this design they have proposed composite HT estimator and two-stage regression estimator of the population total of the interest variable.

Murthy (1964) developed the product estimator for the situation where the correlation between the auxiliary and the interest variable is high but negative. Gattone et al. (2016) proposed two product estimators based on HT and HH estimators where the correlation between the auxiliary and the interest variable is high but negative.

In this paper, we consider the case when the correlation between the auxiliary and the interest variable is high but negative. We propose a logarithmic product estimator based on HT estimator of the population total of the interest variable. Also we compare the proposed estimator with Murthy's product type HT estimator.

In Section 2, the classical ACS design with its corresponding estimators is presented. Section 3 develops the new logarithmic product type HT estimator under ACS for the negatively correlated variables. A simulation study that shows the performance of the proposed estimator is presented in Section 4. Discussion on results and conclusions is given in Section 5.

2. Adaptive Cluster Sampling

Consider a rare and clustered finite population of N units. Let Y be the variable of interest that takes values $Y_1, Y_2, \ldots, Y_N$ corresponding to the N population units respectively. The target parameter is $\tau_Y = \sum_{i=1}^{N} Y_i$, the population total of the interest variable Y. In ACS, an initial sample of n units is selected from the population of size N by using a conventional sampling design. Let C be the pre-specified condition of interest. If any of the initially selected units satisfies C, neighboring units are added to the sample. Neighboring units of an unit are those units that have at least one common boundary. (Refer Thompson 1990). If any of the units so added to the sample satisfies C then its neighbors are also added to the sample. This process of adding the neighboring units continues till there are no units in the neighborhood satisfying C. The set of units satisfying C around a unit included in the initial sample forms a network. A unit included in the initial sample that does not satisfy the condition C is treated as a network of size one. The units that are adaptively added to the sample but do not satisfy the condition C are called as the edge units of the corresponding network. The set of units included in a network along with its edge units is called as a cluster. We assume that the population is partitioned into K distinct networks.

The modified HT estimator of τ_Y is given as:

$$(\hat{\tau}_Y)_{HT} = \sum_{k=1}^{K} \frac{Z_k y_k^*}{\pi_k}$$

Where Z_k is an indicator variable which equals one if any unit of the k[th] network is in the initial sample and equals to zero otherwise. y_k^* is the sum of Y values in the k[th] network. π_k is the probability of including an unit in the k[th] network in the sample. It is given by the following formula:

$$\pi_k = 1 - \frac{\binom{N-m_k}{n}}{\binom{N}{n}}$$

Where m_k is the size of the k[th] network.

Variance of $(\hat{\tau}_Y)_{HT}$ is given by the following expression:

$$V(\hat{\tau}_Y)_{HT} = \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j} y_i^* y_j^*.$$

The unbiased estimator of $V(\hat{\tau}_Y)_{HT}$ is given by:

$$v(\hat{\tau}_Y)_{HT} = \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}\pi_i\pi_j} y_i^* y_j^* Z_i Z_j.$$

Where π_{ij} is the joint probability of networks i and j being intersected by the initial sample. It is given as follows:

$$\pi_{ij} = 1 - \frac{\left[\binom{N-m_i}{n} + \binom{N-m_j}{n} - \binom{N-m_i-m_j}{n}\right]}{\binom{N}{n}}.$$

The modified HH estimator of τ_Y is given by:

$$(\hat{\tau}_Y)_{HH} = \frac{N}{n}\sum_{k=1}^{n} \bar{w}_k,$$

where the summation is over the units sampled and $\bar{w}_k$ is the average of Y values in the k [th] network.

The variance of $(\hat{\tau}_Y)_{HH}$ is given by:

$$V(\hat{\tau}_Y)_{HH} = \frac{N(N-n)}{n}\sum_{k=1}^{N} \frac{(\bar{w}_k-\mu_Y)^2}{N-1}.$$

where $\mu_Y = \frac{\tau_Y}{N}$.

The unbiased estimator of $V(\hat{\tau}_Y)_{HH}$ is given by:

$$v(\hat{\tau}_Y)_{HH} = \frac{N(N-n)}{n}\sum_{k=1}^{n}\frac{(\bar{w}_k-\hat{\mu}_{HH})^2}{n-1}.$$

where, $\hat{\mu}_{HH} = \frac{(\hat{\tau}_Y)_{HH}}{n}$.

3. Logarithmic Product Type HT estimator

Consider a rare and clustered population divided into N units. Let (X_i, Y_i) , $i = 1, 2, \ldots, N$ be the pairs of values of the auxiliary variable X and interest variable Y. The aim is to estimate the population total τ_Y of the interest variable. It is assumed that the population total of the auxiliary variable τ_X is known. Further, it is assumed that X and Y have a high negative correlation. In such a situation Murthy (1964) has defined product estimator of the population mean of Y and generalized by Gattone et al. (2016).

We propose the following logarithmic product type HT estimator of the population total τ_Y as follows:

$$\hat{\tau}_Y = (\hat{\tau}_Y)_{HT}\left[1 + log\left(\frac{(\hat{\tau}_X)_{HT}}{\tau_X}\right)\right]$$

Where $(\hat{\tau}_X)_{HT}$ and $(\hat{\tau}_Y)_{HT}$ are the HT estimators of τ_X and τ_Y respectively.

Define

$e_X = \frac{(\hat{\tau}_X)_{HT}-\tau_X}{\tau_X}$ and $e_Y = \frac{(\hat{\tau}_Y)_{HT}-\tau_Y}{\tau_Y}$

Hence,

$$(\hat{\tau}_X)_{HT} = \tau_X(1 + e_X) \text{ and } (\hat{\tau}_Y)_{HT} = \tau_Y(1 + e_Y)$$

Then we get,

$E(e_X) = E(e_Y) = 0$

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

$$V(e_Y) = \frac{1}{2\tau_Y^2}\sum\sum_{i\neq j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$COV(e_X, e_Y) = \frac{1}{2\tau_X \tau_Y} \sum \sum_{i \neq j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{x_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Hence $\hat{\tau}_Y$ can be written as:

$$\hat{\tau}_Y = \tau_Y (1 + e_Y)[1 + log(1 + e_X)]$$

$$\cong \tau_Y (1 + e_Y) \left[1 + e_X - \frac{{e_X}^2}{2} \right]$$

$$= \tau_Y \left[1 + e_X - \frac{{e_X}^2}{2} + e_Y + e_X e_Y \right]$$

Hence,

$$Bias(\hat{\tau}_Y) = E(\hat{\tau}_Y) - \tau_Y$$

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

$$- \frac{\tau_Y}{4\tau_X^2} \sum \sum_{i \neq j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right)^2$$

The mean squared error of $\hat{\tau}_Y$ is given as follows:

$$MSE(\hat{\tau}_Y) = E(\hat{\tau}_Y - \tau_Y)^2$$

$$= E \left[\tau_Y \left(1 + e_X - \frac{{e_X}^2}{2} + e_Y + e_X e_Y \right) - \tau_Y \right]^2$$

$$\cong E[\tau_Y (e_X + e_Y)]^2$$

$$= \tau_Y^2 \mathrm{E}(e_X + e_Y)^2$$

$$= \tau_Y^2 V(e_X + e_Y)$$

$$= \tau_Y^2 [V(e_X) + V(e_Y) + 2COV(e_X, e_Y)]$$

$$= \frac{1}{2} \sum \sum_{i \neq j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \frac{{\tau_Y}^2}{2\tau_X^2} \sum \sum_{i \neq j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right)^2$$

$$+ \frac{\tau_Y}{\tau_X} \sum \sum_{i \neq j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{x_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

4. Pilot Study

Pilot study was conducted by using NACS (Latpate and Kshirsagar (2019, 2018 a)). The interest was to estimate the total number of ever green plants which are rare in that region due to the presence of Basalt rocks.

The area of 100 acres in the Tamhini Ghat was divided into 100 plots each of size 1 acre and the percentage of silica observed on each of these plots was measured. Time required to measure the percentage of silica in a sample from one acre plot is fairly lesser than the time required to measure the number of evergreen plants in one acre. Secondly, the testing a soil sample for the percentage of silica is much cheaper than the cost incurred in counting the number of evergreen plants in one acre. The cost of testing a soil sample was \$2 and that of counting the number of evergreen plants in one acre was \$20. So, we considered the percentage of silica in one acre as the auxiliary variable.

Figure 1: Silica (S_iO_2)% on the different plots of the square region.

24 *	25	86 *	60	52	35	65	50	60	1
40	30	30	75	18	19	55	30	4	14
45	48 *	56	23	15 *	17	53	30	13	12
47	47	23	25	80	60	45	45	35	70
48	50	25 *	35	57	68	40	23 *	80	40
49	43	36	65	58	58	90	45	90	30
45	35 *	56	85	19	30	18	18	40 *	50
48	53	65	55	13	16 *	15	18	30	60
70	30	17	18 *	15	48	44	44	35	50
30	30	18	17	15	43	36	50	80	36

* in a square indicates selection in initial sample.

The nature of the soil in Western Ghats is of two types: Basalt rocks and Leterite. After studying the nature of the soil we had observed the abundance of evergreen plants whenever the silica content of the soil is 20 percent or less. Hence we considered $C_x = \{X \leq 20\}$ as the condition for adaptation.

A random sample of 10 plots was drawn from this area by using SRSWOR. The plots selected in the initial sample from this population related to the auxiliary variable X(percentage of silica in a plot) are shown by putting $*$in that plot as shown in Figure 1.

Then the procedure, negative adaptive cluster sampling was used. The networks were formed around the plots selected in the initial sample which satisfied the condition C_x. Each plot with $C_x = \{X > 20\}$ and selected in the initial sample formed a network of $size\ 1(shown\ in\ yellow\ colour)$. There were such 6 networks of size 1 selected in the initial sample from the above population of X variable. There was 1 network of size $13(shown\ in\ blue\ colour)$ and another network of size $4(shown\ in\ brown\ colour)$.Thus the total number of distinct networks in the sample was 8.

Figure 2: The values of the number of evergreen plants observed on the plots in the population.

*		*							**405**
				35	**20**			**306**	**130**
	*			**100** *	**65**			**107**	**108**
		*					*		
	*			**15**		**40**	**40**	*	
				120	**65** *	**95**	**30**		
		75	**32** *	**91**					
		36	**35**	**81**					

The clusters were formed by using auxiliary information and domain knowledge of Silica content and evergreen plants. These two variables are negatively correlated. It means that the abundance of Silica in soil leads to the rare evergreen plants. A cluster involves the network units and edge units. The edge units of clusters of size more than 1 were dropped to get the networks. Only those networks which satisfied the condition $C_x = \{X \leq 20\}$ were selected and were measured for the survey variable (number of evergreen plants) as shown in Figure 2. The total number of evergreen plants are 2031.

5. Results and Conclusions

For the computational efficiency in estimation, r number of repetitions were performed; where r varied as 3000, 6000, 9000, 12000, 15000 and 18000. We considered the initial sample sizes as 10, 20, 30 and 40 for each repetition.

The estimated population total over r repetitions is given by

$$\hat{\tau}_Y = \sum_{i=1}^{r} \frac{(\hat{\tau}_Y)_i}{r}$$

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 using Monte Carlo simulation study.

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

Table 1: Results of Simulation Study.

Number of repetitions (r)	Initial Sample Size (n)	Logarithmic Product HT Estimator		Murthy's Product Estimator	
		$\hat{\tau}_y$	SE($\hat{\tau}_y$)	$\hat{\tau}_y$	SE($\hat{\tau}_y$)
3000	5	1727.88	1751.66	1793.65	1798.99
	10	1965.91	1216.69	1997.97	1224.63
	20	1987.81	738.97	2001.99	740.31
	30	2009.27	482.04	2033.47	482.55
	40	2020.34	327.74	2031.32	324.21

6000	5	1676.82	1749.71	1858.17	1797.80
	10	1911.19	1228.06	1959.88	1235.21
	20	1995.22	742.01	2012.97	738.66
	30	2013.35	491.64	2020.18	490.78
	40	2021.52	327.69	2024.79	333.53
9000	5	1739.88	1745.49	1783.01	1776.21
	10	1930.00	1226.49	1934.34	1227.97
	20	1984.67	743.19	1994.93	744.06
	30	2013.69	483.50	2010.27	491.96
	40	2022.69	327.51	2027.47	319.09
12000	5	1697.04	1741.29	1790.97	1763.09
	10	1897.59	1221.03	1955.63	1228.38
	20	1981.20	740.74	2011.19	742.91
	30	2015.87	482.22	2021.91	489.74
	40	2022.84	326.03	2032.87	320.50
15000	5	1715.48	1741.79	1814.86	1761.41
	10	1908.05	1215.16	1944.04	1225.55
	20	1988.08	744.87	2008.52	741.13
	30	2017.02	489.44	2022.62	489.88
	40	2023.30	325.78	2028.86	323.26
18000	5	1705.20	1739.42	1819.27	1758.57
	10	1899.12	1218.43	1940.67	1234.10
	20	1993.31	740.55	2002.70	740.73
	30	2025.22	488.72	2019.14	493.43
	40	2025.83	324.21	2029.51	325.18

The results are shown in the Table 1 above. From that table it can be seen that the accuracy in the estimate of the population total increases with an increase in the number of repetitions and the size of the initial sample. The standard error of the estimator decreases with an increase in the initial sample size. This is consistent with the statistical regularity. Thus, the proposed estimator can be used effectively to estimate the population total of the interest variable when it is negatively correlated with the auxiliary variable. Also, logarithmic product type HT estimator is compared with Murthy's estimator. It is found that logarithmic product type HT estimator is superior for smaller sample size. For larger sample

size, both the estimators perform equally. For smaller sample size, logarithmic product type HT estimator is suitable. Hence, the cost of survey is minimized by using the proposed estimator with higher precision.

References

Bahl, S. and Tuteja, R. K. (1991): Ratio and product type exponential estimators. Journal of Information and Optimization Sciences, 12, pp.159-163.

Chao, C, T. (2004): Ratio estimation in adaptive cluster sampling. Journal of Chinese Statistical Association **42(3)**: 307-327.

Chao, C. T., Dryver, A. L., Chiang, T. C. (2011) : Leveraging the Rao-Blackwell theorem to improve ratio estimators in adaptive cluster sampling. Environmental and Ecological Statistics **18(3)**: 543-568.

Dryver, A. L. and Chao, C. T. (2007): Ratio estimators in adaptive cluster sampling. Environmetrics, **18(6)**, pp. 607-620.

Dryver, A. L., Netharn U. and Smith, D. R. (2012): Partial systematic adaptive cluster sampling. Environmetrics **23**: 306-316.

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

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

Latpate, R. V. and Kshirsagar, J. K. (2020): Two Stage Inverse Adaptive Cluster Sampling With Stopping Rule Depends upon the Size of Cluster: Sankhya-B: The Indian Journal of Statistics, Series B, Vol. 82-B, (1) pp. 70-83.

Latpate, R. V. and Kshirsagar, J. K. (2020). Two Stage Negative Adaptive Cluster Sampling. Communications in Mathematics and Statistics. Vol.8, (1) pp. 1-21.

Latpate, R. V. and Kshirsagar, J. K. (2019): Negative adaptive cluster sampling. Model Assisted Statistics and Applications, 14(1):65–81. DOI 10.3233/MAS-180452.

Kshirsagar, J. K. and Latpate, R. V. (2018a): Adaptive Sampling For the Improving Sample Performance(Ph.D thesis submitted to SPPU, Pune).

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

Lin, F. M. and Chao, C, T. (2014) : Variances and variance estimators of the improved ratio estimators under adaptive cluster sampling. Environmental and Ecological Statistics **221**: 285-311.

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

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

Thompson, S. K. (1990): Adaptive cluster sampling. JASA 85(412) pp. 1050-1058.

Thompson, S. K. (1991b): Stratified adaptive cluster sampling. Biometrika, **78(2)**, pp. 389-397.

Thompson, S. K. (2002). Sampling. Second Edition, Wiley Publications.