A COMPARISON OF SEVERAL TECHNIQUES FOR ESTIMATING THE AVERAGE VOLUME PER ACRE FOR MULTIPANEL DATA WITH MISSING PANELS[1]

Dave Gartner and Gregory A. Reams[2]

Abstract—As Forest Inventory and Analysis changes from a periodic survey to a multipanel annual survey, a transition will occur where only some of the panels have been resurveyed. Several estimation techniques use data from the periodic survey in addition to the data from the partially completed multipanel data. These estimation techniques were compared using data from two periodic surveys from Georgia. The comparison criteria were based on (1) an estimated mean within the confidence interval derived from using the complete multipanel data set and (2) a small, estimated standard error that does not underestimate the complete data standard error. Multiple imputation matching performed best; the double sampling ratio estimator also performed well. Two methods—single imputation using group means and single imputation using matched stands—both underestimated the standard error. Replacing the missing observations with growth model predictions using SETWIGS caused an overestimation of the mean.

INTRODUCTION

The USDA Forest Service, Forest Inventory and Analysis (FIA) Units have been conducting surveys of commercial forest land in the continental United States since the 1930s. Traditionally, FIA has conducted surveys on a State level with a cycle from 6 to 15 years with a mode of about 10 years in the South. Prior to a tightening of the supply and demand relationship for wood fiber in the South, the 10-year cycle was considered timely enough (Reams and others 1999).

With the growing demand for wood products from the South, the need for more current inventory information has become apparent. To meet this need, the Agricultural Research, Extension, and Education Reform Act (Public Law 105–185): (The Farm Bill) of 1998 mandated FIA to implement an annual inventory system Nationwide.

Southern FIA is changing from single-panel (periodic) whole-State surveys to an interpenetrating five-panel annual survey (Reams and Van Deusen 1999). The latter design divides the large periodic survey into five repeated smaller samples, called panels (Reams and Van Deusen 1999). By providing information about the variations between years, the separate annual samples are able to estimate annual and secular trends.

The new annual five-panel design will give rise to new estimation techniques. The new official FIA estimate will be a moving average using the annual survey data (Reams and others 1999). The moving average is operationally convenient, requires a minimum of assumptions, and is basically design-based as opposed to model-based.

To understand how the moving average will be implemented, consider the following situation: (1) the last full periodic survey has been completed; (2) starting immediately afterwards, the five-panel annual system has been implemented; and (3) three panels have been completed and now an estimate of live standing volume per acre for Georgia is needed. The official FIA estimate will be the average, using the annual survey data from the plots in panels one through three and the closeout periodic survey data for plots in panels four and five. Note that plots from panels four and five have yet to be measured under the annual system; therefore, the plot attributes are at least 3 years old.

Some users of the annual survey data suggested using statistical modeling techniques to update the data values. Some of these techniques replace the missing data in the unsurveyed panels with estimates from the surveyed panels. In the statistical literature, this replacement of missing data with modeled data is called imputation (Rubin 1987). After imputing data values for old or unmeasured plots, it would be tempting to analyze a simulated-complete data set as a complete data set. However, this approach tends to understate the true variance in the estimates (Little and Smith 1987, Van Deusen 1997).

This study compares the performance of several techniques. In addition to imputation, the double sampling ratio estimator was used as a comparison. Because multiple imputation is conservative in its estimate of the variance (Rubin 1987), the variance estimate for double sampling is expected to be lower than for multiple imputation.

METHODS

Data

We simulated the end of the third panel, having access to the data from the first three panels and the last periodic survey. The variable in the comparison is the statewide average volume of live trees in cubic feet per acre. To compare predicted values with observed values, we used the 1988 and 1996 periodic surveys from Georgia. To simulate conditions at the end of year three, we deleted the

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Mathematical Statistician and Supervisory Mathematical Statistican, USDA Forest Service, Southern Research Station, 2730 Savannah Highway, Charleston, SC 29414, and USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804, respectively.

stand volume data from 40 percent of the 1996 plots. We used only unit, county, plot, forest type, and stand origin from 1996 plots with deleted stand volume data.

Operational Information Assumptions

We assumed that the regional FIA units know which areas have been harvested and the volumes of any remnant stands. In the simulation, we coded as "cut" all stands with trees that were marked as cut in the 1996 survey. We put these cut stands in a separate data set and did not delete any of their 1996 volume data. The prediction methods could have been modified to handle harvested stands by including a probability of being harvested and a prediction of harvesting intensity, but we decided against this approach.

Data Preparation

Two major changes in the plot sampling protocols occurred between the 1988 survey and the 1996 survey. One was the change from using variable-radius plots to fixed-radius plots. The second was the handling of plots that contained more than one stand. During the 1988 survey, if any of the subplots fell into a different stand than the central subplot, that subplot was rotated until it fell within the same stand as the central subplot. During the 1996 survey, if any subplot fell into a different stand or stands, the subplots were not rotated but the different stands were given different codes, called condition codes. To make sure the 1988 data used to predict 1996 volumes matched the observed 1996 stands, only pure stands were used. If the trees measured in 1988 appeared in more than one 1996 condition code, we removed the plot from the data set. If a 1996 condition code did not have any 1988 trees, we removed the condition code from the data set.

After we removed these data, 3,749 plots remained. We calculated the live tree volumes in cubic feet per acre for each plot for both survey years. We then placed the 1,194 stands with cut trees in a separate data set. We simulated the two unsurveyed panels by deleting the 1996 volume data for 40 percent (1,020 out of 2,555) of the remaining stands. To suit the estimation techniques being run by forest type, we placed the forest types with fewer than seven plots in the 60 percent of the data that represent the three surveyed panels.

Estimation Techniques

Three-panel method—The first estimation technique uses only the 1996 volume data for the cut stands and the three surveyed panels. We calculate the means and standard errors for both groups, and then combine them, weighting the uncut stands to include the number of stands with missing volumes.

Single imputation group means—Adding information on the forest types yields the second estimation technique. With single imputation group means, the missing volumes are replaced with the average observed volume for that forest type. We then recombine the cut and uncut data sets and use standard estimation procedures.

Single imputation matching—Adding the information on the 1988 volumes yields two more types of estimation techniques: single imputation matching, and multiple imputation matching. For single imputation matching, we find stands with 1988 volumes and forest types that match those of the stands with missing 1996 volumes. Once we find a matching stand, we replace the missing 1996 volume data with the data from the matched stand. Then we recombine the cut and uncut data sets and use standard estimation procedures.

Multiple imputation matching—Multiple imputation (Rubin 1987) matching differs from single imputation matching, in that a set of possible donor plots is sought for each missing value. A separate donor stand is then randomly chosen for each missing value from its donor pool. We repeat this process of randomly choosing donor plots several times, and combine the results from the repetitions.

For each imputed data set, we calculate the statistic of interest (mean live tree volume per acre), denoted as $\hat{Q}_{*l}$. The variance of $\hat{Q}_{*l}$ is denoted as U_{*l}. In this case, U_{*l} is the standard error. The function for the estimated mean is

$$\overline{Q}_m = \sum_{l=1}^{m} \hat{Q}_{*l} \Big/ m \tag{1}$$

where m is the number of repetitions of the imputation process. The estimator for the variance of $\overline{Q}_m$ has two components. The first component is the average of the variances of this mean:

$$\overline{U}_m = \sum_{l=1}^{m} U_{*l} / m. \tag{2}$$

The second component of this variance estimator is the variance of the $\hat{Q}_{*l}$'s:

$$B_m = \sum_{l=1}^{m} (\hat{Q}_{*l} - \overline{Q}_m)^2 / (m-1). \tag{3}$$

These two components are combined in the following manner:

$$T_m = \overline{U}_m + (1 + m^{-1}) B_m. \tag{4}$$

When standard errors are mentioned in the results for multiple imputation techniques, we use the square root of T_m. The estimated overall mean has a t distribution with mean $\overline{Q}_m$ and standard error of the square root of T_m. The degrees of freedom according to Rubin (1987) is

$$\nu_m = (m-1)\left[1+\frac{\overline{U}}{(1+m^{-1})B_m}\right]^2 \quad (5)$$

This degrees of freedom has been given a modifier for possible small sample sizes (Barnard and Rubin 1999). The modifier is

$$\hat{\nu}_{obs} = (1-\gamma)\nu_0(\nu_0+1)/(\nu_0+3), \quad (6)$$

where $\gamma = (1+m^{-1})B_m / T_m$ and ν_0 is the degrees of freedom of the full sample if no data values are missing. The final degrees of freedom is

$$\nu_m^* = \left[\frac{1}{\nu_m}+\frac{1}{\hat{\nu}_{obs}}\right]^{-1}. \quad (7)$$

The main advantages of multiple imputation over single imputation are that the variance caused by the process of randomly choosing donor plots is empirically estimated (eq. 3) and is explicitly included in the estimate of the overall variance.

Multiple imputation modeling—The assumption that the 1996 volume is an approximately linear function of 1988 volume for uncut stands yields another estimation technique. Multiple imputation modeling estimates the parameters from a linear regression of 1998 volume on 1996 volume. We modify these parameters by adding a random error term, determined by decomposing the parameters' variance-covariance matrix. Using a Cholesky decomposition turns the variance-covariance matrix into a set of variances for independent normal variables. We then multiply random normal variates by these variances and add them to the parameter estimates. We use these modified parameters to estimate the missing values. We calculate an imputed standard deviation by randomly generating a Chi-square variable and multiplying it by the observed standard deviation. We generate standard normal deviates, multiply them by this imputation standard deviation, and add them to the estimates for the missing values. We repeat this process several times and analyze the results in the same manner as for the multiple imputation matching data. Thankfully, the current multiple imputation software does all of these computations. As with the multiple imputation matching method, the objectives of repeating the process are (1) to empirically estimate the variance of the mean due to randomization, and (2) to incorporate this variance into the total variance for the estimator.

Single imputation growth model—Growth models use the 1988 tree-level information, such as species, diameters, and expansion factors, along with plot site index. We simply replace the missing volumes with the growth model predictions. For this study we used the growth model, SETWIGS (Bolton and Meldahl 1990).

Multiple imputation using growth model predictions—We could incorporate growth model predictions into multiple imputation efforts in two different ways. The first way would be to replace the missing data with the growth model projections. According to Rubin (1987), the proper method for replacing the missing data with the growth model predictions is the same method used for the linear regression predictions in multiple imputation modeling, including decomposing the parameter variance-covariance matrix and imputing new parameters and standard errors. Unfortunately, the current multiple imputation software will not calculate these values. We decided not to use this method because of the effort it required.

The second method of incorporating growth model predictions into multiple imputation is using the growth model predictions as the covariate. We used multiple imputation matching and multiple imputation modeling techniques by replacing the 1988 volume information in the earlier multiple imputations with the growth model predicted volumes.

Double sampling ratio estimator—We also used the classical sampling statistical technique called double sampling using a ratio estimator (Cochran 1977). Double sampling occurs when a sample is taken and the value of one variable (X) is observed. Then a subsample of the first sample is taken and the value of the variable of interest (Y) is observed. The estimated average for Y on the whole sample using a ratio estimator ($\overline{y}_R$) is the average of X for the whole sample ($\overline{x}'$), times the ratio of the average of Y for the subsample ($\overline{y}$) divided by the average of X for the subsample ($\overline{x}$):

$$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\overline{x}'. \quad (8)$$

The variance of this estimator is given by equation 9:

$$V(\overline{y}_R) = \frac{1}{n'}s_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)s_d{}^2, \quad (9)$$

where n is the number of observations in the subsample, n' is the number of observations in the full sample, and

$$s_d{}^2 = \sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2/(n-1). \quad (10)$$

Note this estimator is not the same as regression through the origin. In this instance, we used Y as the 1996 volume for the uncut stands and the growth model-predicted volumes for X. We then combined the estimates and variances for the cut and uncut stands.

Solas Software

We ran all multiple imputations using Solas software (1999). The multiple imputation matching techniques followed the propensity score method, which uses a logistic regression equation to predict the propensity of an X value to correspond with a missing Y value. A donor pool of observations is created from a local neighborhood of propensity values. If only one covariate is used, the predicted propensity is a monotonic function of the covariate, and the neighborhood of predicted propensities is the same as the neighborhood of the covariate values. However, this

condition will not necessarily be true if more than one covariate is used. Because Solas creates a separate equation for each level of a grouping variable, adding grouping variables, such as forest type, will still allow the propensity scoring method to act as a matching method.

Using multiple grouping variables causes a problem for Solas. Solas can run only 30 groups at a time. With more than 30 forest types, we had to break the data set into several parts. We imputed each part separately, then merged them together again. Solas also runs out of memory and has trouble with large data sets. We hope the new SAS multiple imputation procedure will have fewer limitations.

We also ran the single imputation matching using Solas. Instead of limiting the donor pool to just the plot with the next larger and the next smaller 1988 plot volumes, Solas required that the donor pool include at least the next two larger and two smaller plot volumes. In keeping with single imputation, we picked only one value per observation with missing data.

RESULTS

Because the multiple imputation techniques use randomization, we ran all multiple imputation techniques five times to estimate the variability caused by the randomization. We report this variability for the estimated statistics (tables 1 to 3).

Means

Figure 1 shows the relationship between 1988 and 1996 volumes for the stands that were not cut. Several stands along the 1988 volume axis show that criteria for determining cut stands did not catch all of the stands that lost volume. The volume losses were probably the result of natural disturbances as opposed to harvesting.

The full 1996 data have a mean of 1,569.96 ft^3 per ac (table 1) and a standard error of 22.18 (table 2). Because the multiple imputation methods were run five times each, the overall means and standard deviations of the estimated means appear on table 1. All but one of the estimation techniques gave means within 1.15 standard errors of the full data mean. Replacing the missing 1996 volume data with the SETWIGS growth model projections provided a mean of 1,771.63 ft^3 per ac.

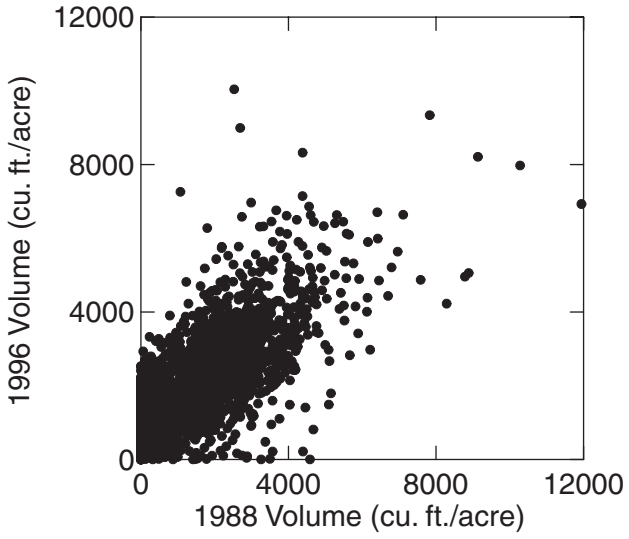

Figure 1—Observed 1988 stand volumes versus observed 1996 stand volumes.

Figure 2 shows the relationship between the volumes predicted by SETWIGS and the 1996 observed volumes for uncut stands. The line on the graph shows where the predicted volume equals the observed volume. Most of the points fall below the line, showing that the SETWIGS-predicted volumes were larger than the observed volumes. This estimated mean is about nine standard errors above the mean found by using all of the data.

Table 1—Estimated mean stand volume by estimation technique

Estimation technique	Means
All of the data	1,569.96
Three surveyed panels only	1,568.60
Group mean imputation	1,577.01
Single imputation matching: volume 1988	1,556.53
Multiple imputation matching: volume 1988[a]	Mean 1,571.57, std. dev. 1.89
Multiple imputation modeling: volume 1988[a]	Mean 1,570.77, std. dev. 17.91
Single imputation: SETWIGS	1,771.63
Multiple imputation matching: SETWIGS[a]	Mean 1,574.68, std. dev. 4.14
Multiple imputation modeling: SETWIGS[a]	Mean 1,575.01, std. dev. 9.61
Double sampling ratio estimator: SETWIGS	1,577.15

[a] Multiple imputation techniques were run five times. The reported results are the mean and standard deviation of the five runs.

Table 2—Standard errors of the various estimation techniques

Estimation technique	Standard error
All of the data	22.18
Three surveyed panels only	25.23
Single imputation group means	19.45
Single imputation matching: volume 1988	21.90
Multiple imputation matching: volume 1988[a]	Mean 23.86, std. dev. 0.34
Multiple imputation modeling: volume 1988[a]	Mean 30.80, std. dev. 3.16
Single imputation SETWIGS	25.43
Multiple imputation matching: SETWIGS[a]	Mean 23.43, std. dev. 0.59
Multiple imputation modeling: SETWIGS[a]	Mean 23.97, std. dev. 1.38
Double sampling ratio estimator: SETWIGS	23.15
All of the data: stratified sample	20.16

[a] Multiple imputation techniques were run five times. The reported results are the means of the standard errors and their standard deviations.

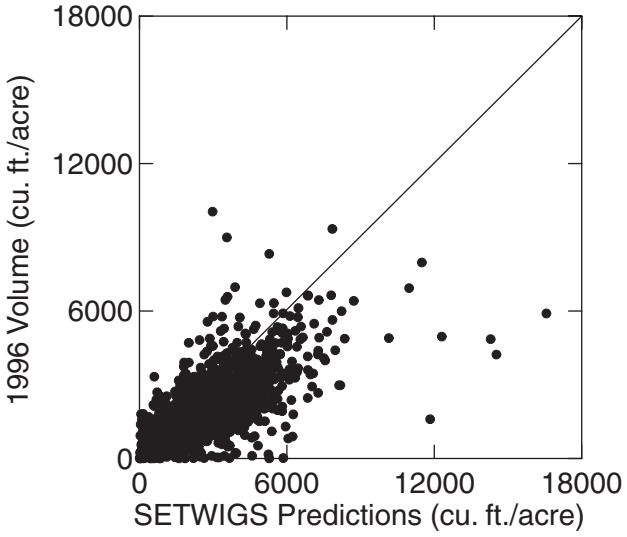

Figure 2—SETWIGS predicted volumes versus observed 1996 stand volumes.

Standard Errors

The estimated standard errors are shown in table 2. As with the estimated means, for the multiple imputation methods, the means and standard deviations of the five estimated standard errors are reported in table 2. The standard errors for the single imputation group means and the single imputation matching are smaller than the standard error found when using the full data set. The double sampling ratio estimator and the multiple imputations have larger standard errors than using the full data set. Generally, the matching techniques outperformed the modeling techniques, and SETWIGS predictions outperformed 1988 volume.

Mean Square Error

The mean square error is the bias squared plus the standard error squared. The mean square error for each method appears in table 3. As with the estimated means, for the multiple imputation methods, the means and standard deviations of the five mean square errors are shown in table 3. Only the single imputation group means

Table 3—Mean square errors of the various estimation techniques

Estimation technique	Mean square error
All of the data	491.95
Three surveyed panels only	638.40
Single imputation group means	428.00
Single imputation matching: volume 1988	659.97
Multiple imputation matching: volume 1988[a]	Mean 581.69, std. dev. 16.23
Multiple imputation modeling: volume 1988[a]	Mean 1361.04, std. dev. 503.69
Single imputation: SETWIGS	41317.47
Multiple imputation matching: SETWIGS[a]	Mean 590.03, std. dev. 66.46
Multiple imputation modeling: SETWIGS[a]	Mean 686.78, std. dev. 100.88
Double sampling ratio estimator: SETWIGS	587.62
All of the data: stratified sample	406.43

[a] Multiple imputation techniques were run five times. The reported results are the means of the mean square errors and their standard deviations.

has lower mean square errors than using all of the data. As with the standard errors, all of the multiple imputation techniques had higher mean square errors, with the matching techniques performing better than the linear modeling techniques.

DISCUSSION

The single imputation group means technique removes the variation between the missing volumes and their means by forest type. Removing this variation causes the standard error and the mean square error to be underestimated. Single imputation matching limits the variation in a similar manner. With a given set of original stand conditions, there is a range of possible ending conditions. Single imputation matching limits the variation less than does single imputation group means and, therefore, underestimates the standard error less than group means imputation. However, it still underestimates the standard error and the mean square error.

The standard error for double sampling ratio estimator is very close to the standard errors for the multiple imputation matching methods. The double sampling ratio estimator required splitting the plots into harvested and unharvested strata, while the multiple imputation methods did not stratify the data. Therefore, the multiple imputation matching methods actually performed slightly better than the double sampling ratio estimator.

Creating an inventory estimate for the harvested stands is more complex, especially for the double sampling ratio estimator and the single imputation growth model method. To use either of these methods on the harvested plots would require either the ability to remotely sense all harvests each year, or the creation of probability-of-harvest models and a method of allocating partial harvests to individual trees. The southern FIA unit currently does not have the budget or the infrastructure to be able to remotely sense harvesting on an annual basis. Some work has been done on probability of harvest models for stands, but no work has been done on methods of allocating partial harvests to individual trees. The multiple imputation methods would not require using remotely sensed information on harvesting or the probability of harvest models. Currently, about 20 percent of the stands are harvested (either clearcut or partial) within a 5-year cycle.

All of the imputation techniques predict plot level data and then calculate overall means and standard errors. The double sample ratio estimator, however, is not an imputation technique because it does not calculate values for the missing observations. Therefore the double sampling ratio estimator may not be suitable for variables that are difficult to model. The number of snags per acre, amount of fallen woody debris, and ownership are examples of such variables. Some tables, such as the diameter distribution tables, may be sensitive to the differences between model predictions and observed data, and may not be fit well by the double sample ratio estimator.

The multiple imputation techniques may have an additional advantage. While each table requires a separate run using double sampling ratio estimator, properly constructed multiple imputation data sets can be used for all tables simultaneously.

ACKNOWLEDGMENTS

We thank Michael Thompson for his assistance with the Georgia data and the volume equations.

REFERENCES

Barnard, J.; Rubin, D. 1999. Small-sample degrees of freedom with multiple imputation. Biometrika. 86: 948–955.

Bolton, R.K.; Meldahl, R.S. 1990. Design and development of a multipurpose forest projection system for southern forests. Alabama Agricultural Experiment Station Bulletin 603. Auburn, AL: Auburn University. 51 p.

Cochran, W.G. 1977. Sampling techniques. New York: John Wiley. 427 p.

Little, R.J.A.; Smith, P.J. 1987. Editing and imputation for quantitative survey data. Journal of the American Statistical Association. 82: 58–68.

Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. New York: John Wiley. 258 p.

Statistical Solutions, Ltd. 1999. SOLAS for missing data analysis 2.0 [User reference]. Cork, Ireland: Statistical Solutions, Ltd. 67 p.

Reams, G.A.; Roesch, F.A.; Cost, N.D. 1999. Annual forest inventory: Cornerstone of sustainability in the South. Journal of Forestry. 97(12): 21–26.

Reams, G.A.; Van Deusen, P.C. 1999. The southern annual forest inventory system. Journal of Agricultural, Biological, and Environmental Statistics. 4(4): 346–360.

Van Deusen, P.C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27: 379–384.