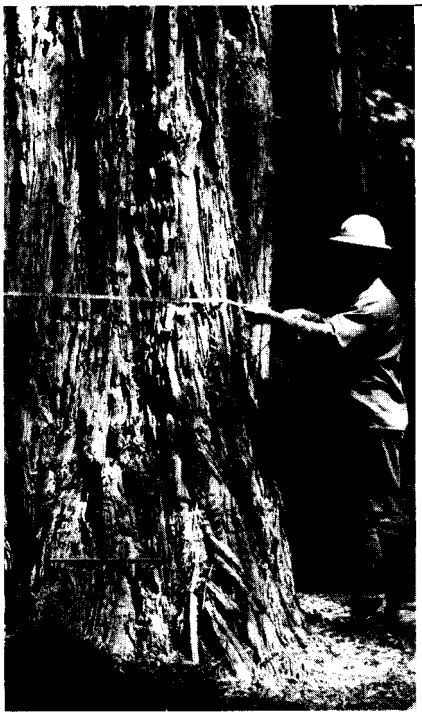

Incorporating Estimates of Rare Clustered Events into Forest Inventories

By Francis A. Roesch

Usually foresters are expected to answer a diverse set of questions about the forest on a very limited budget. The answers to such questions are considered important even though they often deal with conditions that occur only rarely in the forest. If a forester had the financial resources and knew where such conditions existed, the corresponding questions could be answered by designing and conducting an inventory in the vicinity of the rare occurrence. In most cases neither the money nor specific enough knowledge of the condition are available to use this approach.

An alternative approach is to modify an existing inventory to address a question by adapting the field procedure only when the existence of a rare condition is noted. Naive attempts at this sort of strategy will often lead to biased estimates of the condition, so the inventory and analysis must be carefully designed to avoid potential bias. Roesch (1993) showed how to do this for forest inventories using adaptive cluster sampling.

Adaptive cluster sampling is very efficient if the rare condition occurs on clustered trees, and it has two advantages over other methods of estimating rare conditions. The first advantage is that the decision to use an adaptive scheme can be made on a condition-by-condition basis, so adapting the sample for one condition does not affect the cost of estimating other conditions. The second advantage is that only the presence of the condition triggers additional cost. Most other solutions to the rare event problem, such as increasing the size or number of plots, do not have these two desirable properties. This article discusses in further detail the necessary considerations when contemplating an adaptive design and gives estimators for a population-based mean and its variance.

Adaptive Sampling in Forest Inventories

In adaptive cluster sampling, a sample of units in a population is taken and then additional units are selected near those that display any rare condition of interest. Roesch (1993) combined the probability proportional to size sampling schemes common in forestry with an adaptive sampling scheme, resulting in a system that can be applied to many in-place forest inventory systems.

First, sample trees are selected by an inventory rule such as those corresponding to fixed-area plot sampling or point sampling. We will assume that the initial selection of trees is by point sampling. Readers unfamiliar with point sampling can find explanations in standard forest mensuration texts such as Husch et. al. (1982) or Avery and Burkhart (1983). If a sample tree displays some rare condition of interest, additional trees within a fixed radius of the sample tree are examined for that condition. This is repeated for every tree displaying the condition until no new trees are found with the condition.

It is the forester's task to determine a radius that will identify a reasonable number of additional trees for the sample; enough additional trees to provide an estimation advantage and few enough to be considered during the measurement of the field inventory plot. This requires the forester to consider all of the available information on the spatial distribution of the rare condition. Negligence in this task could result in unmanageably large numbers of trees being encountered on some plots. This extra effort in the design stage will be re-

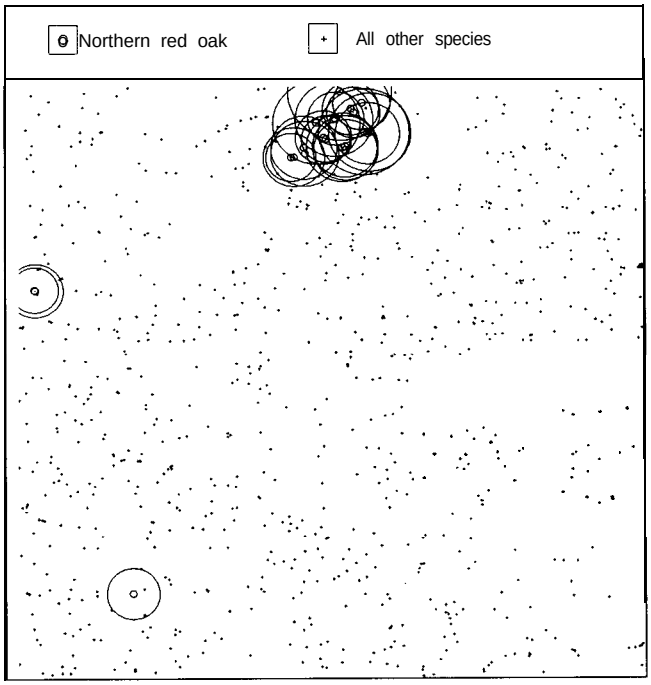

Figure 1. The spatial locations of the trees on a 3.1 l-acre simulated forest, with northern red oak trees differentiated from other tree species. The point sample selection areas at BAF 10 for the northern red oak trees are also shown as circles around each tree. A tree is sekctedfor the point sample if a random point lands within its selection area.

warded by the increased efficiency of a well-planned adaptive sampling survey.

Assume that the tree is the sampling unit and that there are N trees in the forest with labels **1, 2, ... N.** Associated with the N trees are values of a specific characteristic $y=\{y_1, y_2, , y_N\}$. Interest is in estimating the population mean of the y-values $(\bar{y})$. The forester is interested in many different $\bar{y}$s, such as the proportion of trees of a particular species, the average cubic-foot volume of wood, or the proportion of trees supporting a particular parasite, although we only need to consider one $\bar{y}$ at a time.

For example, suppose a forester is interested in the mean percent defoliation by gypsy moth on the northern red oak trees (Quercus ***rubra*** L.) in a forest and the inventory occurs in an area where either northern red oak trees or gypsy moths are rare. If tree ***i*** is a northern red oak tree

Search area (acres)	Network
1/100 1/90	6
1/80	4
1/70 1/10	3

Figure 2. The positive networks formed at each search-area size. The numeral in each plot is the number of positive networks.

that has been defoliated to some extent by gypsy moths, then the condition for tree ***i*** equals 1 (C_i = 1) and y_i = percent defoliated, otherwise both $C_i = 0$ and y_i = 0. The field crew would take the following steps:

1. Conduct the point sample and observe and record C_i for all northern red oak trees i,
2. For all trees with C_i = 1 conduct the adaptive part of the sample:
   a. measure y_i and diameter at breast height (dbh_i) and record the location of tree i;
   b. observe C_i for all northern red oak trees within a circle of radius r from the center of tree ***i*** that have not already been sampled;
   c. i. ignore all new trees for which C_i = 0
      ii. for all new trees with C_i = 1 record C_i and return to (a); and
   d. stop when C_i = 0 for all new trees.

A network is the set of trees such that selection of any tree within the network by the original sample (Step 1 above) will lead to the selection of every other tree in the network. Since selection of trees for which C_i = 0 will not result in the selection of any other trees, these trees are networks of size 1. This procedure maps the population of Ntrees into a population of Mnetworks, conditioned on C= $\{C_1, C_2, , C_x\}$. This can be done for as many conditions as desired.

Trees for which C_i = 0 are ignored unless they are in the original point sample, because only those northern red oak trees that were in the initial sample (Step 1) and those additional northern red oak trees (from Step 2) for which C_i = 1 will be used in the estimators. This results in unbiased estimators, as shown in Thompson (1990) and extended to point sampling by Roesch (1993). The estimators below differ from those in Roesch (1993) in that they are population-based rather than area-based (i.e., the mean per tree rather than the mean per hectare is of interest).

Estimator of the Mean per Tree

The total of the observations over network Kis

$$t_K = \sum_{j=1}^{v_K} y_j$$

where v_K is the number of trees in network K.

Roesch (1993) showed that the probability of tree k, in network K, being included in the sample from at least one of m random points is:

$$\alpha_k = \alpha_K = 1 - \left(1 - \frac{a_K}{L}\right)^m$$

where a_K = union of the point-sample selection areas for the trees in network Kto which tree ***k*** belongs, and L = the total area of the forest. (Note that capitalized subscripts are used when the quantity pertains to the network and lower case subscripts are used when the quantity pertains to the tree.)

To calculate the joint selection areas of trees, the dot count method, which is well known to foresters, could be used. One simply plots the selection areas of trees in the network and randomly

places a square grid of dots over the mapped selection areas. The dots within the selection areas are counted, and the number of dots is multiplied by the area represented by each dot. This method can be used to any desired level of accuracy by adjusting the size of the grid. Ideally the grid should be fine enough so that its error is smaller than the area1 errors associated with the measurement error of tree dbh and location.

A Horvitz-Thompson estimator (Horvitz and Thompson 1952) of the mean defoliation per northern red oak tree is:

$$\bar{y} = \sum_{K=1}^{\kappa} \bar{y}_K$$

where Kequals the number of distinct networks of northern red oak trees in the initial point sample and

$$\bar{y}_K = \frac{\left(\frac{t_K}{\alpha_K}\right)}{\sum_{J=1}^{\kappa}\left(\frac{v_J}{\alpha_J}\right)}$$

There is not a universally "good" estimator of the variance ofj by the most common criteria, such as minimum mean squared error (MSE), etc. Below are two possible variance estimators. The first can be formed by initially noting that the joint probability of including networks /and *Hat* least once from the *m* random points is:

$$\alpha_{JH} = 1 - \left\{\left[1-\frac{a_J}{L}\right]^m + \left[1-\frac{a_H}{L}\right]^m - \left[1-\frac{U_{JH}}{L}\right]^m\right\}$$

where U_{JH} is the union of the selection areas of networks /and *H*. Then a variance estimator is:

$$s_1^2(\bar{y}) = \left(\frac{1}{\kappa^2}\right)\sum_{K=1}^{\kappa}\sum_{H=1}^{\kappa}\bar{y}_K\bar{y}_H\left(\frac{\alpha_{KH}-\alpha_K\alpha_H}{\alpha_{KH}}\right)$$

A variance estimator that is somewhat easier to calculate because it does not require the determination of the joint network probabilities is:

$$s_2^2(\bar{y}) = \left(\frac{1}{R}\right)\sum_{K=1}^{\kappa}\left(\bar{y}_K - \frac{\bar{y}}{\kappa}\right)^2 * n_K$$

where n_K is the number of points from which nerwork Kis chosen (will almost always be equal to 1) and

$$R = \sum_{K=1}^{\kappa} n_K$$

The long-run behaviors of these estimators are examined briefly in the example below.

An Adaptive Sampling Example

The data used in this simulation were a subset of those collected by the USDA Forest Service Northeastern Forest Experiment Station in Hancock County, Maine. Fifty-three circular 1/10-acre plots were established in 1968 and remeasured in 1981. All trees that were at least five inches in dbh were measured. The azimuth and distance from plot center to each tree were recorded to the nearest degree and 1/10 foot respectively. Of the recorded attributes species, location, and dbh from 198 1 were used. A percent defoliation was arbitrarily assigned to each northern red oak tree to estimate the mean percent defoliation.

For the simulation, a highly diverse "forest" was created by cutting the largest square possible out of the circular 1/10-acre plots, each side of the square facing one of the cardinal directions. The first 49 of these squares (by plot number) were used in a 7 x 7 arrangement to simulate a square forest of approximately 3.11 acres. *Figure 1* shows the spatial locations of both the northern red oak trees and all of the other trees in the forest.

Note that in lieu of the adaptive scheme the point sample per-tree estimator would be:

$$\hat{y} = \frac{\sum_{h=1}^{m}\sum_{i=1}^{n_h}\frac{y_i}{a_i}}{\sum_{h=1}^{m}\sum_{i=1}^{n_h}\frac{1}{a_i}}$$

where n_h is the number of trees sampled from point *h*. Since the extra work involved in the adaptive sampling scheme is justified by the estimation advantage over point sampling, we compare the resulting adaptive sampling estimates with the point sample estimates. Five thousand random samples of 30 points each at a basal area factor (BAF) of 10 ft^2/acre were simulated and $\hat{y}$ was calculated for each sample. For 10 different search-area sizes $\bar{y}$ was calculated using the same random points. The search-area sizes ranged from $s = 10$ to $s = 100$ in increments of 10, where $1/s$ equals the area of a search circle in acres. This many search-area sizes were used to show that there is an optimal search-area size for this problem. This will usually be the case, although the forester's prior knowledge will often allow only an approximation to this optimal size. *Figure 2* gives the "positive" networks (i.e., networks of trees for which $C_i = 1$) formed at the different search area sizes.

Results

Table 1 gives the summary statistics over the 5,000 samples. The MSE ratios of MSE $(\bar{y})$/MSE$(\hat{y})$ at each search-area size show the general reduction in MSE due to the additional information from the adaptive part of the sample. These are only 56% for the larger search areas and only 60% even for the smallest search area. In comparing this table with figure 2, we note that no difference occurs in the MSE for the adaptive sampling estimator if the search areas yield the same positive networks, since the estimator does not change. The ratio between the mean of each of the two variance estimators ($s_1^2(\bar{y})$ and $s_2^2(\bar{y})$) and MSE $(\bar{y})$ is also given, as an indication of the variance estimators' performance. This ratio should

Table 1. Statistics from 5,000 simulations of 30 points each.

s[a]	$\frac{MSE(\bar{y})}{MSE(\hat{y})}$	$\frac{mean(s_1^2(\bar{y}))}{MSE(\bar{y})}$	$\frac{mean(s_2^2(\bar{y}))}{MSE(\bar{y})}$	No. A.S.[b] oaks x10^5	No. A.S.[c] others x10^5
10	0.56	1.6	1.2	3.8	8.40
20	0.56	1.8	1.2	3.8	3.40
30	0.56	1.6	1.2	3.8	2.20
40	0.66	1.6	1.2	3.8	1.40
50	0.56	1.6	1.2	3.8	1.00
60	0.56	1.6	1.2	3.8	0.52
70	0.56	1.6	1.2	3.8	0.49
80	0.58	1.8	1.7	3.0	0.28
90	0.60	1.4	1.3	2.3	0.12
100	0.60	1.4	1.3	2.3	0.12

NOTE For comparison 1.1 x 10^5 northern red oaks were encountered on the point sample alone.

[a] The search-area size is 1/s.

[b] The number of northern red oaks encountered in the adaptive part of the sample.

[c] The number of trees other than northern red oaks encountered in the adaptive sample.

be close to 1 over the long-run. Given the rarity of northern red oak trees in the forest, the performance of the variance estimators could be considered fair, although $s_1^2(\bar{y})$ didn't work quite as well as $s_2^2(\bar{y})$ in this case. Further analysis of the spread of the variance estimates not reported here also favored $s_2^2(\bar{y})$. This is due more to the small sample sizes in the simulation than to any fault of $s_1^2(\bar{y})$ since reliable variance estimation requires fairly large sample sizes.

The last two columns of ***table I*** illustrate the true advantage of choosing a reasonably small search area. The number of trees of other species that were encountered during Step 2 over the 150,000 points goes from the very large value of 8.4 x 10^5 at a search area of 1110 of an acre to the much smaller value of 4.9 x 10^4 by 1/70 of an acre, while the number of extra northern red oak trees, and therefore the amount of additional information, remained the same. This extreme difference in the number of extra trees looked at and then ignored for different sized search areas will be observed whenever one is adapting for a truly clustered event. In this case the largest reduction in MSE with the least amount of extra work would have been achieved with a search circle of 1170 of an acre.

Discussion

A major concern when choosing between sampling strategies is cost. The additional cost of the adaptive strategy for a particular application depends on relative cluster size and frequency of occurrence in the sample. These factors can be controlled by the inventory designer, given adequate prior knowledge of the population. In the example above, the additional cost of including extra trees was shown to be controllable by the distance examined for additional gypsy moth damage. Within this distance, a tree's species must be determined and, if the tree is a northern red oak with gypsy moth damage, its dbh and location recorded. If northern red oak trees with gypsy moth damage are truly rare and found in clusters, and an appropriate search area is used, then the additional cost will be small, due to the rareness of the species. The estimate of the per-tree percent defoliation by gypsy moth will be improved, due to the clustering of the species. The size of the search area determines the size of the networks of interest found as well as the number of additional trees encountered. Therefore, for any attribute one would want to select a minimally sized search area dependent on the expected proximity of the target trees to each other.

A question some managers may ask is, "What do we do when, halfway through the field season, our crew runs into a much more concentrated occurrence of something we are adapting for than we ever expected?" This large concentration might result in the field crew failing to complete the plot. Although this situation is best avoided, we might want an ad hoc solution to minimize losses. A very unsatisfactory solution would be to throw out all of the data from the adaptive part of the sample for this condition. A better solution would be to adjust the search-area size down to the largest area that would make this problem plot manageable. Data is eliminated from trees on this and previous plots that would not have been encountered at the new search-area size. The field season is then finished using the new search-area size for this condition. This solution will result in the same estimates that would have been obtained had the smaller search area been used from the start. The analytical expression for the variance of the estimators becomes complicated, however, because there was some probability of not encountering the problem site and completing the field season with the larger search area. This solution will be easier to implement if field crews are in the habit of first completing the nonadaptive part of the sample and then recording the closest trees first in the adaptive part of the sample.

Field foresters are often tempted to "take more data" when they encounter something special during an inventory. This temptation indicates a healthy concern for the resource. Foresters know that today's rare event could be the harbinger of tomorrow's significant effect. Adaptive cluster sampling provides a way for foresters to monitor many of these rare events at once at a relatively small cost. It allows flexibility in the inventory in that once a particular condition becomes less rare, the adaptive sampling procedure can be dropped for that condition, and other conditions can be added to the list almost at will. Perhaps more than other sampling schemes it imparts a responsibility to the forester because its success and efficiency is ensured by and dependent on thorough planning prior to each season's inventory. Some might fault this sample design on that basis; however all inventories require careful planning, and most foresters are up to the task. JOF

literature Cited

AVERY, T.E., and H.E. BURKHART. 1983. Forest measurements. McGraw-Hill, New York. 331 p.

HORVITZ, D.G., and D.J. THOMPSON. 1952. A generalization ofsampling without replacement from a finite universe. JASA 47:663–85.

HUSCH, B., C.I. MILLER, and T.W. BEERS. 1982. Forest mensuration. John Wiley & Sons, New York. 402 p.

ROESCH, EA. JK. 1993. Adaptive cluster sampling for forest inventories. For. Sci. 39(4):655–69.

THOMPSON, S.K. 1990. Adaptive cluster sampling. JASA 85:1,050–59.

ABOUT THE AUTHOR

Francis A. Roesch is a mathematical statistician, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, Room T-10210, 701 Loyola Ave., New Orleans, LA 70113.