Searching for American Chestnut: The Estimation of Rare Species Attributes in a National Forest Inventory

Francis A. Roesch[1] **and William H. McWilliams**[2]

Abstract.—American chestnut, once a dominant tree species in forests of the Northeastern United States, has become extremely rare. It is so rare, in fact, that on completion of 80 percent of the plot measurements of the U.S. Department of Agriculture Forest Service's most recent inventory in Pennsylvania, only 33 American chestnut trees with a diameter at breast height ≥ 1.0 in were found, out of 72,416 sampled trees. This paper discusses auxiliary sampling strategies that allow Forest Inventory and Analysis (FIA) units to estimate rare species in general as a first step in considering the especially difficult problems that American chestnut poses. The strategies involve (1) an increase of the initial plot size, (2) the use of adaptive cluster sampling, and (3) a combination of the first two. Adaptive cluster sampling was developed for the estimation of rare clustered events and is considered here because American chestnut is not only rare but also known to occur almost exclusively in clusters.

American chestnut (*Castanea dentata* (Marsh.) Borkh.), once a dominant tree species in Eastern U.S. forests, has become extremely rare in those same forests (McWilliams *et al*, 2006). It is so rare, in fact, that on completion of 80 percent of the plot measurements of the U.S. Department of Agriculture Forest Service's most recent inventory in Pennsylvania, only 33 American chestnut trees were found out of 72,416 sampled trees. This paper explores adaptations to the Forest Inventory and Analysis (FIA) sample design for estimating attributes of rare species in general as a first step in considering the especially difficult problems that American chestnut poses.

National inventories are best suited to (and funded for) small-scale problems such as the desire to estimate a level of X per million hectares. Related large-scale attributes and rare events, however, are often of disproportionate interest, which results in a general scale problem within the inventory because the rarer an event is, the greater its variance of observation will be and the higher the probability is that the event will be missed entirely by a small-scale inventory.

A few alternative approaches to detecting and estimating rare events would be to increase the sample size, increase the sample complexity (by adding a stage or phase, for example), proportionally or optimally allocate the sample, increase the size of the observation unit, or use adaptive cluster sampling. Here we consider the following options that are readily available to FIA for increasing the sample of American chestnut without increasing the number of sample points:

(1) Increase the size of the sample units utilizing the existing design features, and alter the size distributions selected by the components of FIA's tri-areal design within the natural range of American chestnut (fig. 1), to wit:
   a. Sample chestnut trees with diameter at breast height (d.b.h.) from 1.0 to 5.0 in on the subplot rather than the microplot.
   b. Use the existing design's previously developed macroplot to sample all chestnut trees larger than a breakpoint diameter.

(2) Use adaptive cluster sampling with search circles of a fixed size dependent upon the expected intra-cluster distribution (as in Roesch 1993).

(3) Some combination of options 1 and 2.

Although a detailed discussion of this point is beyond the scope of this article, a minor modification of option 3 could be used for increased efficiency in the estimation of American chestnut.

[1] Mathematical Statistician, U.S. Department of Agriculture (USDA), Forest Service, Forest Inventory and Analysis, Southern Research Station, Asheville, NC.
[2] Supervisory Research Forester, USDA Forest Service, Forest Inventory and Analysis, Northern Research Station, Newtown Square, PA.

1990). The probability (p_i) of using tree i in an estimator is equal to the union of the selection areas of each tree in the network (a_i) to which it belongs, divided by the area of the forest (L_F).

Estimator of the Population Total

Thompson (1990) showed that an unbiased estimator can be formed by modifying the Horvitz-Thompson estimator (Horvitz and Thompson 1952) to use observations not satisfying the condition only when they are part of the initial sample. We can calculate the probability that a tree is used in the estimator even though its probability of being observed in the sample is unknown. The probability of tree k, in network K, being included in the sample from at least one of m plots is:

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

where:

a_K = union of the inclusion areas for the trees in network K to which tree k belongs.

L_F = the total area of the forest.

For the HT estimator, let:

$$J_k = \begin{cases} 0 & \text{if the kth tree does not satisfy the condition and is not selected in the sample, otherwise} \\ 1 & \end{cases}$$

Then sum over the v distinct trees in the sample:

$$t_{HT} = \left(\frac{1}{L_F}\right) \sum_{k=1}^{v} \left(\frac{y_k\ J_k}{\alpha_k}\right)$$

The statistical properties of t_{HT} and other adaptive sampling estimators are discussed in Roesch (1993).

As its name implies, adaptive cluster sampling can be very efficient if the rare condition is distributed in clusters. In adaptive cluster sample designs, a compromise must be found between the level of new knowledge attained and survey cost. Adaptive sampling has at least three advantages: (1) it is efficient because only the presence of American chestnut triggers additional effort and cost; (2) it can be used on an attribute by attribute basis, so adapting the sample for estimation of American chestnut does not affect the cost of other estimates; and (3) it nullifies the weakness of the existing FIA design for the estimation of rare events. Its disadvantages include the potential for field crew confusion with respect to species-specific search rules and the identification of plots in high-probability areas, and the necessity for additional theoretical development and explanation for FIA practitioners and data users.

Simulation

To illustrate the considerations that must be taken into account when choosing between these options for sampling rare events, a simulation utilizing the same population described in Roesch (1993) was conducted. In brief, the simulated population was built using the coalesced 1981 FIA plot data from Hancock County, ME, as seed data. The data were chosen because they were conveniently on hand and were sufficient to illustrate the attributes of these sampling options. Ten sample points were applied to the population 1,000 times and the following four sample designs for eight rare tree distributions within the population were compared:

(1) Bi-areal design.
  - Microplot: d.b.h. < 5.0 in
  - Subplot: d.b.h $\geq$ = 5.0 in

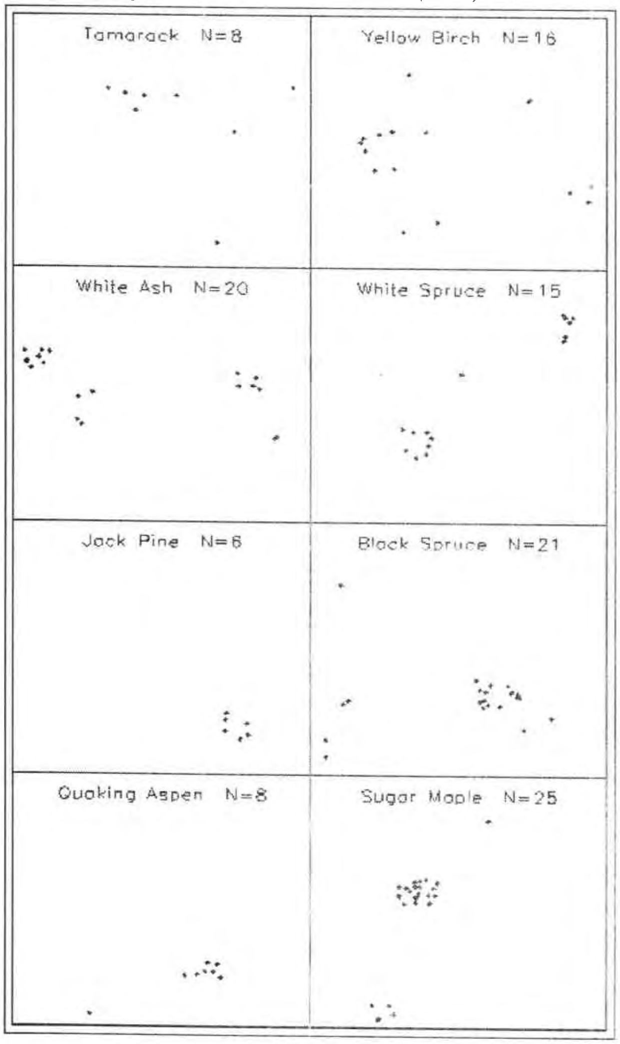
(2) Tri-areal design—breakpoint diameter (9, 12, 15, and 18 in).
  - Microplot: d.b.h. < 5.0 in
  - Subplot: 5.0 in $\leq$ = d.b.h. < breakpoint diameter
  - Macroplot: d.b.h. $\geq$ = breakpoint diameter

(3) Adapted bi-areal design.
  - Search radii of 20, 30, 40, 50, and 60 ft

(4) Adapted tri-areal design.
  - Search radii of 20, 30, 40, 50, and 60 ft

We estimated total basal area and mean squared error (MSE) for the eight rare species whose spatial distributions are plotted individually in figure 3 for each variation of each design. Design 1 is the default design that would be used if no special consideration were given to the rare species. The varying breakpoint diameters affect designs 2 and 4 while the varying search radii affect designs 3 and 4.

Figure 3.—*The spatial locations of the eight rare species in the simulated population described in Roesch (1993).*



Results

Figures 4 through 7 show the simulation's calculated ratios of the MSEs of designs 2, 3, and 4 to design 1 for breakpoint diameters 9, 12, 15, and 18 in, respectively.

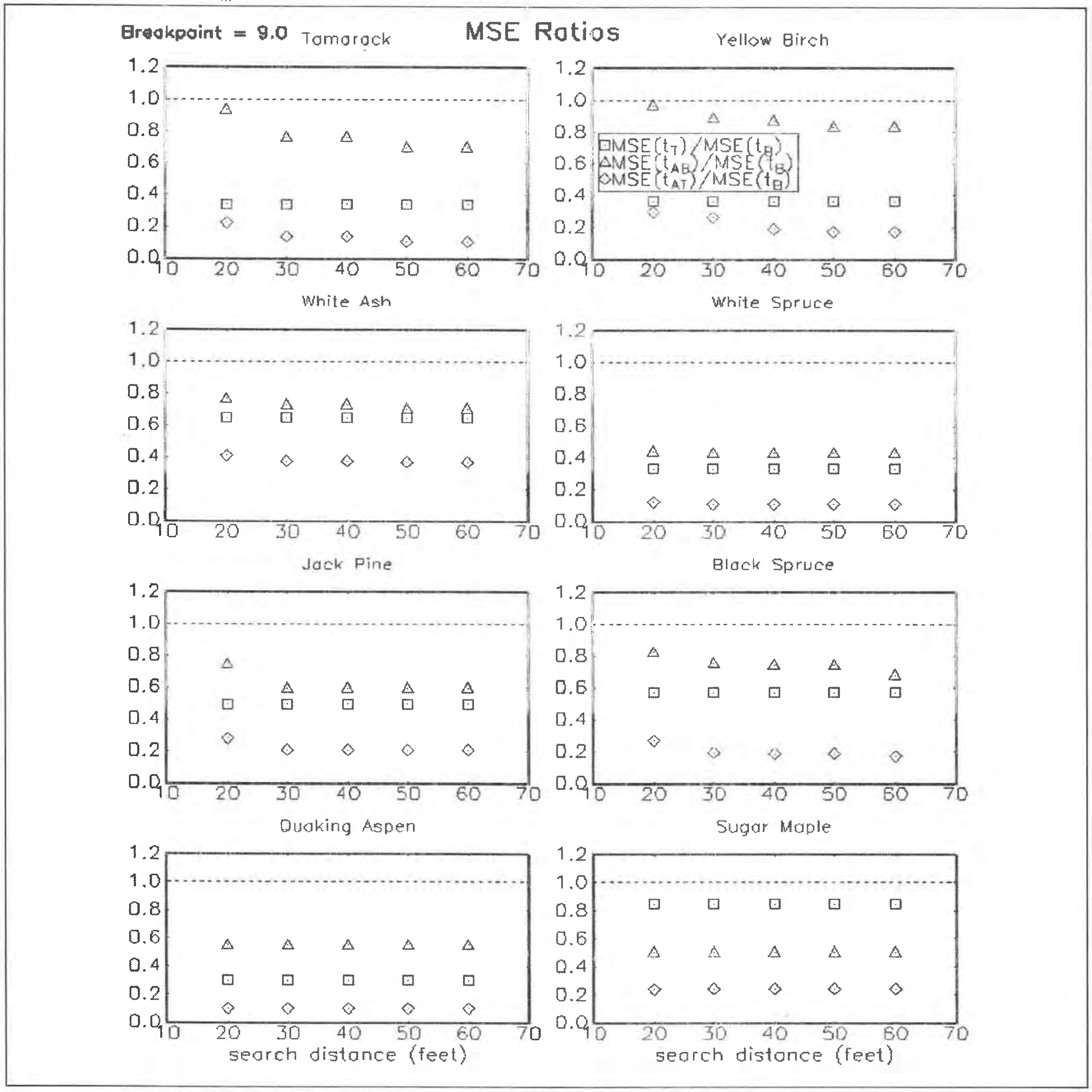
Figure 4 represents the heaviest investment in additional observations on the macroplots for designs 2 and 4 of those studied with a breakpoint diameter of 9 in. For four of the eight distributions (tamarack, yellow birch, white spruce, and quaking aspen), the reduction in MSE for the tri-areal design relative to the bi-areal design is greater than 60 percent; that is, the ratios are less than 40 percent. The tri-areal design has the least advantage over the bi-areal design in the case of the highly clumped sugar maple distribution, with MSE ratios greater than 80 percent. In all instances, the plots for the adapted bi-areal and adapted tri-areal designs show some advantage over their nonadapted counterparts. In all graphs but the sugar maple graph, the tri-areal design shows a greater reduction in MSE over the bi-areal design than does the adapted bi-areal design. The difference is very small in three of the graphs (white ash, white spruce, and jack pine) and fairly small in a fourth (black spruce). The adapted tri-areal in all cases shows the greatest overall reduction in MSE ratios. Note that in most cases a threshold can be discerned, beyond which an increase in search distance for the adapted designs yields little additional MSE reduction. With tamarack and jack pine for example, this appears to happen between search distances of 20 and 30 feet. With quaking aspen and sugar maple, this threshold appears to have occurred before the shortest distance simulated, 20 ft.

Figure 5 represents a smaller investment in additional observations on the macroplots for designs 2 and 4 than did figure 4 with an increased breakpoint diameter of 12 in. For six of the eight distributions, the ratio of tri-areal design MSE to the bi-areal design MSE exceeds the ratio of adapted bi-areal design MSE to MSE for the nonadapted bi-areal design. No advantage can be discerned for the tri-areal design over the bi-areal design for two species (jack pine and quaking aspen). The miniscule advantage noted for sugar maple could hardly be justified by the six-fold increase in plot size. For the remaining species, the tri-areal designs still show a significant advantage over their respective bi-areal counterparts. In all instances the adapted designs outperform their nonadapted counterparts.

The results in figure 6, for the breakpoint diameter of 15 in, show that the diameter distributions of five of the species are such that the tri-areal design provides no advantage. It is at this breakpoint diameter that an advantage of the adapted bi-areal over the unadapted tri-areal is first observed for white spruce.

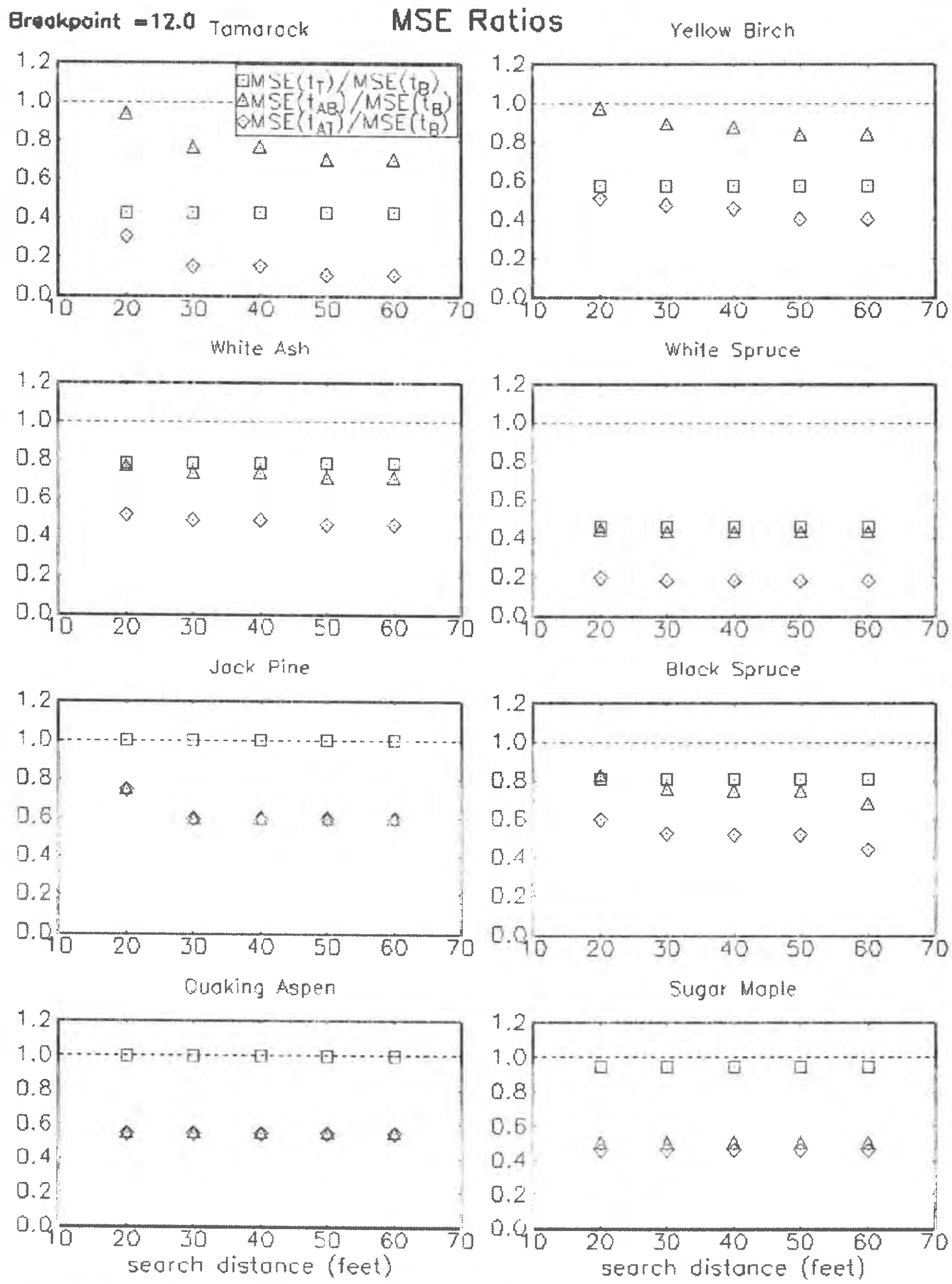
Figure 7 shows that none of the diameter distributions supports an argument for a breakpoint diameter of 18 in or larger.

Figure 4.—*Plots from 1,000 simulations for each species of three mean square error ratios using a breakpoint diameter of 9 in. The denominator in each case is the mean square error of the total basal area estimator from the bi-areal design (t_B). The numerators are (1) the mean square error of the total basal area estimator from the tri-areal design (t_T), (2) the mean square error of the total basal area estimator from the adapted bi-areal design (t_{AB}), and (3) the mean square error of the total basal area estimator from the adapted tri-areal design (t_{AT}).*



MSE = mean square error.

Figure 5.—*Plots from 1,000 simulations for each species of three mean square error ratios using a breakpoint diameter of 12 in. The denominator in each case is the mean square error of the total basal area estimator from the bi-areal design (t_B). The numerators are (1) the mean square error of the total basal area estimator from the tri-areal design (t_T), (2) the mean square error of the total basal area estimator from the adapted bi-areal design (t_{AB}), and (3) the mean square error of the total basal area estimator from the adapted tri-areal design (t_{AT}).*



MSE = mean square error.