# 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.

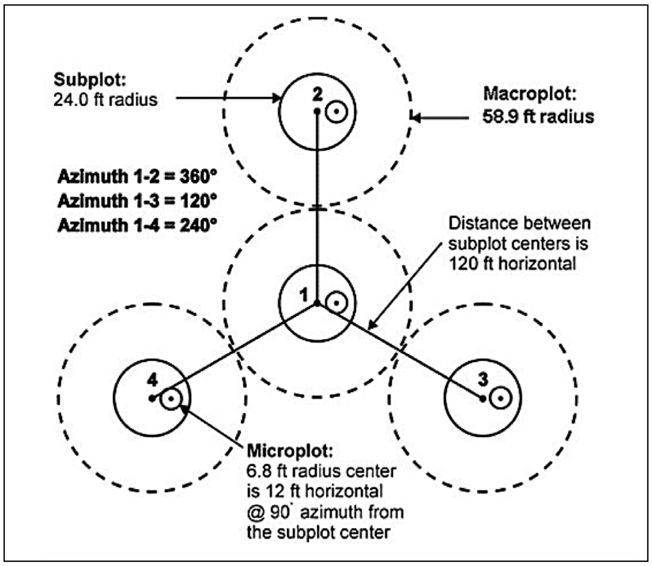
(2) Use adaptive cluster sampling with search circles of a fixed size dependent upon the expected intra-cluster distribution (as in Roesch 1993).

(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.

Figure 1.—*The FIA plot design. All trees greater than 5.0 in d.b.h. are measured on the subplot. Trees greater than 1.0 in d.b.h. are measured on the microplot. The macroplot is an optional feature currently used in the Pacific Northwest as an auxillary sample for large trees.*




*FIA = Forest Inventory and Analysis.*
*Source: Bechtold and Patterson 2005.*

That modification would be to extend the search area for the first member of the network (defined below) by a predetermined portion of the crew members' approach to the plot, thereby increasing the plot size exclusively for American chestnut. The extreme rareness of the American chestnut may warrant such an extended search area, and this extension of the search area would probably not substantially increase observation time because the crew is already traveling to the plot and American chestnut is distinctive enough to usually be recognized immediately. As long as the area of the extended search is known, unbiased estimators can be formed in the same way they are for the options we will consider in greater detail.

Alternatively, if one forsakes the desire for unbiased estimators in favor of any indicator of the presence of American chestnut, crew members could record any observation of the species during the course of their workday. The quality of the resulting information would be comparable to that obtained in almost all of the botanical studies conducted through the middle of the 20th century, and the information obtained could be used as a contemporary update to species distribution maps that were developed in the same way and are still relied on today.

The options considered here are described in detail below.

**Option 1**

Option 1 exploits an adaptable feature of the existing FIA design, first by employing the currently underutilized macroplot to sample this rare species and second by redefining the various plots within the tri-areal design for intensified observations on American chestnut. The macroplot, an optional feature in the FIA sample design, may be used to augment the sample for attributes of regional interest. This option has at least two advantages. First, it is somewhat efficient because increased selection areas could be limited to plots in areas with a high probability of containing the rare event of interest and to observations on the species of interest.

In addition, there would not be any further theoretical development or explanation needed for FIA practitioners and data users, other than a description of the larger selection areas for chestnut trees. The largest disadvantage is that additional costs of observation would be incurred on every plot in all areas of interest. Some potential also exists for field crew confusion with respect to the species-specific plot sizes and the identification of plots in high-probability areas.

**Option 2**

In option 2, the existing inventory is modified by adapting the field procedure when American chestnut is observed. Roesch (1993, 1994), following the work of Kalton and Anderson (1986), Levy (1977), Thompson (1990, 1991a, 1991b), and Wald (1947), showed how to do this for forest inventories using adaptive cluster sampling, in which unique networks of trees are sampled rather than unique trees.

Adaptive designs are usually described as being executed in two stages. First a probability sample of units in a population is taken and then additional units are selected near those units that display a specific condition of interest. Combining probability proportional to size sampling schemes common in forestry with an adaptive sampling scheme results in a system that can be applied to both equal and unequal probability forest inventory systems (Roesch 1993). In this article, the initial selection of trees by FIA's tri-areal design forms the basis of the first stage (Reams *et al.* 2005). In general, if a sample tree displays some rare condition of interest, (e.g., being an American chestnut), a specified area is examined for additional chestnut trees. This is repeated for every new chestnut found. The goal is to choose a distance rule that would identify a reasonable number of additional trees for the sample. Achievement of this goal is facilitated by a foreknowledge of the spatial distribution of the rare event of interest.

Further development assumes that the tree is the sampling unit and that there are N trees in the area of interest with labels 1,2,...,N. Associated with the N trees are values of interest $\mathbf{y} = \{y_1,y_2,...,y_N\}$ and characteristics of interest $\mathbf{C} = \{C_1,C_2,...,C_N\}$ (Roesch 1993). In this case, the species itself is the characteristic of interest; therefore, if tree *i* is an American chestnut tree, $C_i = 1$, and otherwise $C_i = 0$. This study was designed to determine an optimal adaptive sampling design and estimators for the presence, size, and fecundity of American chestnut within most of its natural range. As an example, suppose the variable of interest is total basal area of American chestnut trees. Let $x_i = C_i y_i$, so if tree *i* is an American chestnut tree, $x_i = C_i y_i = 1 * ba_i$.

The field crew would take the following steps:

(1) Conduct the initial sample.

(2) For all American chestnut trees conduct the adaptive part of the sample:

  (a) Measure all desired attributes $y_i$.

  (b) Observe all American chestnut trees that are within a circle of radius *r* from the center of tree *i* and have not already been sampled (i.e., ignore all new trees of other species).

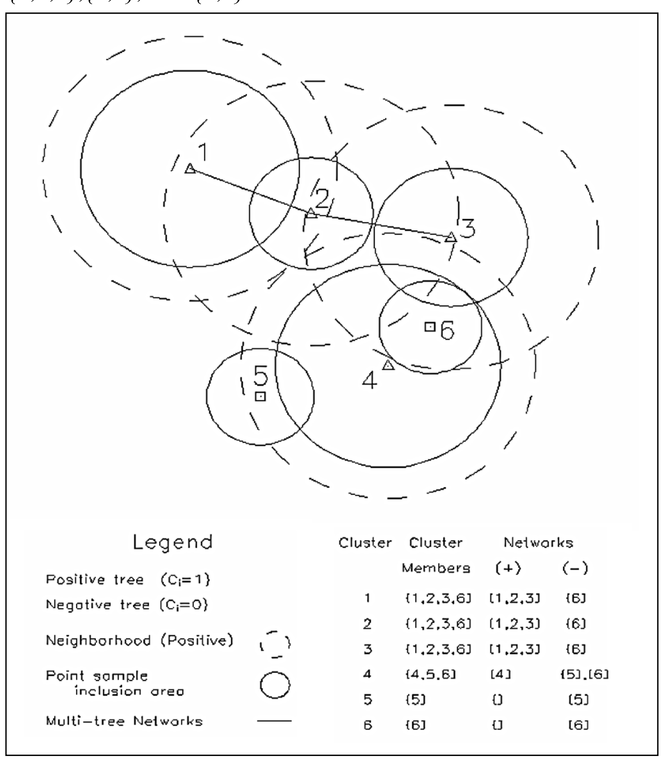
  (c) For all newly observed American chestnut trees, return to (a).

  (d) Stop when no new American chestnut trees are observed.

Similarly, the initial FIA design is extended to a network of chestnut trees. Each tree is surrounded by a circular area of selection. Options for definitions of the radius *r* are discussed below. These options attempt to determine a radius that would identify a reasonable number of additional trees for the sample—that is, enough additional trees to provide an estimation advantage and few enough to be considered during the measurement of the field inventory plot. This requires us to consider all of the available information on the spatial distribution of chestnut trees. It is clear that this extra effort in the design stage will be rewarded by the increased efficiency of a well-planned adaptive sampling survey.

A cluster is the set of all trees included in the sample as a result of the initial selection of tree *i*. A network is the subset of trees within a cluster 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. Because 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 *N* trees into a population of *M* networks, conditioned on *C* (fig. 2). Each network is sampled with known probability because the network population is mapped directly onto the tree population. We ignore trees not displaying the condition (i.e., for which $C_i = 0$) unless they are in the original estimators. This results in unbiased estimators (Thompson

Figure 2.—*Adaptive sampling attributes for a group of six trees in a population. A randomly placed point in the initial sample can select trees from the sets: {},{1},{2},{3},{4},{5},{1,2},{3,4},{3,4,6},{4,5}, and {4,6}.*




*Source: Roesch 1993.*

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:

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

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. ≥ = 5.0 in

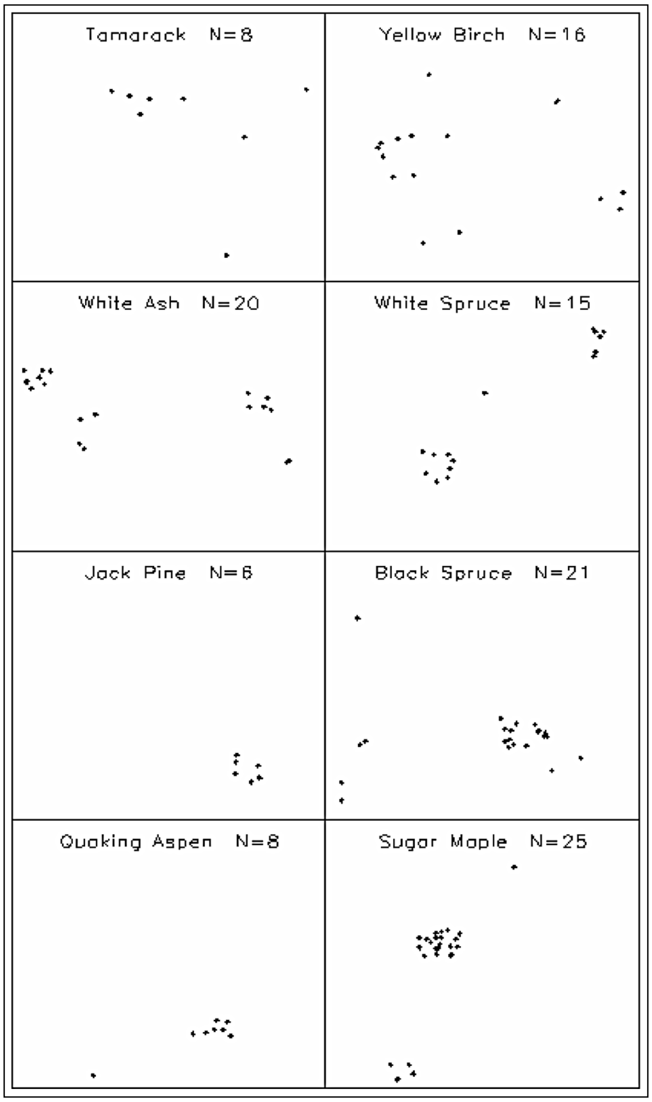
(2) Tri-areal design—breakpoint diameter (9, 12, 15, and 18 in).
  Microplot: d.b.h. < 5.0 in
  Subplot: 5.0 in ≤ = d.b.h. < breakpoint diameter
  Macroplot: d.b.h. ≥ = breakpoint diameter

(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.

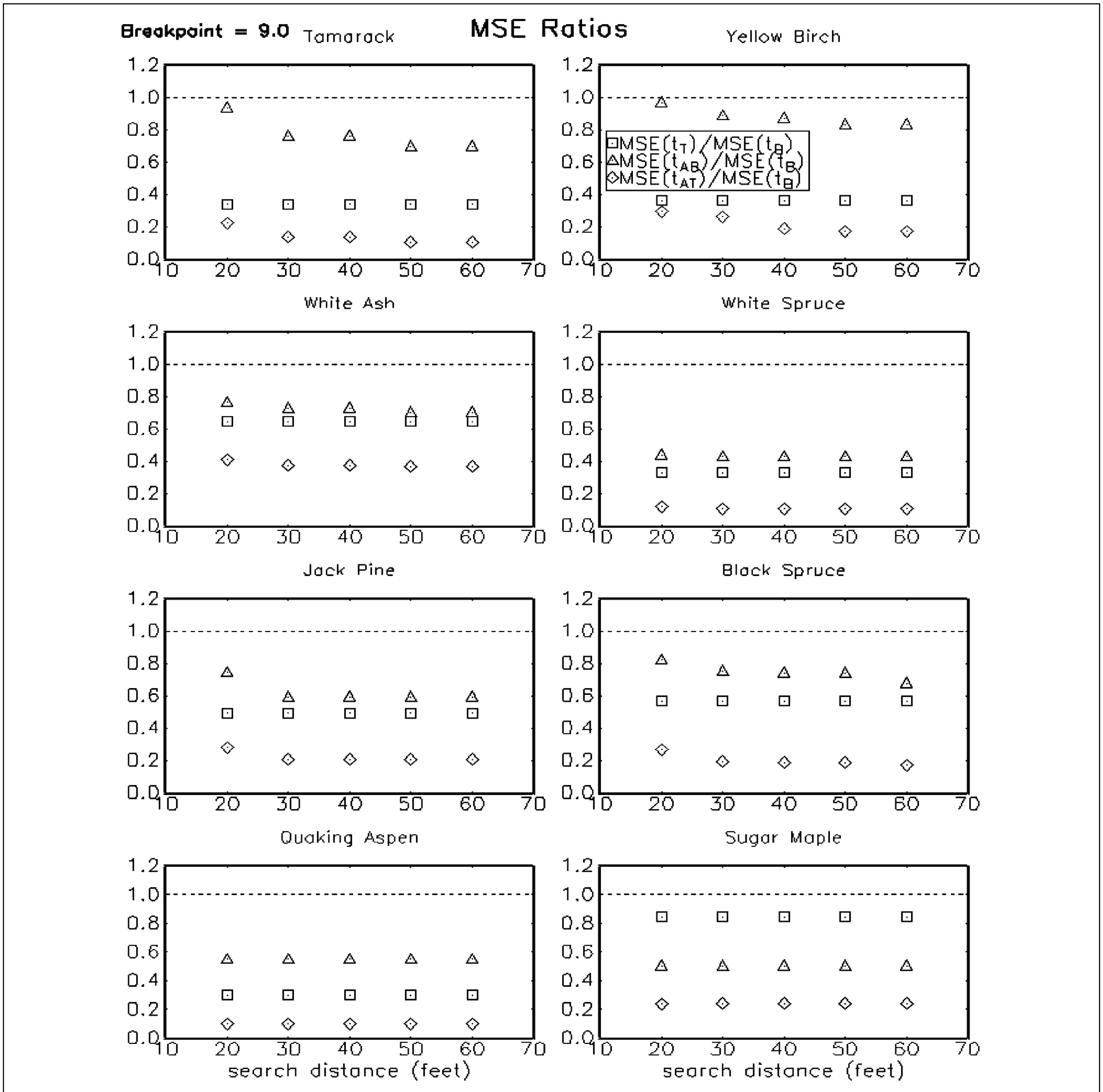
Figure 4 represents the heaviest investment in additional observations on the macroplots for designs 2 and 4 of those studied with a breakpoint diameter of 9 in. For four of the eight distributions (tamarack, yellow birch, white spruce, and quaking aspen), the reduction in MSE for the tri-areal design 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.

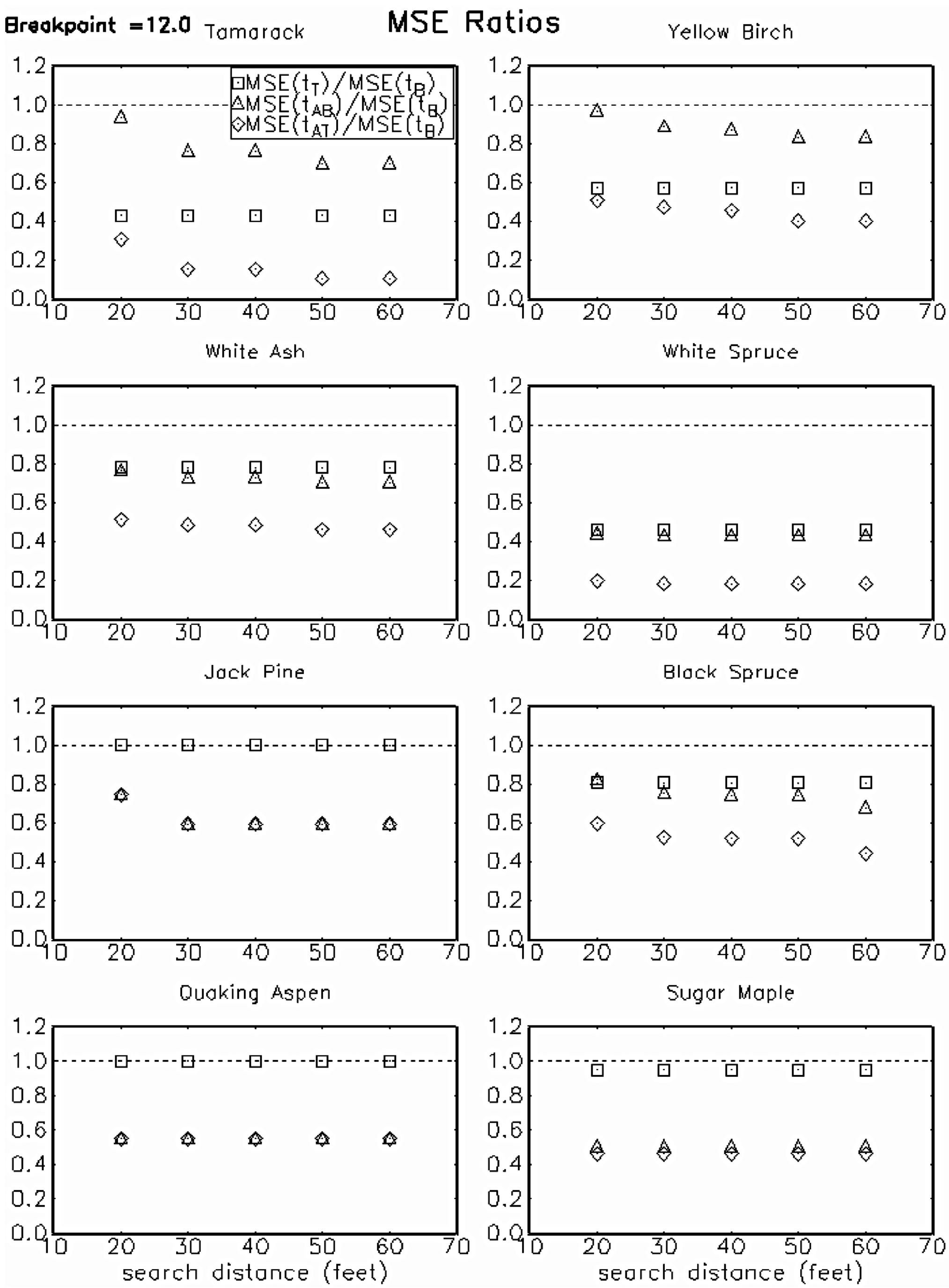
Figure 7 shows that none of the diameter distributions supports an argument for a breakpoint diameter of 18 in or larger.

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 ($t_{AT}$).*



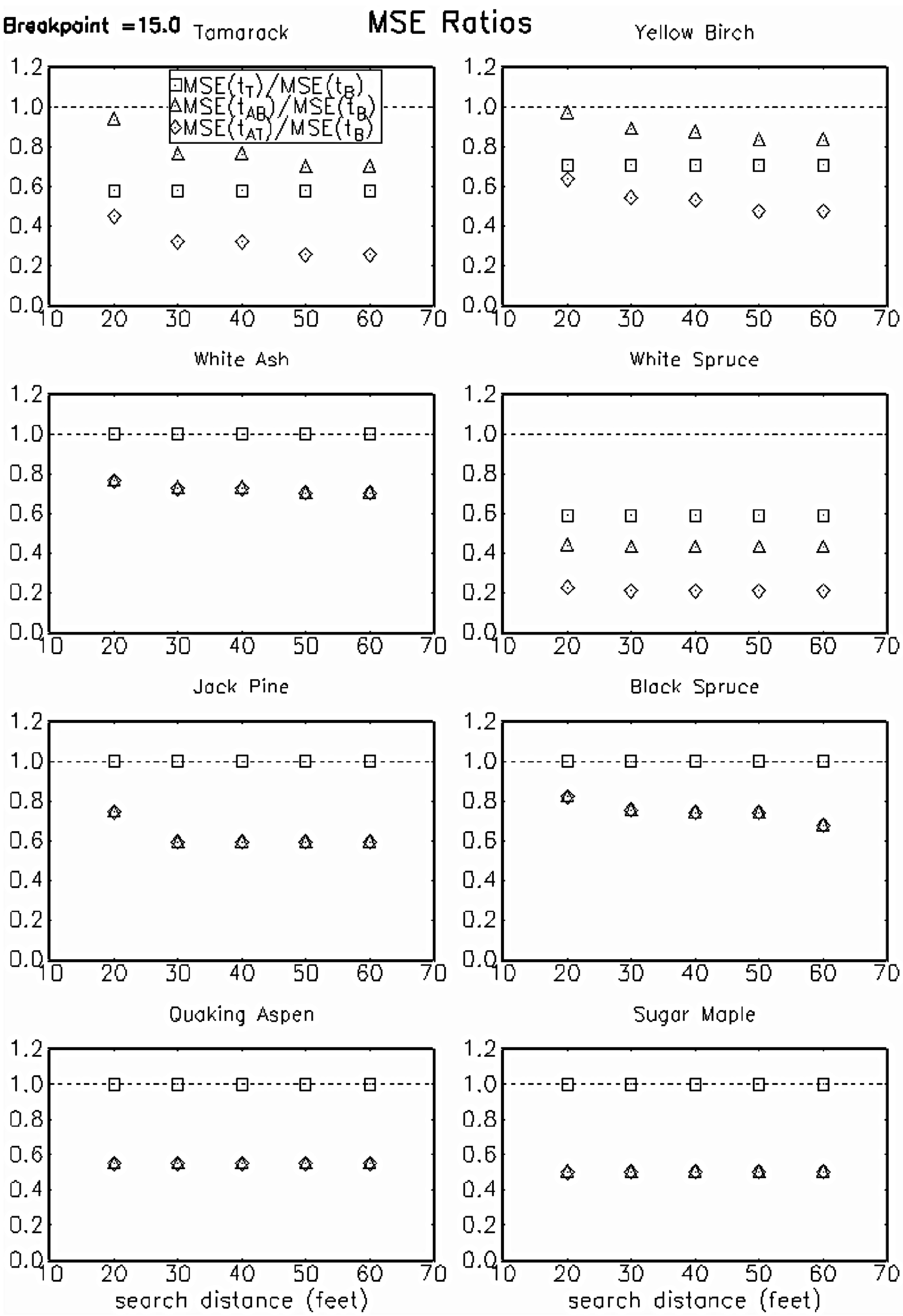

*MSE = mean square error.*

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 ($t_{AT}$).*



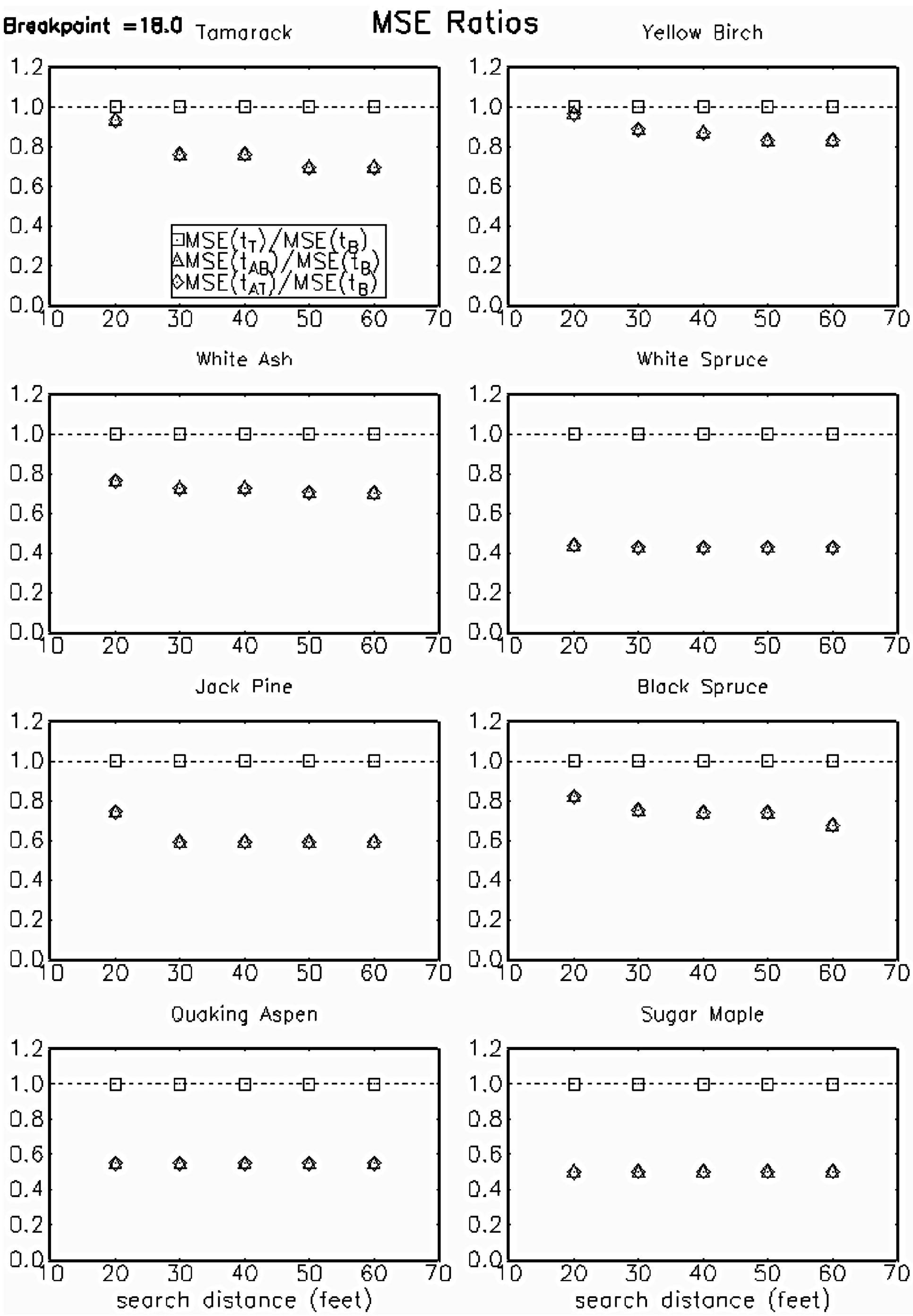

*MSE = mean square error.*

Figure 6.—*Plots from 1,000 simulations for each species of three mean square error ratios using a breakpoint diameter of 15 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 7.—*Plots from 1,000 simulations for each species of three mean square error ratios using a breakpoint diameter of 18 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.*

## Discussion

Estimation of American chestnut attributes is ideal as a test case of adaptive sampling for FIA because the species is so rare, in stark contrast to its previous abundance, and because there is intense scientific interest in the species. A significant advantage of the FIA design for estimation of well-dispersed forest attributes is the intentionally thorough dispersion of the sample plots through the spatial-temporal cube, while this very aspect constitutes a significant weakness for estimation of forest attributes that occur in rare clusters within the cube. An adaptive sampling design laid on top of the FIA design in targeted areas could nullify this weakness of the existing design for the estimation of this rare event.

Relative cost is a major concern when choosing between sampling strategies. The additional monetary cost of the adaptive strategy for a particular application depends on relative cluster size and occurrence in the sample. These factors can be predicted given adequate previous knowledge of the population. In the example in Roesch (1994), the additional cost of including extra trees was shown to be controllable by the distance examined. Within this distance, the species of each tree must be determined and if any tree is an American chestnut, its d.b.h., and location must be recorded. Because American chestnut trees are truly rare and found in clusters, the additional cost will be small and the estimate of the per-tree attributes will be improved. The size of the search area determines the size of the networks of interest found as well as the number of additional other trees encountered. Therefore, for any other specific attribute one would want to select a minimally sized search area dependent on the expected proximity of the target trees to each other.

Adaptive cluster sampling provides a way for FIA to monitor rare events at a relatively small cost. It also 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 of those adapted for.

This direct application of adaptive sampling should yield much-improved estimates of American chestnut attributes; however, it appears that this method could be used even more efficiently for American chestnut with a minor modification. That is, the extreme rareness of the American chestnut may warrant an extended search area for the first member of the network. This is a viable option not discussed in the Roesch or Thompson papers cited. The search area would be extended by a defined portion of the crew members' approach to the plot. This would increase the plot size for American chestnut only, without substantially increasing observation time. That is, the crew is already traveling to the plot, and they generally look around while they are doing that. American chestnut is so distinctive that it is usually recognized immediately. At any rate, the relative advantage of the extended plot may be evaluated in a future study.

## Literature Cited

Bechtold, W.A.; Patterson, P.L., eds. 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Horvitz, D.G.; Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. JASA. 47: 663-685.

Kalton, G.; Anderson, D.W. 1986. Sampling rare populations. Journal of the Royal Statistical Society, Series A (General). 149: 65-82.

Levy, P.S. 1977. Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. JASA. 72: 758-763.

McWilliams, W.H.; Lister, T.W.; LaPoint, E.B.; Rose, A.K.; Vissage, J.S. 2006. Current status of chestnut in eastern U.S. forests. Natural Resources Report NPS/NCR/CUE/NRR 2006/001. In: Steiner, K.C.; Carlson, J.E., eds. Restoration of American chestnut to forest lands. Proceedings of a conference and workshop. Washington, DC: U.S. Department of the Interior, National Park Service: 31-39.

Reams, G.A.; Smith, W.D.; Hansen, M.H.; Bechtold, W.A.; Roesch, F.A.; Moisen, G.G. 2005. The Forest Inventory and Analysis sampling frame. In: Bechtold, W.A.; Patterson, P.L., eds. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Roesch, F.A. 1994. Incorporating estimates of rare clustered events into forest inventories. Journal of Forestry. 92(11): 31-34.

Roesch, F.A., Jr. 1993. Adaptive cluster sampling for forest inventories. Forest Science. 39(4): 655-669.

Thompson, S.K. 1990. Adaptive cluster sampling. JASA. 85(412): 1050-1059.

Thompson, S.K. 1991a. Stratified adaptive cluster sampling. Biometrika. 78: 389-397.

Thompson, S.K. 1991b. Adaptive cluster sampling: designs with primary and secondary units. Biometrics. 47: 1103-1115.

Wald, A. 1947. Sequential analysis. New York: John Wiley. 212 p.