Adaptive Cluster Sampling – A Review

Okafor F. C.

Department of Statistics,
University of Nigeria, Nsukka

Abstract

Efficiency of random sampling for rare and spatial clustered populations is usually improved by either increasing the sample size or plot size in area sampling or both. Recently Thompson (1990) introduced a sampling method, called adaptive cluster sampling, which gives more precise estimate of parameters of rare events. In adaptive cluster sampling, in addition to the initial sample units, further units are canvassed depending on whether or not the required event is found in the initial sample unit or its neighbourhood. In this paper, our main aim is to introduce this method and to give a brief review of some research works in the area.

1.0 Introduction

The purpose of any sample survey is to estimate the population characteristics of interest. In conventional sample survey a probability sample is taken and each sample unit is observed for the variable (s) of interest and estimate(s) of the population parameter(s) is (are) formed based on the observed sample values. Any other population units outside the selected ones are not observed. In a population that has rare characteristics that are widely sparse or empty in some areas but dense or clustered in some other areas of the population or region, (for example fish or snail population or incidence of a rare contagious disease, land or water pollution, endangered species, in general populations that exhibit natural aggregation) the conventional sampling may not yield a good estimate of the population density of such rare characteristic because most of the sample units (areas) will be empty. In conventional sampling the solution to the rare event problem is to increase the sample size or the plot size in area sampling. This procedure seems not to be generally adequate. This led to search for a better sampling design outside of the conventional designs.

For population of this nature, Thompson [1] came up with a new sampling method, which produces an efficient estimate for rare characteristic, termed adaptive cluster sampling (ACS). In adaptive cluster sampling, an initial random sample of a given size is selected. A selected unit is observed, if it contains the characteristic of interest that satisfies a specified condition, the adjacent units of this initial unit (called neighbourhood units) are observed. Neighbourhood may be social or institutional relationship or may be based on degree of affinity or participation. If any other unit in the neighbourhood has the variable of interest that satisfies the set condition, its neighbourhood is also observed and added in the sample and so on, until no new unit satisfies the condition. These units including the edge units (units not meeting the set condition) form a cluster of units around the initial sample unit. Thus the addition of an adaptive unit depends on the observed value of the variable of interest and the set condition. All the units satisfying the set condition associated with the initial sample unit form what is called a network. Clearly, a network is a subset of the cluster. Basically, adaptive cluster sampling makes use of information obtained from an initial selected unit to determine whether to canvass for additional information from the adjacent units, not minding if this unit was originally in the sample or not.

2.0 Advantages and Disadvantages of Adaptive Cluster Sampling

The advantage of ACS is that in a multi-characteristic survey, a characteristic of interest may be rare and spatially clustered while others are not; the use of adaptive cluster sampling for one such characteristic will not affect the cost of observing and estimating other characteristics. Secondly, additional cost of sampling is triggered only when the characteristic of interest is observed. The third advantage is that ACS may help in identifying area of high abundance thereby increasing the information provided by the sample especially where disturbance does not initiate dispersion of species of interest.

Corresponding author: **-**, E-mail: -, Tel. +234 8057282029

The disadvantages include: In ACS the final sample size is not known until further addition of more units is terminated as a result the final sample size is random. Because of the random nature of the sample size, it is not feasible to determine the cost of the survey in advance. According to Smith et al [2] sampling of edge units is a penalty imposed by ACS, because edge units add to the cost of the survey but are not used in estimation unless the edge unit is part of the initial sample.

In the next sections we shall present some of the works done in ACS based on element sampling and other sampling designs. The articles are not exhaustive, only very few are presented here just to give an insight into the method of ACS.

3. ADAPTIVE CLUSTER SAMPLING PROPOSED BY THOMPSON [1]

Thompson [1] is the pioneer of ACS, which involves the selection of initial sample of *n* units by simple random sampling with or without replacement.

He gave two unbiased estimators of the population mean by modifying Hansen-Hurwitz and Horvitz-Thompson estimators. Since selection probability is not known for every unit in the final sample his estimators made use of observations not satisfying the condition only when they are selected as part of the initial sample. This gives the ACS modified Hansen-Hurwitz estimator of the mean as

$$t_1 = \frac{1}{n}\sum_{k=1}^{n}\bar{y}_k^*, \text{ where } \bar{y}_k^* = \frac{1}{m_k}\sum_{j\in A_k}^{m_k} y_j \tag{1}$$

A_k denotes the network that includes unit k and m_k is the number of units in A_k network. This estimator is similar to mean of cluster mean in single stage cluster sampling. The sampling variance of t_1 is given by

$$V(t_1) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\bar{y}_i^* - \mu\right)^2 \tag{2}$$

for simple random sampling without replacement.
And

$$V(t_1) = \frac{1}{n(N-1)}\sum_{i=1}^{N}\left(\bar{y}_i^* - \mu\right)^2 \tag{3}$$

for simple random sampling with replacement. The modified Horvitz-Thompson estimator of the mean is

$$t_2 = \frac{1}{N}\sum_{k=1}^{\upsilon} y_k^* \big/ \alpha_k \tag{4}$$

y_k^* is the sum of the y-values in the k^{th} network; υ is the number of distinct networks in the sample. α_k is the probability that k^{th} network is included in the sample. This probability is the same for all units in the i^{th} unit's network.

$$\alpha_k = 1 - \binom{N-m_k}{n}\Big/\binom{N}{n} \tag{5}$$

for simple random sampling without replacement.

$$\alpha_k = 1 - \left(1 - \frac{m_k}{N}\right)^n \tag{6}$$

for simple random sampling with replacement.
Thompson [1] gave the sampling variance of t_2 as

$$V(t_2) = \frac{1}{N^2}\left[\sum_{k=1}^{\xi}\sum_{j=1}^{\xi} y_k^* y_j^* \left(\frac{\alpha_{jk} - \alpha_j\alpha_k}{\alpha_k\alpha_j}\right)\right] \tag{7}$$

ξ is the number of mutually and exhaustive networks in the population.

$$\alpha_{kj} = \alpha_k + \alpha_j - (1 - p_{jk}) \tag{8}$$

is the joint inclusion probability of networks k and j.

p_{jk} is the probability of excluding networks k and j in the sample.

$$p_{jk} = \binom{N - m_k - m_j}{n} \Big/ \binom{N}{n}$$

for simple random sampling without replacement.

$$p_{jk} = \left(1 - \frac{m_j + m_k}{N}\right)^n$$

for simple random sampling with replacement.
Thompson [1] has shown that ACS can be more efficient than ordinary simple random sampling.

4.0 Illustration

Figure A shows an initial sample of n =8 plots (in bold and numbered a, b, c, d, e, f, g, and h) selected from N=528 plots in the population by simple random sampling without replacement.

Each sample plot is searched, if the number of diseases trees, $y_i > 1$ is found in the plot its adjacent plots forming a plus sign called neighbourhood are also searched. If any of these adjacent plots satisfies the condition it is adaptively added in the sample; while its neighbourhood plots are canvassed and added if the condition is met. This continues until no further plot meets the condition. All the adaptively added plots and those at the adjacent boundaries not meeting the condition make up the adaptive cluster sample. The plots at the boundary of the adaptive cluster not fulfilling the condition are termed the edge units. Adaptive cluster minus the edge units constitute a network. Figure B shows the networks in heavy bold. There are 8 distinct networks intersected by the 8 sample plots. We shall number the networks serially from *a* through *h*.

From Figure B we generate the following information in Table 1.
Table 1: Information generated from figure B on networks a to h.

Network, *k*	1 (a)	2 (b)	3 (c)	4 (d)	5 (e)	6 (f)	7 (g)	8 (h)
Size, m_k	1	1	7	1	17	1	13	1
y_k	0	1	31	0	81	0	67	0
α_k	0.01515	0.01515	0.1019	0.01515	0.2317	0.01515	0.1819	0.01515

Estimate of the mean number of disease trees per plot will be obtained using (4) as follows.

$$t_2 = \frac{1}{528}\left[0 + \frac{1}{0.01515} + \frac{31}{0.1019} + 0 + \frac{81}{0.2317} + 0 + \frac{67}{0.1819} + 0\right] = \frac{1088.15066}{528} = 2.0609.$$

					A											3							
																	1 b						
			4	5	1																		
			2	1																			
		3	1												d				2	3	1		3
		4	3 c	8															9	4	5	7	2
		4	7	2														1	5	3		8	5
																			7	6 e	2	2	
		1																	8	1			
										2	3	1								2			
										3	1												
										1	9	4											
										3	2	5	3	4									
							f			6	7 g	3	9	1									
										1	5	7		2									
5	9											1	7	1									
5	1												6										
9	8	1	3																				
			h																				

Figure A. Initial Sample of size 8 bold plots.

The variance is calculated with sample equivalent of (7) given by

$$\hat{V}(t_2) = \frac{1}{N^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{y_j y_k}{\alpha_j \alpha_k}\left(1-\frac{\alpha_j\alpha_k}{\alpha_{jk}}\right) \qquad (9)$$

$$= \frac{1}{N^2}\left[\left(\frac{y_2}{\alpha_2}\right)^2(1-\alpha_2)+\left(\frac{y_3}{\alpha_3}\right)^2(1-\alpha_3)+\left(\frac{y_5}{\alpha_5}\right)^2(1-\alpha_5)+\left(\frac{y_7}{\alpha_7}\right)^2(1-\alpha_7)\right.$$

$$+2\left\{\frac{y_2y_3}{\alpha_2\alpha_3}\left(1-\frac{\alpha_2\alpha_3}{\alpha_{23}}\right)+\frac{y_2y_5}{\alpha_2\alpha_5}\left(1-\frac{\alpha_2\alpha_5}{\alpha_{25}}\right)+\frac{y_2y_7}{\alpha_2\alpha_7}\left(1-\frac{\alpha_2\alpha_7}{\alpha_{27}}\right)\right.$$

$$\left.\left.+\frac{y_3y_5}{\alpha_3\alpha_5}\left(1-\frac{\alpha_3\alpha_5}{\alpha_{35}}\right)+\frac{y_3y_7}{\alpha_3\alpha_7}\left(1-\frac{\alpha_3\alpha_7}{\alpha_{37}}\right)+\frac{y_5y_7}{\alpha_5\alpha_7}\left(1-\frac{\alpha_5\alpha_7}{\alpha_{57}}\right)\right\}\right]$$

The joint probability of a network with $y_k = 0$ and any other network need not be computed because the result of zero divided by a number other than zero is zero. The computed joint probabilities from (8) are given in Table 2.

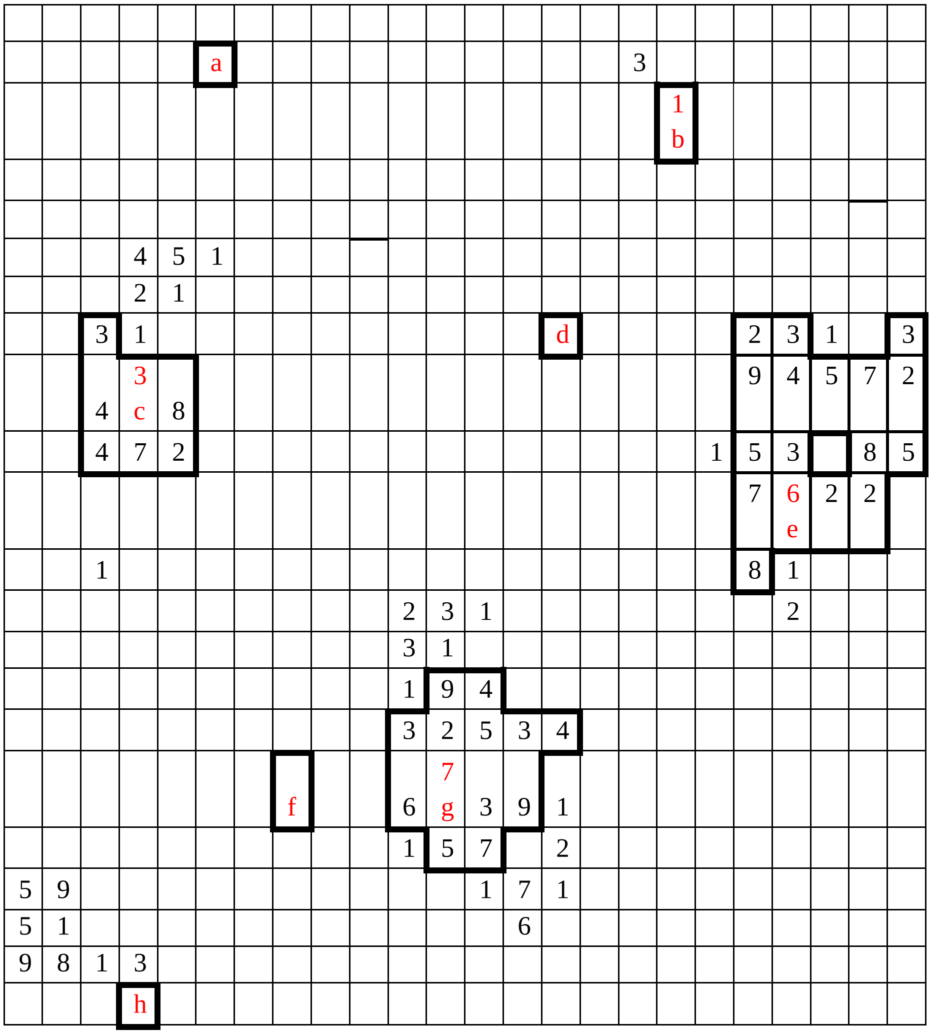

Figure B. Networks for the Initial Sample

Table 2. Computed Joint Probability

Joint Prob.	α_{23}	α_{25}	α_{27}	α_{35}	α_{37}	α_{57}
Value	0.00135	0.00315	0.00245	0.0211	0.0165	0.0379

Hence, the variance of the estimated mean is

$$\hat{V}(t_2)=\frac{1}{528^2}\Big[4290.8647+83118.8864+93896.3701+110991.7247 +2\{-2882.4466-2639.0390-3034.4471-12652.5352 -13824.2075-14426.6466\}\Big]=\frac{193379.202}{528^2}$$

= 0.69365

5.0 Element Sampling

Roesch [3] applied ACS in forest inventories to estimate the average percent defoliation per tree. Smith et al [2] evaluated the use of adaptive cluster sampling for three species of waterfowl using simulation study based on samples drawn from an enumeration of waterfowls in 5000 km^2 of central Florida. Two quadrant sizes of 25 and 100 km^2 were used in their study. According to the authors, the efficiency of ACS depends on the spatial distribution of the population sampled, quadrant size, initial sample size and condition to adapt. Some of these factors led to improvement of precision of the estimate while others did not. Christman and Pontius [4] investigated four bootstrap methods (Sitter's Mirror-Match, Gross, Rao and Wu's Rescaling and McCarthy and Snowden's Bootstrap with replacement) methods in setting up confidence interval. Overall, the study shows that the best methods depend on the type of population under study and the sample size. Using waterfowl data from Smith et al [2], Christman and Pontius [4] recommended the use of McCarthy and Snowden's method with percentile intervals when the modified Hansen-Hurwitz estimator is used. Dryver and Thompson [5] proposed two improved unbiased estimators of ACS conditional on sufficient statistic which is not minimal sufficient. These estimators make use of the edge units. Using Rao-Blackwell theorem the authors showed that the estimators are unbiased. Ojiambo and Scherm [6] investigated the efficiency of ACS in estimating the mean plant disease incidence. Using simulated data they demonstrated that ACS was most precise than simple random sampling for low level of disease incidence especially when diseased plants were highly aggregated and when the most liberal condition to adapt (CA =1 plant) was used. They also indicated that for ACS to be efficient, the final sample size should not be excessively larger than the initial sample size. Arabkhedri et al [7] applied ACS to the estimation of total suspended sediment load. The result of their study suggests that ACS can produce accurate estimate. Coggins et al [8] used ACS to improve the estimate of low level mountain pine beetle damage on trees using samples of airborne imagery.

6.0 ACS in Cluster Sampling

Thompson [9] considered the ACS design where the initial sample of *n* primary units is selected by simple random sampling without replacement with subsequent inclusions to the samples at the secondary unit level. Two estimators of the population mean were suggested. One is

$$t_3 = \frac{1}{Mn}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{y_k I_{ik}}{x_k} \text{ with variance given by}$$

$$V(t_3) = \frac{N-n}{M^2 nN(N-1)}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\frac{y_k I_{ik}}{x_k} - M\mu\right)^2$$

N is the total primary units in the population; *M* is the number of secondary units in each primary units.

x_k is the number of primary units in the population that intersect the k^{th} network. The draw-by-draw probability that the selected primary unit will intersect the k^{th} network is x_k/N .

$$I_{ik} = \begin{cases} 1 \text{ if } i^{th} \text{primary unit intersect the } k^{th} \text{network} \\ 0 \text{ otherwise} \end{cases}$$

The second is

$$t_4 = \frac{1}{MN}\sum_{k=1}^{K} y_k z_k / \pi_k \text{ with variance}$$

$$V(t_4) = \frac{1}{M^2 N^2}\sum_{k=1}^{K}\sum_{j=1}^{K} y_k y_j \left(\frac{\pi_{kj}}{\pi_k \pi_j} - 1\right)$$

$$z_k = \begin{cases} 1 \text{ if one or more primary units that intersect the } k^{th} \text{network} \\ \text{are included in the initial sample} \\ 0 \text{ otherwise} \end{cases}$$

π_k denote the probability that one or more of the primary units that intersect network *k* is included in the initial sample; and is computed from the sample.

$\pi_{kj} \neq 0$ is the probability that one or more of the primary units that intersect both networks *k* and *j* are included in the initial sample.

Salehi and Seber [10] developed the idea of a two-stage ACS. They used two designs – one where clusters are permitted to grow across primary sampling unit boundaries (overlapping scheme); second where clusters are truncated at the boundaries (non-overlapping scheme). According to authors the non-overlapping scheme appears to be more efficient.

7.0 Use of Auxiliary Information

Roesch [11] was first to combine the probability proportional to size sampling scheme with adaptive sampling scheme in forest survey of trees. An initial sample of trees was selected with probability proportional to basal area of the tree bole. If an initial sample tree j has pollution damage (i.e., $y_i = 1$) then tree *j*'s centre is used as the centre of circle with radius *r*. The value of *y* is observed for all trees within this circle which have not already been selected for the sample from this point. If for any of these trees $y_i = 1$, then we carry out the same procedure as for the first set of trees with $y_i = 1$. Otherwise adaptive is stopped when we encounter tree with $y_i = 0$ i.e. no pollution damage. Pontius [12] applied ACS to a case where the primary sampling units were selected with probability proportional to the sizes of the primary units. In each selected primary unit, every secondary unit is inspected for rare event; if condition is satisfied it is adaptively added. Thompson [13] used stratified random sampling scheme to select initial sample of units and adaptive procedure is then used to select additional units once a set condition is satisfied. He showed with illustrative small data set that stratified adaptive cluster sampling is more efficient than the conventional stratified random sampling for estimating rare and clustered population. Finally Felix-Medina and Thompson [14] proposed adaptive cluster double sampling and used regression estimator to obtain the estimate of the population mean.

8.0 Conclusion

From the researches carried so far it is clear that ACS was developed largely for ecological studies. It has been applied to survey on birds, amphibians, marine and aquatic species as well as tree species and tree diseases. Most of the work done so far on ACS is based on one item study. In a multi-subject survey where the survey design is generally based on the important species, species of secondary interest are estimated poorly. ACS involves extra cost in adding additional units in the neighbourhood that meets the condition, as a result adaptive strategy should be reserved for important characteristics that are known to be rare and found in clumps. While the conventional sampling based on the initial sample should be applied to common and abundant species in other to save cost. One setback in the use of ACS is that it may not be reliable for invasive species where the method of data collection disturbs the species in the neighbouring units that have the target organisms.

However, simulation studies on ACS have shown that increase in precision of ACS over simple random sampling depends on the spatial distribution of the population, initial sample size, sampling unit size and condition determining when to canvass neighbouring units.

Finally, going by the work done in ACS so far, much study has not been done in the application of ACS in some household (human population) surveys, like agricultural surveys, disease prevalence survey in humans and animals; and also in the use of auxiliary information. Presently ACS has not been used in Nigeria by researchers.

References

[1] Thompson, S.K. (1990): Adaptive cluster sampling: *Journ of Amer. Stat. Assoc.85, 412, 1050 – 1059.*

[2] Smith, D. R., Conroy, M.J. and Brakhage, D.H. (1995): Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics 51, 777 – 788.*

[3] Roesch, F. A. (1994): Incorporating estimates of rare clustered events into forest inventories. *Journal of Forestry 92, 12, 31- 34.*

[4] Christman, M, C and Pontius, J.S. (2000): Bootstrap confidence intervals for adaptive cluster sampling. *Biometrics 56, 503 – 510.*

[5] Dryver, A.L. and Thompson, S.K. (2005): Improved unbiased estimators in adaptive cluster sampling. *J. of Royal Stat. Soc series B 67, 1, 157 – 166.*

[6] Ojiambo, P.S. and Scherm, H. (2010): Efficiency of adaptive cluster sampling for estimating plant disease incidence. *Phytopathology 100, 7, 663 -670.*

[7] Arabkhedri, M, Lai, F.S., Noor-Akma, I. and Mohamad-Rosla, M.K. (2010): An application of adaptive cluster sampling for estimating total suspended sediment load. *Hydrology Research 41, 1, 63 – 73.*

[8] Coggins, S.B., Coops, N.C. and Wulder, M.A. ((2010): Improvement of low level bark beetle damage estimates with adaptive cluster sampling. *Silva Fennica 44(2) 290 – 301.*

[9] Thompson, S.K. (1991a): Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics 47, 1103 – 1115.*

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

[11] Roesch, F. A. (1993): Adaptive cluster sampling for forest inventories. *Forest Science, 39, 4, 655 – 669.*

[12] Pontius, J.S. (1997): Strip adaptive cluster sampling: probability proportional to size selection of primary units. Biometrics 53, 1092 – 1096.

[13] Thompson, S.K. (1991b): Stratified adaptive cluster sampling. *Biometrika 78, 2, 389 – 397.*

[14] Felix-Medina, M.H. and Thompson, S.K. (2004): Adaptive cluster double sampling. *Biometrika 91, 4, 877 – 891.*