

EVALUATING IMPUTATION AND MODELING IN THE NORTH CENTRAL REGION

Ronald E. McRoberts

ABSTRACT.—The objectives of the North Central Research Station, USDA Forest Service, in developing procedures for annual forest inventories include establishing the capability of producing annual estimates of timber volume and related variables. The inventory system developed to accomplish these objectives features an annual sample of measured field plots and techniques for updating data for plots measured in previous years. This paper describes and evaluates the feasibility of updating techniques and compares the bias and precision of the annual estimates they produce. The analyses indicated that simple, plot-level imputation and modeling techniques produced adequately unbiased and precise estimates of basal area per acre for large area estimates.

INTRODUCTION

The Renewable Forest and Rangeland Resources Planning Act of 1978 requires that the USDA Forest Service conduct inventories of forest land in the United States to determine its extent and condition and the volume of standing timber, timber growth, and timber removals. Passage of the Agricultural Research, Extension, and Education Reform Act of 1998 further requires that the Forest Service conduct annual forest inventories in all states with 20 percent of plots to be measured in each state each year.

Forest Inventory and Analysis (FIA) precision standards (USDA-FS 1970) require a sampling intensity of one plot for approximately every 6,000 acres in the North Central region. To satisfy this requirement, the geographical sampling hexagons established for the Forest Health Monitoring Program (White *et al.* 1992) were divided into 27 smaller FIA hexagons, each containing approximately 5,937 acres. An equal probability grid of field plots, designated the Federal base sample, was constructed by establishing a plot in each FIA hexagon. The Federal base sample was systematically divided into five interpenetrating, non-overlapping panels. Each year the plots in a single panel are selected for measurement with panels selected on a 5-year, rotating basis.

At least three approaches to calculating annual FIA estimates from the Federal base sample

have been considered. The simplest approach is to use the data from the 20-percent panel of plots measured in the current year. Although these estimates reflect current conditions, their precision may be unacceptable for some variables due to the small annual sample size. An alternative is to use the data for all plots obtained from the five most recent panels of measurements and employ a moving average estimator. This alternative increases precision because data for all plots are used for estimation; the disadvantage is that the estimates do not reflect current conditions but rather a moving average of conditions over the past 5 years. A third approach is to update to the current year data for plots measured in previous years and then base estimates on the data for all plots. If the updating procedures are unbiased and sufficiently precise, this alternative provides nearly the same precision as the average of all plots but without the adverse effects of using out-of-date information. Two categories of updating techniques, imputation (Rubin 1987) and modeling, are of general interest and were evaluated using a specially created annual database of tree information.

ANNUAL DATABASE

Observations of the same 101,398 trees on 5,086 FIA plots for both the 1977 (Spencer 1982) and 1990 (Miles *et al.* 1995) Minnesota inventories were used to evaluate the updating techniques. These plots represent approximately 14.7 million acres of timberland. (In an

FIA context, timberland is defined as forest land that is capable of producing in excess of 20 ft³ per acre per year of industrial wood crops under natural conditions and that is not associated with urban or rural development (Miles *et al.* 1995.) Plots included in the 1977 inventory were measured between 1974 and 1978; plots included in the 1990 inventory were measured between 1986 and 1991. These plots are termed variable radius plots due to the use of point sampling techniques that select trees with probability proportional to cross-sectional area rather than proportional to the frequency of occurrence in the population (Myers and Beers 1971). Thus, the number of trees in the population represented by a sample tree, termed the tree factor, varies by tree and is calculated as a scaling constant divided by the square of the tree diameter. Tree factors are used to expand the measurements of sample trees to per unit area estimates.

Based on observations of the individual trees, an 11-year database of annual diameters at breast height (DBH) (4.5 ft) and annual status with respect to survival, mortality, and harvest for each tree was created. Construction of the database required distributing total growth between inventories over varying numbers of years for individual trees in each of three categories: (1) trees alive at both inventories; (2) trees that died between inventories due to causes other than harvest; and (3) trees that were harvested between inventories. For trees alive in both inventories, average annual DBH growth was calculated by dividing the total growth in DBH over the measurement interval by the number of years between measurements. Measured DBH for the 1977 inventory was assigned to year 0, and DBHs for the 10 subsequent years were calculated by adding the average annual growth to the previous year's DBH. For trees that died due to causes other than harvest, a year of mortality between 1 and 10 was randomly selected and assigned to the tree independently of years of mortality assigned for other trees on the same plot. For harvested trees, a year of harvest between 1 and 10 was randomly selected and assigned to the tree but with the provision that all trees harvested on the same plot were harvested in the same year. For both mortality and harvested trees, measured DBH for the 1977 inventory was assigned to year 0, and DBHs for subsequent years up to the year of mortality or harvest were calculated by adding previous year's DBH and predictions of annual diameter

growth obtained from individual tree diameter growth models (McRoberts and Lessard 2000). Although these procedures create greater uniformity in annual DBH growth than would be observed, the effects of differences between actual and calculated growth are expected to have minimal impact on evaluations of the updating techniques. Alternatives would require either annual measurement or destructive sampling of all trees. The former alternative would be prohibitively expensive and would risk the masking of actual DBH growth by DBH instrument measurement error; the latter alternative would be prohibited by landowners if not also ecologically disastrous.

Evaluations of the updating techniques were based on plot basal area per acre (BA), a variable representing the sum, scaled to a per acre basis using tree factors, of the cross-sectional areas of live tree boles at breast height.¹ Calculation of unbiased estimates of change in basal area per acre (Δ BA) is difficult using data from variable radius plots (Van Deusen *et al.* 1986). One technique fixes tree factors at the time of the first measurement and bases estimates of Δ BA on the increase in the cross-sectional areas of surviving trees and losses in BA due to mortality. This technique excludes contributions to Δ BA of new trees entering the sample. A second technique recalculates tree factors at every measurement, thus allowing new trees entering the sample to contribute to the Δ BA estimates. However, recalculation of tree factors excludes contributions to Δ BA of the growth of surviving trees, because the product of their cross-sectional areas and tree factors remains constant. A consequence of both techniques is that Δ BA is underestimated.

Although complex approaches to unbiased estimation of Δ BA using variable radius plots have been proposed, they were not considered for this study because evaluation of the updating techniques did not require absolutely precise Δ BA values. For this study, the constant tree factor technique was selected because it incorporates the growth of surviving trees, a primary interest in the construction of these updating techniques. Therefore, using

¹ Unless otherwise noted, all future references to basal area (BA) and annual change in basal area (Δ BA) are understood to be on a per acre basis.

the database of annual tree diameter values and tree factors corresponding to the year 0 DBHs, BA was calculated each year for each plot, and Δ BA was calculated each year for each plot as the difference between BA for the current and previous years.

UPDATING TECHNIQUES

Both imputation and plot-level models were investigated as a means of updating data for plots measured in previous years. Imputation for this application was a three-step process: (1) plots measured in the current year were placed into similarity groups; (2) plots measured in previous years were matched to a group of similar plots measured in the current year; and (3) values from the group of similar plots measured in the current year were selected to replace missing values for plots measured in previous years. For this application, plots were grouped on the basis of similarity in previous year's BA. The groups were created by first ordering all plots measured in the current year with respect to previous year's BA and then creating groups of 20 consecutive plots beginning with the plot with lowest previous year's BA. Plots measured in previous years were then matched to a group of plots measured in the current year on the basis of previous year's BA, whether it was obtained as a measurement or as an updated estimate. For each plot measured in a previous year, a plot was randomly selected with replacement from the group of 20 similar plots measured in the current year, and the latter plot's average annual Δ BA since last measurement was imputed to the former plot; this technique is hereafter referred to as IMPUTE.

Two model-based updating techniques were also investigated. For both modeling techniques, Δ BA for a plot was assumed to be

related to both previous year's BA and to the current survival, mortality, or harvest status of trees on the plot. Thus, based on the annual status of trees, all plots were placed into one of three categories: (1) survival (no mortality or harvest); (2) mortality (at least one mortality tree); and (3) harvest (at least one harvested tree). There were no plots in the 1990 inventory data that had experienced both mortality and harvest since the 1977 inventory. For each category of plots, a simple model of the relationship between Δ BA and previous year's BA was selected, and its parameters were estimated using weighted regression techniques (table 1). In practice, the annual survival, mortality, and harvest status of plots will not be known. Thus, models for predicting the status of plots were also developed. First, all plots in the annual database were ordered with respect to previous year's BA and then placed into groups of 250 consecutive plots beginning with the plot with the lowest previous year's BA. For each group, the proportions of plots in the survival, mortality, and harvest categories were calculated. Simple models of the probabilities of survival, mortality, and harvest were then selected, and their parameters were estimated using maximum likelihood procedures (table 1). With this technique, hereafter referred to as PREDICT, the survival, mortality, and harvest status of each plot measured in a previous year was predicted using random numbers and the status models. Then, given the predicted status, Δ BA for the plot was predicted using the Δ BA models.

Although model predictions of survival, mortality, and harvest status were expected to be unbiased, the combined effects of their uncertainties and those of the Δ BA predictions risked increasing the variability of the annual mean estimates around the means of the annual database values, the standard errors of these

Table 1.—*The models*

Prediction	Category	Model form
Change in annual basal area, Δ BA	Survival	$E(\Delta BA) = \beta_1 [1 - \exp(\beta_2 BA)]$
Change in annual basal area, Δ BA	Mortality	$E(\Delta BA) = \beta_1 + \beta_2 BA$
Change in annual basal area, Δ BA	Harvest	$E(\Delta BA) = \beta_1 + \beta_2 BA$
Change in annual basal area, Δ BA	Disturbed	$E(\Delta BA) = \beta_1 + \beta_2 BA$
Annual probability, P_{surv}	Survival	$E(P_{\text{surv}}) = \exp(\beta_1 BA^{\beta_2})$
Annual probability, P_{mort}	Mortality	$E(P_{\text{mort}}) = 1 - \exp(\beta_1 BA^{\beta_2})$
Annual probability, P_{harv}	Harvest	$E(P_{\text{harv}}) = 1 - E(P_{\text{surv}}) - E(P_{\text{mort}})$

means, or both. Thus, a second model updating technique, based on the assumption that satellite-based remote sensing techniques can be used to accurately detect plots that have experienced substantial disturbance, was developed (Befort 2000). Disturbance for this technique may be due to either mortality or harvest; no distinction is made. Disturbance using remote sensing techniques can be confidently detected for plots satisfying two criteria: previous year's $BA \geq 30 \text{ ft}^2/\text{acre}$, and $(\Delta BA/BA) \leq -0.3$ (Befort, pers. comm.²). Using the annual database values, a simple model of the relationship between ΔBA and previous year's BA for plots satisfying these criteria was selected, and its parameters were estimated using weighted regression techniques (table 1). With this technique, hereafter referred to as REMOTE, updating again involves prediction of both status and ΔBA . First, plots measured in previous years that satisfied the remote sensing disturbance detection criteria were identified, and their ΔBA was predicted using the model constructed for this technique. For the remaining plots measured in previous years, survival, mortality, and harvest status and ΔBA were predicted in the same manner as for the PREDICT technique. However, considerably fewer plots required status prediction with the REMOTE technique.

For both modeling techniques, the uncertainty due to the residual variation around the estimated ΔBA curves was incorporated into the ΔBA predictions. For each estimated curve, distributions of the residuals for narrow categories of predicted ΔBA were estimated. In application, whenever a value of ΔBA was predicted, a corresponding residual from the appropriate distribution was randomly generated and added to the prediction. Thus, the estimates of standard errors of mean BA estimates obtained using the model updating techniques include the uncertainty of the model predictions due to residual variation.

SIMULATING THE INVENTORY

The feasibility of the updating techniques and the bias and precision of their annual BA estimates were evaluated by using the annual database as the basis for simulating the process of annually inventorying the 14.7 million

timberland acres in Minnesota. For each of 250 simulations, 2,476 plots from among the 5,086 timberland plots were randomly selected to mimic the annual inventory intensity of 5,937 acres per plot. Each simulation was initiated with a simulated complete inventory of the 2,476 timberland plots by beginning with the annual database year 0 values. On a rotating basis, 20 percent of these plots were selected for measurement each year. Simulated measurement of a plot consisted of replacing its estimated BA value with the value for the appropriate year in the annual database. For the remaining 80 percent of plots for which measurement was not simulated in the current year, data were updated using each of the three inventory techniques. Each year, the mean BA across all plots and the standard error of the mean were calculated for each updating technique and for annual database of values; the latter estimates were designated TRUE. Following the simulations, the median values of the distributions of the annual means and the standard errors of the means were determined for each technique.

RESULTS

Evaluations of the updating techniques entailed comparing the median values for the 250 simulations of the estimated means of annual BA across all plots and the standard errors of the means obtained using the three updating techniques to the corresponding annual means and standard errors obtained from the annual database values. A comparison of the TRUE and IMPUTE annual means revealed that the imputation technique produced estimates that exhibited negligible bias with respect to the TRUE values (Table 2). In addition, the similarity between the TRUE and IMPUTE standard errors indicated that the IMPUTE technique quite accurately estimated the uncertainty in the TRUE means. A comparison of the median values of the TRUE, PREDICT, and REMOTE annual means revealed that neither modeling technique exhibited conspicuous bias (table 2). A comparison of the median standard errors of the means indicated that both modeling techniques adequately estimated the TRUE standard errors. As expected, the variability of the PREDICT annual means around the TRUE means was greater than for the REMOTE means.

A further comparison of the updating techniques was made by calculating the root mean

² William Befort, Division of Forestry, Minnesota DNR, November 9, 1999.

Table 2.—Median values of annual means and standard errors of means for 250 simulations

Year	TRUE		IMPUTE		PREDICT		REMOTE	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	56.24	0.74	56.12	0.76	56.38	0.73	56.09	0.74
2	55.53	0.73	55.23	0.75	55.39	0.73	55.22	0.73
3	54.07	0.73	54.28	0.75	54.34	0.72	53.88	0.73
4	53.14	0.72	53.41	0.74	53.30	0.72	52.93	0.72
5	51.64	0.72	52.14	0.73	52.16	0.72	51.64	0.72
6	50.80	0.72	51.31	0.72	51.31	0.71	50.75	0.72
7	49.72	0.72	49.66	0.72	50.05	0.71	49.55	0.72
8	48.48	0.72	48.73	0.72	49.12	0.71	48.47	0.71
9	45.97	0.72	47.28	0.72	47.75	0.71	46.59	0.71
10	45.24	0.72	45.91	0.73	46.54	0.71	45.52	0.71

square error of the squared deviations of the updated annual means from the corresponding TRUE annual means for years 5-10 (table 3). The first 4 years were excluded in this comparison, because annual means for these years retained a component of the year 0 complete inventory. The resulting 5th percentile, median, and 95th percentile values for distributions of root mean square errors indicated that the differences between the IMPUTE and REMOTE means were small with respect to root mean square deviation, although the REMOTE results were somewhat better than the IMPUTE results. The similarity of results for these updating techniques may be partially attributed to the large area represented by the aggregation of data over this large number of plots; it is yet to be determined if these results hold for smaller areas.

CONCLUSIONS

Several conclusions emerged from these analyses. First, the simple, plot-level updating techniques were not only feasible, but they produced acceptable estimates of both annual BA means and standard errors of the means for large area estimates. Second, although the

REMOTE technique produced somewhat better results, the quality of the means obtained with the modeling and imputing techniques relative to the TRUE means was similar. Third, because 5-year Δ BA is usually small compared to BA 5 years in the past and because the uncertainty in Δ BA predictions is small compared to the natural variation in BA among plots, Δ BA appears to be an appropriate quantity to use as the basis for updating. Fourth, but less conclusively, a combination of disturbance detection using remote sensing procedures and model predictions of survival, mortality, and harvest status appeared to be a better alternative than using only model predictions of status. Finally, additional testing is appropriate to determine if these large area results hold for smaller numbers of plots representing smaller timberland areas.

LITERATURE CITED

Befort, W. 2000. **Change detection, stratification, and mapping for continuous forest inventory**. 1999 proceedings of the Statistics and the Environment Section; annual meeting of the American Statistical Association; 1999 August 8-12; Baltimore, MD.

Table 3.—Root mean square deviation: updated versus TRUE annual means

Statistic	TRUE	IMPUTE	PREDICT	REMOTE
5th percentile	0.00	0.58	0.66	0.29
Median	0.00	0.68	1.01	0.33
95th percentile	0.00	0.77	1.34	0.54

Alexandria, VA: American Statistical Association.

McRoberts, R.E.; Lessard, V.C. 2000. **Estimating uncertainty in annual forest inventory estimates**. 1999 proceedings of the Statistics and the Environment Section; annual meeting of the American Statistical Association; 1999 August 8-12; Baltimore, MD. Alexandria, VA: American Statistical Association.

Miles, P.D.; Chen, C.M.; Leatherberry, E.C. 1995. **Minnesota forest statistics, 1990, revised**. Resour. Bull. NC-158. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station.

Myers, C.C.; Beers, T.W. 1971. **Point sampling and plot sampling compared for forest inventory**. Res. Bull. 877. August, 1971. West Lafayette, IN: Purdue University, Agriculture Experiment Station.

Rubin, D.B. 1987. **Multiple imputation for nonresponse in surveys**. New York: Wiley.

Spencer, J.S., Jr. 1982. **The fourth Minnesota forest inventory: timber volumes and**

projections of timber supply. Resour. Bull. NC-57. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station.

USDA Forest Service (USDA-FS). 1970. **Operational procedures**. For. Serv. Handb. 4809.11, Chapter 10:11.1-1—11.1-3. Washington, DC: U.S. Department of Agriculture, Forest Service.

Van Deusen, P.C.; Dell, T.R.; Thomas, C.E. 1986. **Volume estimation from permanent horizontal points**. Forest Science. 32(2): 415-422.

White, D.; Kimerling, J.; Overton, S.W. 1992. **Cartographic and geometric components of a global sampling design for environmental monitoring**. Cartography and Geographic Information Systems. 19: 5-22.

ABOUT THE AUTHOR

Ron McRoberts is a Mathematical Statistician and Group Leader with the Forest Inventory and Analysis program of the North Central Research Station, USDA Forest Service, St. Paul, MN.