Statistical Properties of Alternative National Forest Inventory Area Estimators

Francis A. Roesch, John W. Coulston, and Andrew D. Hill

Abstract: The statistical properties of potential estimators of forest area for the USDA Forest Service's Forest Inventory and Analysis (FIA) program are presented and discussed. The current FIA area estimator is compared and contrasted with a weighted mean estimator and an estimator based on the Polya posterior, in the presence of nonresponse. Estimator optimality is evaluated both theoretically and via simulation under bias and mean squared error criteria. The results indicate that, under realistic conditions, the current FIA area estimator can sometimes result in substantial bias and have a higher mean squared error than both of the alternative estimators. This finding is of special interest because the same factor that contributes to this increased bias and variance applies to all area-based FIA estimates. The weighted mean and Polya posterior estimators gave similar results for estimating the total area of a domain. It is concluded that the main advantage of the latter approach is that many other statistics are obtainable because the entire population distribution is estimated from the same sampling effort. The cost of this advantage for the Polya posterior approach is that a single result requires many more computer operations, a cost that has become virtually ignorable over the past decade. FOR. SCI. 58(6): 559–566.

Keywords: nonresponse, Polya posterior

THE MOST BASIC and sometimes the most problematic variable that national forest inventories such as the US Forest Service's Forest Inventory and Analysis (FIA) program attempt to measure is the land area of a particular condition or domain, such as forest area within a limited geographical area. Note that it is common in multiresource inventories to have multiple populations of interest because different attributes require different realizations of the population. Although the areal dimension of the population of interest for FIA is the land area within the boundaries of the United States, this article is concerned with estimation within smaller subdivisions (or domains) of the overall area. In particular, we explore the estimation of the area of a domain at a fixed point in time. At present, the FIA program uses an internally developed, stratum-based estimator for all area estimation. This estimator is based on Equations 4.1, 4.2, 4.3, and 4.5 of Scott et al. (2005). Here we show that the estimator, as defined, is design biased. We show how the bias is introduced into this area estimator and then compare and contrast that estimator with a weighted mean estimator in the presence of nonresponse. We then evaluate the optimality of each estimator with respect to a mean squared error (MSE) criterion.

An alternative presentation of Equation 4.1 in Scott et al. (2005) is

$$P_i^{hd} = \frac{a_i^{do}}{\bar{\mathbf{a}}_h^o}, \tag{1}$$

where P_i^{hd} is the ratio of the area observed to be in domain d on plot i to the average plot area within stratum h (note that P_i^{hd} will often be greater than 1 because of plots that were partially unobserved), a_i^{do} is area (ac) of plot i observed to be in domain d, and $\bar{\mathbf{a}}_h^o$ is average observed plot area in stratum h (within the population, see Equation 2). That is,

$$\bar{\mathbf{a}}_h^o = \frac{1}{n_h}\sum_{i=1}^{n_h} a_i^o, \tag{2}$$

where n_h is the number of ground plots with phase 1 assignments to stratum h (for total area, this includes all plots with any portion in the population; for subsequent estimates, any entirely nonsampled plot is excluded) and a_i^o is the observed area (ac) of plot i (in the population). The reader should note that in the equations above and in the following, a nonnumeric character in the superscript position will be used to indicate context-specific class membership rather than to indicate exponentiation. For instance, an "o" in the superscript indicates membership in the observed group, whereas a "d" in the superscript position indicates membership in the domain d. For the reader's convenience, the most often repeated notational elements in this article are collected in the Appendix.

By this notation, Equation 4.3 in Scott et al. (2005) was

$$\overline{P_h^d} = \frac{\sum_i^{n_h} P_i^{hd}}{n_h},$$

and the astute reader will notice that three equations could

Manuscript received January 14, 2011; accepted July 11, 2011; published online February 2, 2012; http://dx.doi.org/10.5849/forsci.11-008.

Francis A. Roesch, USDA Forest Service, Southern Research Station, 200 WT Weaver Blvd., Asheville, NC 28804-3454—Phone: (828) 257-4871; froesch@fs.fed.us; John W. Coulston, USDA Forest Service—jcoulston@fs.fed.us. Andrew D. Hill, USDA Forest Service—adhill@fs.fed.us.

Acknowledgments: This work was funded by the USDA Forest Service Forest Inventory and Analysis Program. The authors thank the Associate Editor and two anonymous reviewers for their helpful comments.

have been combined to obtain the proportion of stratum h plot area observed to be in domain d, and simply written as

$$\bar{P}_h^{do} = \frac{\sum_{i=1}^{n_h} a_i^{do}}{\sum_{i=1}^{n_h} a_i^{o}}. \quad (3)$$

The estimator for the total area of the domain can be written as

$$\hat{A}_d = \sum_{h=1}^{H} \bar{P}_h^{do} A_h, \quad (4)$$

where A_h is the area of stratum h and H is the number of strata.

In the presence of nonresponse, Equation 3 is a ratio estimator (because the a_i^o values are estimates and therefore cannot be assumed to be known). Ratio estimators have long been known to be design biased, except under special conditions, (e.g., Pathak 1964, Cochran 1963, Chapter 6, Rao 1966, 1977, Srivenkataramana 1980). In the case at hand, these special conditions would be met if $a_i^{do} = P_h^d a_i^o + \varepsilon_i$, where P_h^d is the true proportion of stratum h area in domain d and the ε_i are independent with an expectation of zero. This condition would seem tenable when there cannot be any nonresponse. Otherwise, when one is willing to assume that this condition is true, this design-biased estimator would be model unbiased (or conditionally unbiased). The bias of $\hat{A}_d$ will be a direct result of the bias in $\bar{P}_h^{do}$. Exactly how much bias will be present is application dependent.

As in most surveys, portions of the intended sample are not observed for a variety of reasons. The most obvious reasons are the following: first, denial of access to the location by a landowner; second, the presence of a hazard in the vicinity of the location; and third, insufficient allocation of human resources. Some of these reasons are truly reflective of the population. For the first reason, for instance, we can assume that there is some group of landowners that, if given the choice, would always deny access to land. In addition, for the second reason, certain hazards would always be avoided, and therefore land in the vicinity of those hazards is not observable. This effective reduction in the sampled population relative to the target population is ignored in many public surveys. Here we assume that all failure to observe is a result of a partitioning of the target population by an interaction of (potentially random) factors and the sample design. In areas where this assumption is highly untenable, it would be preferable to sever the area from the target population.

Assume that we have divided (uniquely and completely) the entire land area into a number of strata. Further assume that our sampling mechanism partitions each stratum into areas of specific sampling intensity (substrata). In general, the proportion of stratum h in domain d can be found by multiplying the proportion of each partition in domain d by the proportion of the stratum covered by the partition and summing the results over all partitions, i.e.,

$$P_h^d = \sum_{j=1}^{J_h} P_{j|h} \Pi_{d|j},$$

where $P_{j|h}$ is the proportion of stratum h that is partition j, $\Pi_{d|j}$ is the proportion of partition j in domain d, and J_h is the number of partitions in stratum h.

Likewise, the proportion of the domain over the entire area can also be calculated using a weighted mean

$$P_d = \sum_{h=1}^{H} P_h^d \Pi_h = \sum_{h=1}^{H} \left[\Pi_h \left(\sum_{j=1}^{J_h} P_{j|h} \Pi_{d|j} \right) \right],$$

where P_h^d is the proportion of stratum h in domain d, and Π_h is the proportion of the area in stratum h. Of the quantities above, Π_h is known, whereas $P_{j|h}$ may or may not be known and $\Pi_{d|j}$ must be estimated from the sample:

$$\hat{P}_d = \sum_{h=1}^{H} \hat{P}_h^d \Pi_h = \sum_{h=1}^{H} \left[\Pi_h \left(\sum_{j=1}^{J_h} \hat{P}_{j|h} \hat{\Pi}_{d|j} \right) \right],$$

where $\hat{P}_h^d$ is the sample estimate of the proportion of stratum h in domain d, $\hat{P}_{j|h}$ is the sample estimate of the proportion of stratum h that is partition j, and $\hat{\Pi}_{d|j}$ is the sample estimate of the proportion of partition j in domain d.

If both $\hat{P}_{j|h}$ and $\hat{\Pi}_{d|j}$ are unbiased and have zero covariance or $\hat{P}_h^d$ can be unbiasedly estimated directly, then $\hat{P}_d$ will be unbiased.

As discussed above, the approach of Scott et al. (2005) calculates the average sample intensity across the stratum, regardless of the actual substrata inclusion probabilities with a ratio estimator, $\bar{P}_h^{do}$, and the estimator for the area of the domain $\hat{A}_d$.

The expected value of $\hat{A}_d$:

$$\begin{aligned} E[\hat{A}_d] &= E\left[\sum_{h=1}^{H} \bar{P}_h^{do} A_h \right] \\ &= \sum_{h=1}^{H} A_h E[\bar{P}_h^{do}] \\ &= \sum_{h=1}^{H} A_h E\left[\frac{\sum_{i=1}^{n_h} a_i^{do}}{\sum_{i=1}^{n_h} a_i^{o}} \right] \\ &= \sum_{h=1}^{H} A_h \left[E\left[\frac{1}{\sum_{i=1}^{n_h} a_i^{o}} \right] E\left[\sum_{i=1}^{n_h} a_i^{do} \right] + \operatorname{cov}\left[\frac{1}{\sum_{i=1}^{n_h} a_i^{o}}, \sum_{i=1}^{n_h} a_i^{do} \right] \right], \end{aligned}$$

The covariance term may not be safely ignorable because both terms are affected by nonresponse. Regardless of how the nonresponse arises, we define $a_i^o = I_i a_i$ and $a_i^{do} = I_i^d a_i^d$,

where a_i and a_i^d are the areas of plot i and domain d on plot i, which would be observed, respectively, whereas I_i and I_i^d are indicator variables equal to 1 if plot i and domain d on plot i, respectively, are observable and 0 otherwise.

In the event of 100% response only, in the case at hand, the realizations of all indicator variables are equal to 1, and the expected value becomes the expected value of a proportion and the estimator can be shown to be unbiased. Otherwise set

$$A_h^{do} = \sum_{j=1}^{J_h} \sum_{i=1}^{m_j} I_i^d a_i^d, \text{ and } A_h^o = \sum_{j=1}^{J_h} \sum_{i=1}^{m_j} I_i a_i,$$

where m_j is the number of plots in partition j, J_h is the number of partitions in stratum h, A_h^{do} is the observed area in domain d in stratum h, and A_h^o is the observed area in stratum h. Then

$$E[\hat{A}_d] = \sum_{h=1}^{H} A_h E\left[\frac{A_h^{do}}{A_h^o}\right]. \quad (5)$$

Intuitively, examination of Equation 5 suggests that if nonresponse is random, the bracketed proportion should deviate little from the desired proportion, and, therefore, the bias in $\hat{A}_d$ should be small. Furthermore, Equation 5 can be expanded to

$$E[\hat{A}_d] = \sum_{h=1}^{H} A_h \left[E\left[\frac{1}{A_h^o}\right] E[A_h^{do}] + \operatorname{cov}\left[\frac{1}{A_h^o}, A_h^{do}\right]\right]. \quad (6)$$

If the missing values occur completely at random within strata, the terms $1/A_h^o$ and A_h^{do} in Equation 6 are independent, and the covariance term is equal to 0, resulting in a bias that will be equal to 0. Otherwise, the expected value can be estimated via approximation by Taylor series expansion (Mood et al., 1974, p. 181):

$$E[\hat{A}_d] \approx \sum_{h=1}^{H} A_h \left[\frac{E(A_h^{do})}{E(A_h^o)} - \frac{\operatorname{cov}(A_h^{do}, A_h^o)}{[E(A_h^o)]^2} + \frac{E(A_h^{do})\operatorname{var}(A_h^o)}{[E(A_h^o)]^3}\right]. \quad (7)$$

For completeness, Mood et al. (1974, p. 181, Theorem 4) also gives the corresponding variance estimator, which in our case is the within-stratum variance estimator

$$\operatorname{var}[\hat{A}_h^d] \approx A_{dh}\left[\frac{E(A_h^{do})}{E(A_h^o)}\right]^2 \cdot \left[\frac{\operatorname{var}(A_h^{do})}{[E(A_h^{do})]^2} + \frac{\operatorname{var}(A_h^o)}{[E(A_h^o)]^2} - \frac{2\operatorname{cov}(A_h^{do}, A_h^o)}{[E(A_h^{do})E(A_h^o)]}\right] \quad (8)$$

If the number of plots in the population was finite, we could set

$$\sum_d = \sum_{h=1}^{H} \frac{1}{A_h} \sum_{k=1}^{N_h} a_k I_{kd} a_k^d, \quad \text{and} \quad \sum_a = \sum_{h=1}^{H} \frac{1}{A_h} \sum_{k=1}^{N_h} a_k I_k a_k,$$

with the index k going from 1 to the total number of plots in the stratum (N_h). The approximation to the expected value in 7 would then result in

$$E[\hat{A}_d] \approx \sum_{h=1}^{H} A_h \left[\left(\frac{\sum_d}{\sum_a}\right) - \frac{[A_h^{do} - \sum_d][A_h^o - \sum_a]}{[\sum_a]^2} + \frac{(\sum_d)[A_h^o - \sum_a]^2}{[\sum_a]^3}\right]. \quad (9)$$

Note that in Equation 9, we have an additional approximation over and above what is discussed in Mood et al. (1974). The approximation concerns the expression N_h in the definitions of $\sum_d$ and $\sum_a$. Of course, obtaining the number of population elements requires a finite population, whereas the FIA sample of land area is actually drawn from an infinite population. Because the sample is very sparse, there is probably no practical consequence to this further approximation, but to make it without acknowledgment would be a mistake, as pointed out by Williams (2001). Although it is beyond the scope of this article, many infinite populations of forest attributes can be mapped over the land area (and even through time) into a finite population (Roesch et al. 1993, Roesch 2008). In this case, we note that there are a finite number of segments in the population and each segment has associated with it the domain attribute $d = 1$ or $d = 0$, corresponding to the segment being in the domain or not, respectively. The infinite population of possible points over the land area is thus mapped into a finite population of segments. Each segment is sampled with a probability proportional to its size.

Alternatively, there is a finite (but possibly very large) number N_g of starting points for a randomly placed grid. Each grid yields the observation

$$\bar{P}_{hdg} = \frac{\sum_{i=1}^{n_h} a_i^{do}}{\sum_{i=1}^{n_h} a_i^o}$$

with probability equal to

$$\frac{\sum_{i=1}^{n_h} a_i}{A_h}.$$

The probability proportional to size estimator for the area of stratum h that is domain d is

$$P_{hd}^{ppz} = \left(\sum_{i=1}^{n_h} a_i/A_h\right)\left[\left(\sum_{i=1}^{n_h} a_i^{do} \Big/ \sum_{i=1}^{n_h} a_i^o\right) \Big/ \left(\sum_{i=1}^{n_h} a_i/A_h\right)\right] = \left(\sum_{i=1}^{n_h} a_i^{do} \Big/ \sum_{i=1}^{n_h} a_i^o\right),$$

the ratio of means estimator with variance is

$$V(P_{hd}^{pprz}) = (1/N_g)\sum_{g=1}^{N_g}\left(\sum_{i=1}^{n_{hg}} a_i/A_h\right) \cdot\left(\left(\left(\sum_{i=1}^{n_{hg}} a_i^{do}\Big/\sum_{i=1}^{n_{hg}} a_i^{o}\right)\Big/\left(\sum_{i=1}^{n_{hg}} a_i/A_h\right)\right) - P_h^d\right)^2.$$

Approximate Bias, Variance, and MSE of $\hat{\mathrm{A}}_d$

From Equation 9, we see that the bias, variance, and MSE of $\hat{A}_d$ are approximately

$$\text{bias}\,[\hat{A}_d] = E[A_d - \hat{A}_d] \approx A_d - \sum_{h=1}^{H} A_h\left[\left(\frac{\sum_d}{\sum_a}\right) - \frac{[A_h^{do} - \sum_d][A_h^{o} - \sum_a]}{[\sum_a]^2} + \frac{(\sum_d)[A_h^{o} - \sum_a]^2}{[\sum_a]^3}\right],$$

$$\text{var}[\hat{A}_d] = [\hat{A}_d - E[\hat{A}_d]]^2 \approx \left[\sum_{h=1}^{H}\bar{P}_h^d A_h - \left(\sum_{h=1}^{H} A_h\left[\left(\frac{\sum_d}{\sum_a}\right) - \frac{[A_h^{do} - \sum_d][A_h^{o} - \sum_a]}{[\sum_a]^2} + \frac{(\sum_d)[A_h^{o} - \sum_a]^2}{[\sum_a]^3}\right]\right)\right]^2,$$

$$\text{MSE}[\hat{A}_d] = \text{var}[\hat{A}_d] + [\text{bias}[\hat{A}_d]]^2 = [\hat{A}_d - E[\hat{A}_d]]^2 + [E[A_d - \hat{A}_d]]^2$$

$$\approx \left[\sum_{h=1}^{H}\bar{P}_h^d A_h - \left(\sum_{h=1}^{H} A_h\left[\left(\frac{\sum_d}{\sum_a}\right) - \frac{[A_h^{do} - \sum_d][A_h^{o} - \sum_a]}{[\sum_a]^2} + \frac{(\sum_d)[A_h^{o} - \sum_a]^2}{[\sum_a]^3}\right]\right)\right]^2 + \left[A_d - \sum_{h=1}^{H} A_h\left[\left(\frac{\sum_d}{\sum_a}\right) - \frac{[A_h^{do} - \sum_d][A_h^{o} - \sum_a]}{[\sum_a]^2} + \frac{(\sum_d)[A_h^{o} - \sum_a]^2}{[\sum_a]^3}\right]\right]^2,$$

respectively.

Missing at Random

The estimator $\hat{A}_d$ therefore appears to be unbiased when observations are truly missing at random within strata. On the other hand, Patterson et al. (2011) present evidence strongly indicating that observations are often not missing at random within FIA strata but might be missing at random within the identifiable partitions discussed here. In that case, $\hat{A}_d$ would be biased proportional to the covariance term in Equation 6.

Weighted Mean Estimator for the Area of Domain d

We contrast $\hat{A}_d$ with a weighted mean estimator

$$\bar{A}_d = \sum_{h=1}^{H} A_h \sum_{j=1}^{J_h}\frac{A_j}{A_h}\frac{\sum_{i=1}^{m_j} a_i^{do}}{\sum_{i=1}^{m_j} a_i^{o}},$$

which differs from $\hat{A}_d$ only in its recognition of the partitions creating the substratification. In addition, we note that the weighted mean estimator is unbiased under more general conditions, i.e.,

$$E[\bar{A}_d] = E\left[\sum_{h=1}^{H} A_h \sum_{j=1}^{J_h}\frac{A_j}{A_h}\frac{\sum_{i=1}^{m_j} a_i^{do}}{\sum_{i=1}^{m_j} a_i^{o}}\right].$$

Set

$$S_j^{do} = \sum_{i=1}^{m_j} I_i^d a_i^d \quad \text{and} \quad S_j^{o} = \sum_{i=1}^{m_j} I_i a_i,$$

then

$$E[\bar{A}_d] = \sum_{h=1}^{H}\sum_{j=1}^{J_h} A_j E\left[\frac{S_j^{do}}{S_j^{o}}\right].$$

Therefore,

$$E[\bar{A}_d] = \left[\sum_{h=1}^{H}\sum_{j=1}^{J_h} A_j\left[E\left[\frac{1}{S_j^{o}}\right]E[S_j^{do}] + \text{cov}\left[\frac{1}{S_j^{o}}, S_j^{do}\right]\right]\right]. \tag{10}$$

If the missing values occur completely at random within the partitions (within substrata), the terms $1/S_j^o$ and S_j^{do} are independent and the covariance term is equal to 0, resulting in a bias that will equal 0. That is, $\bar{A}_d$ is unbiased under this more robust condition because the missing at random assumption needs only to be applied within the identifiable partitions, and we can assume that the covariance term is 0 in Equation 10. An example of a potential partitioning that is almost always identifiable in the United States is one based on ownership category.

Variance of $\bar{\mathrm{A}}_d$

$$\text{var}[\bar{A}_d] = [\bar{A}_d - E[\bar{A}_d]]^2 = \left[\sum_{h=1}^{H}\sum_{j=1}^{J_h} A_j\frac{S_j^{do}}{S_j^{o}} - A_d\right]^2.$$

MSE of $\bar{\mathrm{A}}_{\mathrm{d}}$

$$\text{MSE}[\bar{A}_d] = \text{var}[\bar{A}_d] + [\text{bias}[\bar{A}_d]]^2 = \text{var}[\bar{A}_d].$$

Because bias is generally considered the worst of the evils

in multiresource monitoring efforts, we certainly should prefer to use the estimator $\bar{A}_d$ over the estimator $\hat{A}_d$ when $\text{MSE}[\bar{A}_d] \leq \text{MSE}[\hat{A}_d]$. Of course, unknown events occurring in a sampled population can impart bias to any estimator. The reason $\hat{A}_d$ is characterized as a biased estimator in this work whereas $\bar{A}_d$ is not is that $\hat{A}_d$ is defined to ignore potentially known events leading to nonresponse within strata.

Domain Estimator Based on the Polya Posterior

All estimation efforts, either explicitly or implicitly, involve a model of how the observed relates to the unobserved. Suppose that we did observe all of what we intended to observe in our sample and we did not have a supplemental sample to "make up for" the missing observations? The frequentist literature is replete with discussions concerning comparisons to what would have been observed at the "missing data" locations. It is rare outside of the Bayesian literature to find an acknowledgment that intended but unobserved sample elements should play absolutely no role in inference over and above the role of elements that were never intended to be in the sample. That is, for inference purposes, a finite population of size N is partitioned into $s = o$ "observed" elements and $s' = (m + n')$ "unobserved" elements. The utility of this acknowledgment became obvious subsequent to Rubin's (1987) work in multiple imputation (Ghosh and Meeden 1997). Multiple imputation was devised to allow the use of complete data methods for incomplete data through the creation of a set of complete pseudosamples by modeling the relationship of the unobserved to observed sample elements. In contrast, the Polya posterior provides a mechanism to create a set of pseudopopulations using a model of the relationship of unobserved population elements to observed population elements. Surprisingly, little use of the Polya posterior method is found in the forestry literature, with notable exceptions being Magnussen and Köhl (2002), Magnussen et al. (2004), and Magnussen et al. (2010).

The Polya posterior is based on the Polya urn model. In it, we assume that we have an urn containing s_i balls of color b_i, $i = 1, \ldots, k$ and $s = s_1 + \cdots + s_k$. In keeping with the notation above, let $N = s + s'$. Draw a ball at random from the urn and observe its color. Replace it and place one more ball of the same color into the urn. Do this s' times. The probability of getting r_i balls of color b_i in a specified order is

$$\left\{\prod_{i=1}^{k}[\Gamma(s_i + r_i)/\Gamma(s_i)]\right\}/\{\Gamma(s + s')/\Gamma(s)\}, \quad \text{where } s' = \sum_{i=1}^{k} r_i.$$

The limiting distribution of the vector of proportions of each color of ball is Dirichlet $(s_1, \ldots, s_k)$, as s' goes to infinity (Ghosh and Meeden 1997, p. 42). If we repeat this R times, we will have R simulated copies of the Polya posterior for the entire population and when R is large, we can estimate almost any characteristic of the population that we desire.

Example of Use of a Polya Urn: Mapped Plots and Forest Fragmentation

Suppose we are only interested in forested conditions and we used a Polya Urn model to generate copies of the entire distribution of forest proportion in an area? That is, suppose that there are a finite number of plots and that each plot has an area of forest ranging from zero to the maximum plot size, We throw all our s plots into an urn, and pull one out, determining its value of a_i^{do}. We place it and an identical plot back into the urn. We do this s' times, resulting in one copy of the simulated population. We repeat this R times, and we will have R simulated copies of the Polya posterior for the entire population. We can then estimate most population parameters. For instance, the area of forest would be estimated by

$$\tilde{A}_d = R^{-1}\left(\sum_{k=1}^{R}\left[\sum_{h=1}^{H}\sum_{j=1}^{p_h}\left(\sum_{i=1}^{s_i} a_i^{do} + \sum_{i=1}^{s_i'} \hat{a}_i^{do}\right)\right]\right),$$

where $\hat{a}_i^{do}$ is an a_i^{do} from the original sample selected to be repeated by the simulation.

Simulation

To compare these estimators, we wrote a simulation in R and ran it in R version 2.8.1 (R Development Core Team 2010). For the simulation, we defined an area of 1,000,000 units, consisting of two equally sized strata, A and B. Within each stratum, there are two equally sized partitions, 1 and 2. From this basic framework, we defined 12 different populations. These populations were constructed from every combination of two different proportions of forest in each stratum. Stratum A could be 60% forested or 80% forested (P1 = 0.6 or 0.8) and stratum B could be 20% forested or 40% forested (P2 = 0.2 and 0.4). For each of these four combinations, we defined three different allocations of the proportion of the stratum's forestland to each of the equally sized partitions shown in Table 1.

On each of these 12 populations, we applied three different vectors (V1, V2, and V3) for the probability of missing in each partition. These vectors are given in Table 2. This resulted in 36 simulations of 1,000 iterations (complete samples) each. The total area intended for each sample was equivalent to the area sampled by FIA on a subplot basis. That is, the intended sampling effort was 1:36,000. For each iteration, conditions were sampled until enough area was collected to form 167 "plots" (of ⅙ acre each) over the 1,000,000 units (assume acres). Condition sizes were randomly drawn from an empirical distribution of condition sizes observed on all FIA plots (on a subplot basis) since the inception of the mapped plot approach. That distribution is

Table 1. Proportion of forest land in each partition by allocation set.

Allocation set	Partition 1	Partition 2
1	0.52	0.48
2	0.54	0.46
3	0.56	0.44

Table 2. Vectors V1–V3 giving the probability of missing in each partition of each stratum.

Stratum and partition	Vector		
	V1	V2	V3
A1	0.01	0.01	0.5
A2	0.01	0.30	0.0
B1	0.01	0.01	0.1
B2	0.01	0.01	0.0

given in Table 3. For completeness, we also give the equivalent empirical distribution for all plots observed on a macroplot basis. A description of the distinction between plots consisting of four subplots as opposed to plots consisting of four of the larger macroplots can be found in Roesch (2007). Each condition observation was assigned "forest" or "nonforest" by comparison against a uniform random iterate with "p" equal to the percent forest for each partition. The analogous procedure was used to assign "observed" or "not observed. " Each estimator was computed for each iteration. The empirical bias and mean squared error was then calculated for each estimator. For Polya posterior approach, $\tilde{A}_d$ was calculated as described above, with the exception that condition sizes on the plot were used as weights and the parameter R was set equal to 500 using the function wtpolyap in the R library polyapost (Meeden and Lazar 2006).

Discussion of Simulation Results

To place all of the comparisons on an even footing, Table 4 gives the empirical bias calculated over 1,000 samples as a percentage of the true value for $\hat{A}_d$, $\bar{A}_d$, and $\tilde{A}_d$. We see in Table 4 that, in all cases, the biases for $\bar{A}_d$ and $\tilde{A}_d$ are very low, never reaching an absolute value of 1% and seldom exceeding 0.5% in absolute value. We note that the same is true for $\hat{A}_d$ in every allocation set when the vector of missing values V1 is used. This verifies that, with respect to bias, the assumption of missing at random within each stratum is a safe one to make when it is actually true. That is, the significant aspect in that observation for the vector V1 is that the probability of missing is the same in each partition. From that same table, we see that the effect of ignoring the partitions is a dramatic increase in percent bias for $\hat{A}_d$ when vectors V2 and V3 apply, a condition that worsens as the differential, in the proportion of forest, between partitions, gradually increases with allocation set.

Table 3. Distribution of condition sizes on FIA mapped plots (empirical cumulative distribution function).

Condition proportion of plot	Subplot basis	Macroplot basis
<0.25	0.048994569	0.114391608
0.25	0.122260536	0.131038382
>0.25 and <0.5	0.157512839	0.177900622
0.5	0.182438849	0.181732521
>0.5 and <0.75	0.216289726	0.227401219
0.75	0.27638506	0.239650732
>0.75 and <1.0	0.312315026	0.321816697
1.0	1.000000000	1.000000000

The significant aspect of this latter observation is that there is a large difference in the probability of missing values between partitions within each stratum. Because we used actual data to inform the simulation parameters, we know that the simulated differences are not impossible. In fact, even larger differences between identifiable partitions can be observed in the FIA database.

Table 5 gives the MSE calculated over 1,000 samples for each estimator in each of the 36 cases. Note that these values are usually very close, indicating that in many of the cases noted above, in which ignoring the partitions led to an increase in bias, there was a sufficient lowering of variance in $\hat{A}_d$ resulting from drawing sample size strength from the joined partitions to offset the increase in the contribution of bias squared to the MSE. This is to be expected because we are only reporting on subtle differences in the true proportion of forest between partitions within strata. The exception to this successful offset can be seen in the rows of the table corresponding to vector 3 for allocation sets 2 and 3. We see that if differences in both the proportion of forest and the probability of observation between partitions are great enough, the MSE of $\bar{A}_d$ and $\tilde{A}_d$ will be significantly lower than the MSE of $\hat{A}_d$. Even when these differences are very low, the MSEs of $\bar{A}_d$ and $\tilde{A}_d$ are never much higher than the MSE of $\hat{A}_d$. We would expect the MSE of $\bar{A}_d$ to rise relative to the MSE of $\hat{A}_d$ when the partitions become small enough to result in an inadequate sample within the partitions, whereas there remains an adequate sample over the combined partitions.

The simulation reported on here used a small subset of potential parameter settings. We actually ran many other simulations in which the parameter settings varied much more widely. We chose these particular parameter settings because they are either close to those that have been observed in FIA data or could realistically be expected to be observed (if that were possible). Given that any simulation could inadvertently favor one estimator over another, we explain our choices below.

The original simulations contained both widely varying stratum sizes and partition sizes and more choices for percent forest in each stratum and between partitions within each stratum. For verification, we also ran the simulation for the case in which the probability of missing and percent forest was equal across partitions and obtained the expected results with respect to a lowering of MSE when the partitions are ignored. The R code for the simulations can be obtained from the first author on request.

The differences that were observed in simulations based on those other parameter settings, not reported here, were for the most part predictable. Varying the stratum sizes in favor of the more heavily forested stratum, while keeping everything else constant, increases the effect of ignoring partitions. Likewise, varying the stratum sizes in favor of the less heavily forested stratum, while keeping everything else constant, decreases the effect of ignoring partitions. Varying the partition sizes in favor of the more heavily observed partition, while keeping everything else constant, decreases the effect of ignoring partitions. In addition, varying the partition sizes in favor of the less heavily observed partition, while keeping everything else constant, increases

Table 4. Percent bias over 1,000 simulated samples.

Missing values vector	Estimator	Proportion forest											
		Stratum A: 0.6						Stratum A: 0.8					
		Stratum B: 0.2			Stratum B: 0.4			Stratum B: 0.2			Stratum B: 0.4		
		1[a]	2	3	1	2	3	1	2	3	1	2	3
V1	$\hat{A}_d$	−0.48	−0.36	−0.16	−0.22	−0.32	−0.16	−0.23	−0.24	0.14	−0.06	−0.23	0.09
	$\bar{A}_d$	−0.52	−0.42	−0.27	−0.25	−0.37	−0.25	−0.27	−0.32	0.01	−0.09	−0.29	−0.02
	$\tilde{A}_d$	−0.45	−0.43	−0.23	−0.21	−0.40	−0.26	−0.22	−0.30	0.04	−0.06	−0.30	−0.02
V2	$\hat{A}_d$	−0.16	−1.03	−1.00	−0.12	−0.59	−1.23	−0.11	−1.15	−1.37	−0.08	−0.76	−1.50
	$\bar{A}_d$	0.32	−0.08	0.46	0.26	0.17	−0.06	0.42	−0.12	0.19	0.35	0.09	−0.20
	$\tilde{A}_d$	0.31	−0.08	0.43	0.26	0.18	−0.09	0.45	−0.17	0.18	0.39	0.05	−0.21
V3	$\hat{A}_d$	0.67	1.96	3.13	0.21	1.81	2.39	0.88	2.11	3.36	0.46	1.96	2.71
	$\bar{A}_d$	−0.49	−0.20	0.00	−0.77	−0.01	−0.22	−0.25	−0.13	0.06	−0.53	0.02	−0.13
	$\tilde{A}_d$	−0.48	−0.17	0.01	−0.76	0.00	−0.21	−0.25	−0.13	0.07	−0.52	0.01	−0.13

[a] Allocation sets 1–3.

Table 5. MSE ($\times 10^{-9}$) over 1,000 simulated samples.

Missing values vector	Estimator	Proportion forest											
		Stratum A: 0.6						Stratum A: 0.8					
		Stratum B: 0.2			Stratum B: 0.4			Stratum B: 0.2			Stratum B: 0.4		
		1[a]	2	3	1	2	3	1	2	3	1	2	3
V1	$\hat{A}_d$	1.02	1.13	1.15	1.20	1.35	1.42	0.76	0.86	0.88	0.97	1.03	1.03
	$\bar{A}_d$	1.02	1.13	1.15	1.20	1.35	1.41	0.76	0.87	0.87	0.97	1.03	1.02
	$\tilde{A}_d$	1.01	1.16	1.16	1.20	1.38	1.43	0.76	0.87	0.87	0.98	1.04	1.03
V2	$\hat{A}_d$	1.21	1.25	1.10	1.38	1.28	1.29	0.97	0.95	0.89	1.20	1.07	1.11
	$\bar{A}_d$	1.22	1.26	1.13	1.38	1.30	1.28	1.00	0.97	0.91	1.22	1.10	1.09
	$\tilde{A}_d$	1.21	1.27	1.14	1.37	1.30	1.30	1.00	0.98	0.92	1.23	1.11	1.10
V3	$\hat{A}_d$	1.21	1.34	1.36	1.51	1.66	1.65	1.01	1.08	1.24	1.26	1.38	1.55
	$\bar{A}_d$	1.36	1.35	1.35	1.68	1.67	1.62	1.05	0.94	0.89	1.30	1.20	1.20
	$\tilde{A}_d$	1.37	1.36	1.38	1.68	1.68	1.64	1.04	0.94	0.91	1.29	1.21	1.21

[a] Allocation sets 1–3.

the effect of ignoring partitions. An analogous but more subdued set of effects is observed by more widely varying the difference of percent forest between strata. Although we do have examples of greater differential in percent forest between partitions within a stratum, we are not sure how prevalent they are. However, the fact that we have found these examples in conjunction with the study presented here suggests that further investigation is warranted.

Conclusions

Tables 4 and 5, for the most part, reveal very similar results for the weighted mean estimator ($\bar{A}_d$) and the estimator based on the Polya posterior ($\tilde{A}_d$). This was to be expected as it is in keeping with the theoretical development of the latter estimator and underlines the point that if one were merely interested in a population total or mean, there would not be much justification in accepting the computational overhead corresponding to use of the Polya posterior. The advantage of the approach lies in the ability to obtain a probabilistic picture of the entire population distribution. Virtually any distributional measure can be calculated once one has obtained the posterior distribution. The interested reader can explore this topic further in the Bayesian literature, starting with the texts by Ghosh and Meeden (1997), Carlin and Louis (2000), Leonard and Hsu (1999), and Bernardo and Smith (1994). The potential for providing insight coupled with the ever-lowering cost of computer operations suggests that the Polya posterior approach will, and should, become more heavily used.

Literature Cited

BERNARDO, J., AND A. SMITH. 1994. *Bayesian theory.* John Wiley and Sons, New York. xiv + 586 p.

CARLIN, B., AND T. LOUIS. 2000. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall, New York. xvii + 419 p.

COCHRAN, W.G. 1963. *Sampling techniques*, 2nd ed. John Wiley and Sons, New York. xvii + 413 p.

GHOSH, M., AND G. MEEDEN. 1997. *Bayesian methods for finite population sampling*. Chapman and Hall, New York. viii + 289 p.

LEONARD, T., AND J. HSU. 1999. *Bayesian methods: An analysis for statisticians and interdisciplinary researchers*. Cambridge University Press, Cambridge, UK. xiv + 333 p.

MAGNUSSEN, S., AND M. KÖHL. 2002. Polya posterior frequency distributions for stratified double sampling of categorical data. *For. Sci.* 48(3):569–581.

MAGNUSSEN S., B. SMITH, C. KLEINN, AND I.F. SUN. 2010. An urn model for species richness estimation in quadrat sampling from fixed-area populations. *Forestry* 83(3):293–306.

MAGNUSSEN, S., S.V. STEHMAN, C. PIERMARIA, AND M.A. WULDER. 2004. A Pòlya-urn resampling scheme for estimating

precision and confidence intervals under one-stage cluster sampling: application to map classification accuracy and covertype frequencies. *For. Sci.* 50(6):810–822.

MEEDEN, G., AND R. LAZAR. 2006. *Polyapost: Simulating from the Polya posterior.* R package version 1.1.

MOOD, A.M., F.A. GRAYBILL, AND D.C. BOES. 1974. *Introduction to the theory of statistics*, 3rd ed. McGraw-Hill, New York. xvi + 564 p.

PATHAK, P.K. 1964. On sampling schemes providing unbiased ratio estimators. *Ann. Math. Stat*. 35(1):222–231.

PATTERSON, P., J. COULSTON, F. ROESCH, J. WESTFALL, AND A. HILL. 2011. A primer for nonresponse in the US Forest Inventory and Analysis program. *Environ. Monitor. Assess*. DOI: 10.1007/s10661-011-2051-5.

R DEVELOPMENT CORE TEAM. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAO, T.J. 1966. On the variance of the ratio estimator for Midzuno-Sen sampling scheme. *Metrika* 10:89–91.

RAO, T.J. 1977. Estimating the variance of the ratio estimator for the Midzuno-Sen sampling scheme. *Metrika* 24:203–208.

ROESCH, F.A. 2007. Location uncertainty and the tri-areal design. P. 221–226 in *Proc. of the Seventh Annual Forest Inventory and Analysis Symposium, Oct. 3–6, 2005, Portland, ME*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and W.H. McWilliams (eds.). USDA For. Serv. Gen. Tech. Rep. WO-77. vii + 319 p.

ROESCH, F.A. 2008. An alternative view of continuous forest inventories. *For. Sci.* 54(4):455–464.

ROESCH, F.A., JR., E.J. GREEN, AND C.T. SCOTT. 1993. An alternative view of forest sampling. *Surv. Methodol*. 19(2): 199–204.

RUBIN, D.B. 1987. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, New York. xxvii + 291 p.

SCOTT, C.T., W.A. BECHTOLD, G.A. REAMS, W.D. SMITH, J.A. WESTFALL, M.H. HANSEN, AND G.G. MOISEN. 2005. Sample-based estimators used by the forest inventory and analysis national information management system. In *The enhanced Forest Inventory and Analysis program—National sampling design and estimation procedures*, Bechtold, W.A., and P.L. Patterson (eds.). USDA For. Serv. Gen. Tech. Rep. SRS-80. Southern Research Station, Asheville, NC. vii + 85 p.

SRIVENKATARAMANA, T. 1980. A dual to ratio estimator in sample surveys. *Biometrika* 67(1):199–204.

WILLIAMS, M.S. 2001. Performance of two fixed-area (quadrat) sampling estimators in ecological surveys. *Environmetrics* 12:421–436.

Appendix

Notation used in this article is as follows:

P_i^{hd}	ratio of the area observed to be in domain d on plot i to the average plot area within stratum h
a_i^{do}	area (ac) of plot i observed to be in domain d
a_i^o	observed area (ac) of plot i (in the population)
$\bar{\mathbf{a}}_h^o$	average observed plot area in stratum h
n_h	number of ground plots in stratum h
$\bar{P}_h^{do}$	proportion of stratum h plot area observed to be in domain d
P_h^d	proportion of stratum h in domain d
$P_{j\|h}$	proportion of stratum h that is partition j
$\Pi_{d\|j}$	proportion of partition j in domain d
J_h	number of partitions in stratum h
$\hat{P}_h^d$	sample estimate of the proportion of stratum h in domain d
$\hat{P}_{j\|h}$	sample estimate of the proportion of stratum h that is partition j
$\hat{\Pi}_{d\|j}$	sample estimate of the proportion of partition j in domain d
I_i	indicator variable equal to 1 if plot i is observable
I_i^d	indicator variable equal to 1 if domain d on plot i is observable
m_j	number of plots in partition j
A_h^{do}	observed area in domain d in stratum h
A_h^o	observed area in stratum h