Iranian Journal of Fisheries Science 3(1) 77-84 (2001)

Application of Adaptive Sampling in Fishery Part 2: Truncated Adaptive Cluster Sampling Designs

M. Salehi M.

School of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran

Email: salehi_m@cc.iut.ac.ir

Abstract: There are some experiences that researcher come across quite number of time for very large networks in the initial samples such that they can not finish the sampling procedure. Two solutions have been proposed and used by marine biologists which we discuss in this article: i) Adaptive cluster sampling based on order statistics with a stopping rule, ii) Restricted adaptive cluster sampling. Until recently, the unbiased estimators were not available for both sampling. Restricted adaptive cluster sampling was used (Lo *et al.*, 1997) under investigation in US Southwest Fisheries Science Center which was reasonably efficient even with a biased estimator. Salehi & Seber (2002) propose an unbiased estimator for this sampling design. They show that the unbiased estimator is shown to compare very favourably with the standard biased estimators, using simulation.

Key Words: Sampling design; Clump fish population; Abundance estimation

Introduction

The final sample size is random for adaptive cluster sampling designs, which adds uncertainty to survey planning. Misapplication of adaptive cluster sampling can result in failure to achieve desired precision or in a final sample size that exceeds a survey's budget. Excessive final sample size can result when the biological population is not as rare as or is more widely distributed than anticipated; practitioners of adaptive cluster sampling refer to this problem colloquially as encountering "the cluster from hell". In fishery study, two main

different truncated adaptive cluster sampling designs have been introduced in practice to control the final sample size which are discussed in two sections.

Adaptive cluster sampling based on order statistics with a stopping rule:

According to Quinn (1999), rockfish in the northeastern Pacific are notoriously difficult to sample because of their aggregated population. Trawl sampling with a stratified sampling design were used to assess their population status. However, the variance estimator were high. A cooperative project between the University of Alaska Fairbanks, Auke Bay National Laboratory, and the fishing industry commenced in 1998, seeking to improve estimates of rockfish abundance .A two-week survey was conducted in August 1998 according to an adaptive cluster sampling design in six strata. Ten to fifteen random tows were made in each stratum and the additional samples were added on order statistics (Thompson & Seber 1996; Chapter 6).To have some degree of controlling the final sample size, a stopping rule was implemented to curtail sampling after three to six level of additional sampling were conducted. They use the Hanson-Hurwitz (HH) estimator of Adaptive Cluster Sampling (ACS) based order statistics which is biased because of the implemented stopping rule. Adaptive cluster sampling showed to be a good method for rockfish populations, because high abundances were encountered in adjacent tows. The adaptive sampling design was easy to carry out and usually resulted in lower error than simple random sample. In adaptive cluster sampling based on order statistics (acsord), an initial simple random sample of size n is selected, producing an ordered list of sample values

$$y_{(1)} \leq y_{(2)} \leq \cdots y_{(n-r)} \leq y_{(n-r+1)} \leq \cdots y_{(n)}$$

An adaptive sampling phase is then carried out in the neighborhoods of the top r sample units whose y-values are greater than the condition $y > y_{(n-r)} = c$.

For Adaptive cluster sampling based on order statistics (acsord), the condition for additional sampling changes with different sample so that all the initial sample

must be observed before starting the adaptive sampling phase. It could require too much travel time back to the high density unit. As Quinn *et al.* have mentioned they will be looking at alternative cluster sampling designs for which the condition for adaptive sampling is determined before the survey. We of course reduce the travel time further using initial strip sampling rather simple random sample.

In addition to the above problem, the introduced estimator is biased. In spite of sampling problem of the bias method, it was relatively satisfactory.

Restricted adaptive cluster sampling:

Lo *et al.* (1997) used Restricted Adaptive Cluster Sampling (RACS) to estimate Pacific hake larval abundance. The RACS has been described by Brown (1994) in which selection of units for initial sample continues sequentially until a specified final sample size is reached. With this design, each time a unit in the initial sample is selected and observed, the associated adaptive additions of surrounding unit are carried out. If the cumulative total number of units in the sample equals or exceeds the sample size limit, no more initial units are selected and the sample size can then vary only by the size of the cluster associated with the last selected initial unit.

Brown (1994) applied the modified Horvitz-Thompson (HT) and Hanson-Hurwitz (HH) estimators by Thompson (1990), which are biased in this case. Brown & Manly (1998) estimated the biases by using the bootstrap method, and evaluated their method using simulation. They generated 27 populations with different degrees of patchiness. Bootstrapping was successful for removing the bias in only 8 populations using the HT estimator and 17 populations using the HH estimator. Salehi & Seber (2000) present an unbiased estimator for the RACS. In this article we are going to introduce the new notation and estimators of adaptive cluster sampling.

We need to define three things in order to use adaptive cluster sampling: (1) the condition C, usually at the form of $C = \{x_i > c\}$, where C is a constant; (2) the neighborhood of a unit, which includes the unit itself; and (3) an initial sample. The neighborhood of an initial unit is sampled whenever the unit satisfies

C .If any of these additional units satisfy C ,their neighborhoods will be added to the sample as well. This process continues until all the units in the network determined by the initial unit are sampled. If the initial unit does not satisfy C ,no further units are added, and the network consists of just the initial unit. We can therefore associate with each unit i a network A_i ,though some of these networks will be the same as several units which can belong to the same network. The distinct networks, labeled by the subscript $k(k=1,\ldots,K)$,will form a partition of the N units. The initial sample is selected at random and without replacement, and suppose f_i is the number of units in the initial sample which intersect A_i .Let m_i denote the number of units in A_i , and let m_i^* be m_i plus the number of edge units of A_i . The variable of interest associated with A_i is $y_i=\sum_{j=1}^{m_i} x_j$, called the y -value of the network, and $w_i=y_i/m_i$ is the mean of the x -values in A_i . We are interested in estimating $\tau=\sum_{l=1}^{N} x_i=\sum_{k=1}^{K} y_k$ using a restricted adaptive scheme which we now describe. Let ν_n^* be the total number of distinct units observed up to and including the n th initial selection. We stop selecting when $\nu_n^*\geq M$, where M is a predetermined number. Let k be the number ofdistinctnetworks intersected by the initial sample. The last selected network must not be already selected otherwise it will add nothing to the total number of distinct observed units. This network must also satisfy $m_n^*\geq M-\nu_{n-1}^*>0$. Suppose that the number of networks in the sample satisfying these two conditions is l , and these network ks are indexed as $k=1,\ldots,l$. Let L be the set of these l networks. The remaining networks are indexed as $k=l+1,\ldots,k$.

We also index the l units intersected by the initial sample as $i=1,\ldots,l$, and the remaining units as $i=l+1,\ldots n$. Thus, an unbiased estimator of population total, say τ , is

$$\hat{\tau}_R=\frac{N}{n-1}\left(\sum_{i=1}^{l}\frac{(l-1)}{l}w_i+\sum_{i=n+1}^{n}w_i\right) \qquad (1)$$

The unbiased variance estimator of (1) is given by

$$\hat{\text{var}}(\hat{\tau}_R) = \frac{l(l-2)N(N-n+1)-N^2(n-2)}{l^2(n-1)^2(n-2)} \sum_{i=1}^{l} \sum_{j \prec i}^{l} (w_i - w_j)^2$$

$$+ \frac{N(N-n+1)}{(n-1)^2(n-2)} \left(\frac{l-1}{l} \sum_{i=1}^{l} \sum_{j=l+1}^{n} (w_i - w_j)^2 + \sum_{i=l+1}^{n} \sum_{j>i}^{n} (w_i - w_j)^2 \right) \tag{2}$$

Example: Consider the population in Figure 1 containing $N = 400$ units, and a unit satisfies the criterion C if x, the number of objects in the unit, is greater than 0. We choose $M = 40$. The neighborhood of each unit consists of all adjacent units and the unit itself. Suppose, for demonstration purposes, that the ordered sample $((1,1,1,0),(2,6,13,36),(3,1,1,0),\ldots,(9,1,1,0),(10,11,24,107))$,
Where the components of the vectors are respectively the order of sample, size of network size of cluster, and the y_i^*. The sample set contains 8 networks of size 1, one network of size 6 and one network of size 11.
We have $n = 10$, $l = 2$, $w_1 = 6$, $w_2 = 107/11$ and $w_i = 0$ for $i = 3,\cdots 10$.
Thus we have,

$$\hat{\tau}_R = \frac{400}{9} \left(\sum_{i=1}^{2} \frac{(2-1)}{2} w_i + \sum_{i=2+1}^{10} w_i \right)$$

$$= \frac{400}{9} \left(\frac{1}{2} 6 + \frac{1}{2} 9.73 + 0 \cdots + 0 \right) = 349.5 \tag{3}$$

Thus unbiased variance estimator can easily evaluated.

As we see the unbiased estimator can be simply calculated. We can avoid repeated network in the sample which causes the estimator to be more efficient (less variance).

Restricted ACS with networks selected without replacement:

In the previous scheme, more than one initial unit can fall in a network so that a network can be selected more than once. Such a scheme might be called adaptive cluster sampling with networks selected with replacement. We now consider a restricted version of the design considered by Salehi & Seber (1997) in which networks are selected without replacement; we shall find that it is more convenient to work with networks rather than individual units. If the cost of `edge units' differs from non-edge units (Thompson & Seber 1996), we can work with the cost rather than the number of selected units. The scheme is now to continue sampling until the total cost reaches some predetermined value. Associated with each network k we define a cost, say, which can be a linear combination of and the number of edge units of network k, including just the sum of them. With the k th network we associate the vector (k, m_k, c_k, y_k).

Choosing the first unit leads to the selection of a network, which may just be the unit itself. We then `remove' that network, but not its edge units, from the population and choose a second initial unit at random from the remaining units. This will lead to a second network which is also `removed'. If $c_1 + c_2 \geq C_T$ we stop sampling. Otherwise we sample sequentially until $\sum_{k=1}^{n} c_k \geq C_T$, where n is the number of selected networks, and then stop. Note that we check the cost of the selected units after the second selection, as the variance estimator will only exist if at least two networks are selected.

Calculation of Estimators for this design is complicated. On the other hand, Salehi & Seber (2002) showed that there is not significant difference between the efficiencies of with and without replacement sampling designs so that we do not discuss them and we refer readers to Salehi & Seber (2002).

Discussion

Adaptive cluster sampling is an efficient method relative to conventional methods (e.g. simple random sample, stratified sampling etc.) for sampling rare

and clustered populations when cluster sizes are large relative to unit sizes (Seber & Salehi, 2002). This is the case for marine organisms so that we expect that adaptive cluster sampling designs are suitable ones. But, there is a question to be answered, that if this design is more appropriate for rare and clustered biological organisms such as fishes. It seems that the answer depends on the sort of marine organism and many other local parameters. To explore which design is most appropriate, we should locally run an experiment(s). However, we should look at strip versions of ACS designs so that we reduce the traveling cost for a fishery study.

The need to sequentially select initial units at random precludes using the most efficient travel path between selected units (strips) in the study region. We suggest select an initial sample of size, the predetermined final sample size (η) and, find the best route of shipping for all this selected initial sample. We then go to their sites in the study region and measure the variable of interest. If the variable of interest satisfy the condition C we carry out the adaptive phase. The process of measuring the "η" selected initial sample units will stop when the number of final sample size pass M for first time.

In order to investigate optimal properties of a sampling design we must consider a model for population. This approach is known as "model-based" or "superpopulation" approach. For this model-based approach, Thompson & Seber (1996) indicate which of the results for conventional designs carry over to adaptive designs and which of those do not. They also showed that optimal designs tend to be adaptive (Chapter 10).

Acknowledgment

I would like to express my sincere thanks to Dr. B. M. Amiri and three anonymous referees for their comments and suggestions.

I also would like to acknowledge financial support provided for this research by the BENEFIT institution, South Africa.

References

Brown, J.A. , 1994. The application of adaptive cluster sampling to ecological studies. *In:* statistics in ecology and environmental monitoring (eds. D.J. Fletcher and B.F.J. Manly) University of Otayo Press. pp.86-97.

Brown, J. A. & Manly, B. F. J. , 1998. Restricted adaptive cluster sampling. Environmental and Ecological Statistics. **5**. pp. 49-63.

Lo, N. C.H. ; Griffith, D. and Hunter, J.R. , 1997. Using a restricted adaptive cluster sampling to estimate Pacific hake larval abundance. Calif. Coop. Oceanic Fish. Invest. Rep. **38:**103-113.

Quinn II, T.J. ; Hanselman, D.H. ; Clausen, D.M. ; Heifetz, J. and Lunsford, C. 1999. "Adaptive cluster sampling of rockfish populations" Proceedings of the American statistical Association, Joint Statistical Meetings, Biometrics Section. pp.11-20.

Salehi M., M. and Seber, G.A.F. , 1997. Adaptive cluster sampling with networks selected without replacement. Biometrika. **84:** 209-219.

Salehi M., M. and Seber, G.A.F. , 2002. "Unbiased Estimators for Restricted Adaptive Cluster Sampling " {\it Australian and New Zealand Journal of Statistics. **44(1):** 63-74.

Seber, G.A.F and Salehi M., M. , 2002. Adaptive Sampling, Encyclopedia of biostatistics Volume 1. 2nd edition. (Eds. Peter Armitage and Theodore Colton). John Wiley & Sons, Ltd, Chichester, in press.

Thompson, S.K. , 1990. Adaptive cluster sampling. Journal of the American Statistical Association. **85:**1050--1059.

Thompson, S.K. and Seber, G.A.F. 1996. Adaptive Sampling, Wiley , New York. U.S.A. 265 P.