A COMPARISON OF SEVERAL TECHNIQUES FOR IMPUTING TREE LEVEL DATA

David Gartner and Greg Reams [1]

ABSTRACT.—As Forest Inventory and Analysis (FIA) changes from periodic surveys to the multipanel annual survey, new analytical methods become available. The current official statistic is the moving average. One alternative is an updated moving average. Several methods of updating plot per acre volume have been discussed previously. However, these methods may not be appropriate for updating more detailed data such as diameter distribution and species composition tables. Several methods for updating these more detailed data will be compared. Methods to be compared include imputing whole tree lists from donor plots and imputing all the individual trees for each plot. The data from the last periodic survey of Georgia and Georgia's first panel will be used to compare the different imputation methods.

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

With the growing demand for wood products from the South, the need for more current inventory information has become apparent. This need is evidenced by the Agricultural Research, Extension, and Education Reform Act (PL 105-185) (The Farm Bill) of 1998, which congressionally mandates 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). Panel denotes a sample in which the same plots are measured on two or more occasions. For those familiar with the longstanding 10-year periodic FIA survey, the interpenetrating annual panel design is analogous to taking the large periodic survey and dividing it into five repeated smaller samples (Reams and Van Deusen 1999). The chief advantage of the annually repeated survey over the traditional periodic design is that the separate annual samples provide information about variations that occur between the periods. This results in the ability to estimate annual and secular trends.

The official FIA estimate will be a moving average using the annual survey data (Reams and others 1999). Equation 1 is the formula for the moving average, where τ is the value for the different panels.

$$\sum_{i=1}^{5} \frac{\tau_{T-i}}{w_i} \qquad (1)$$

However, some users of the annual survey data have suggested they would like to use data values that can be considered either made current or updated in some fashion via a statistical modeling approach. These approaches would project the oldest data forward in time by replacing the missing future panels with estimates based on data from existing panels, e.g., an updated moving average. Equation 2 is the formula for the updated moving average.

$$\sum_{i=1}^{3} \frac{\tau_{T-i}}{w_i} + \sum_{i=4}^{5} \frac{\hat{\tau}_{T-i+5}}{w_i} \qquad (2)$$

Operational implementation of this updated moving average would not occur before year 8. Changes from panel 1 to panel 6, from panel 2 to panel 7, and from panel 3 to panel 8

[1] Mathematical Statistician, Southern Research Station, U.S. Department of Agriculture, Forest Service, 4700 Old Kingston Pike, Knoxville, TN 37919. Phone: (865) 862-2066, e-mail: dgartner@fs.fed.us; and Section Head for Methods and Techniques, Southern Research Station, U.S. Department of Agriculture, Forest Service, 200 WT Weaver Boulevard, Asheville, NC 22804.

can be used to predict the changes from panel 4 to panel 9 and from panel 5 to panel 10. Figure 1 shows what data would be used, with the dotted line under the label "Updated Moving Average" representing data to be predicted.

This replacement of missing data with modeled data is referred to as imputation in the statistical literature (Rubin 1987). After data values for old or unmeasured plots have been imputed, it would be tempting to analyze a simulated-complete data set as if it were a complete data set. However, this approach would tend to understate the true variance in the estimates (Little and Smith 1987, Van Deusen 1997).

Several methods of updating plot per acre volume have been discussed (Gartner and Reams 2000). However, these methods would become very cumbersome to use for tables with many entries, such as the diameter distribution by species composition tables. Methods that impute tree-level values will be needed to create these tables.

METHODS

Data

The data used for this study are from the seventh Georgia statewide survey completed in 1997 and the first panel of the new rotating panel system. The same plot locations are used in both systems, so that plots can be compared at two different times. The same fixed-radius plot designs were used in both surveys, allowing the tree lists from the two surveys to be compared.

The list of plots started with the plots occurring in both the seventh Georgia statewide survey and the first annual panel. All plots with more than one condition code were removed from the data set, as were all nonforested plots and all plots with harvesting. This left 419 plots available for this analysis.

Because we are attempting to simulate the conditions for the updated moving average, where two of the five panels are going to be projected forward in time, we removed the first panel data from 40 percent of the plots. We were then able to use the other 60 percent to impute the updates to the plots with missing data and to compare these imputations with the results from the full data set.

General Description of Multiple Imputation

The intent of imputation is to generate replacements for missing values from the same posterior distribution as the missing values. This is done in this study by either matching methods or modeling.

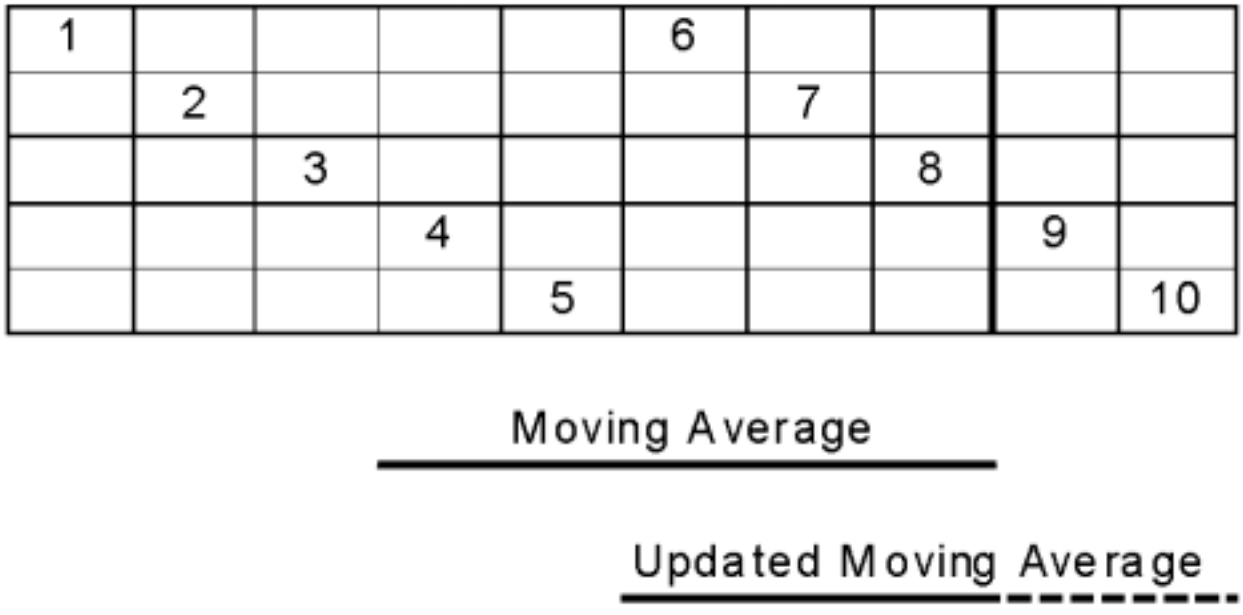

Figure 1.—*Differences in the data used at year 8 for the moving average and the updated moving average.*

Matching imputation methods simulate this posterior distribution by finding time 2 values in the neighborhood of the available time 1 values and randomly choosing one. This is very similar to the k-nearest neighbor approach. However, where the k-nearest neighbor uses just the mean of the neighboring values, matching imputation methods can loosely be thought of as using this mean value and adding a randomly chosen observed error component.

Figure 2 shows how matching imputation works. In the column on the left are data points containing time 2 data in the time 1 neighborhood of the data point missing time 2 data (on right). Because we are using multiple imputation, values from the column are repeatedly drawn at random with replacement to fill in the missing data. This process is repeated for each missing data point.

Modeling imputation methods use a regression model to estimate the mean value and add an error term. However, because the regression parameters are also random variables, for the replacement values to match the posterior distributions, error terms also have to be added to the estimated parameters.

Figure 3 shows how modeling imputation works. The estimated standard deviation is the distance between the outer two lines. The estimated standard deviation does not include the variance associated with the standard error of the prediction, the distance from the center prediction line and the inner dashed lines. To incorporate this variation, most imputation programs first add random error terms to the parameters using the variance-covariance matrix, before calculating the estimated values for the missing data. Then random error terms are generated from the predicted standard error.

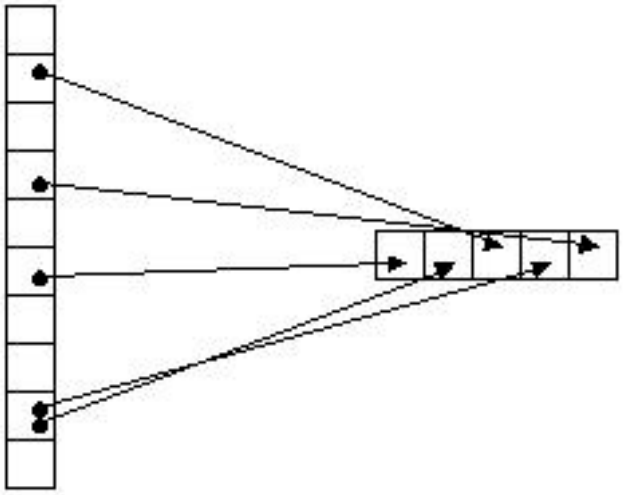

Figure 2.—*How matching imputation works.*

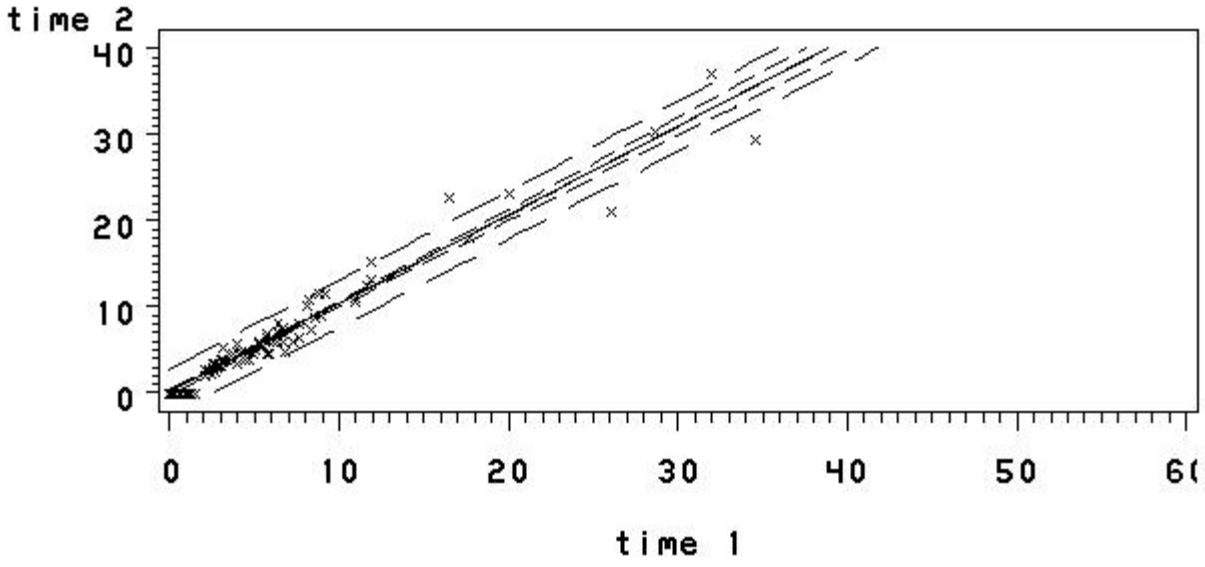

Figure 3.—*Modeling imputation using linear regression, including the prediction line, and lines for the standard error of prediction and the standard deviation of the estimate.*

For each run, the imputation process is repeated for all of the missing data set five times. Then the five repetitions are combined in the following manner. For each imputed data set, we calculate the statistic of interest denoted as $\hat{Q}_{*l}$. The variance of $\hat{Q}_{*l}$ is denoted as U_{*l} The function for the estimated mean is

$$\overline{Q}_m = \sum_{l=1}^{m} \hat{Q}_{*l} \Big/ m \qquad (3)$$

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 within-repetition variances of this mean,

$$\overline{U}_m = \sum_{l=1}^{m} U_{*l} / m \qquad (4)$$

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

$$B_m = \sum_{l=1}^{m} (\hat{Q}_{*l} - \overline{Q}_m)^2 /(m-1) \qquad (5)$$

These two components are combined in the following manner:

$$T_m = \overline{U}_m + (1+m^{-1})B_m \qquad (6)$$

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) are

$$\nu_m = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B_m}\right]^2 \qquad (7)$$

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

$$\hat{\nu}_{obs} = (1-\gamma)\nu_0(\nu_0+1)/(\nu_0+3) \qquad (8)$$

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 were missing. The final degrees of freedom are

$$\nu_m^* = \left[\frac{1}{\nu_m} + \frac{1}{\hat{\nu}_{obs}}\right]^{-1} \qquad (9)$$

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 (equation 5) and is explicitly included in the estimate of the overall variance.

Description of Specific Imputation Techniques Used

The imputation methods used are divided into two main categories depending on the level of the variables being imputed: plot level or tree level. These two main categories are further broken down by which independent variables were used in the imputation and whether the imputation was done by matching or modeling.

Three different methods of imputing data at the plot level were used. The first method was imputing plot-level data by matching on just forest type. The second method was imputing plot-level data by grouping by forest types and matching on initial stand volumes. The third of these methods was grouping by forest type and matching by final stand volumes as predicted by SETWIGS (Bolton and Meldahl 1990). Whenever a donor plot was chosen, the donor-plot's tree list was read into the plot with missing data.

Four different methods of imputing tree-level data were used. With the tree-level imputation, three variables had to be imputed: diameter, volume, and mortality. Unfortunately, the current imputation software is not designed to impute correlated multivariate missing data. Therefore, each variable had to be imputed separately. The first method for imputing tree-level data was matching by species and on either initial diameter or volume, with final diameters and mortality being matched on initial diameter, and final volume being matched on initial volume. The second method was matching by species and on either diameter or volume like the previous method, except instead of matching on initial conditions, we matched on SETWIGS predicted values. The third method was modeling by species and on either initial diameter or initial volume. Because mortality is a discrete variable and is not very amenable to modeling by linear regression, mortality was imputed by matching on initial diameter. The final method was modeling by species on SETWIGS predictions for either diameter or volume, with mortality being imputed by matching on predicted percent mortality.

Because all of these imputation processes include randomization processes, each method was run five times and each run included five repetitions.

Tables

The different methods all generate sets of tree-level data. The performances of the different methods were compared using two of the volume tables generally provided in the State resource bulletins. Specifically, table 19 (volume by species and diameter class) and table 24 (volume by county and species type).

Comparison Statistics

The above referenced output tables 19 and 24 have several hundred entries each. Not only will each entry have its own sum, but each entry will also have a separate variance. To condense this information into something recognizable, we used the sum of the ratios of predicted mean squared error to observed variance. The formula for the sum of the ratios of mean squared error is

$$\sum_{i=1}^{n}\left\{\left[\frac{(\overline{Q}_i - X_i)^2}{Var_i}\right]+\frac{T_i}{Var_i}\right\}\Big/ n \qquad (10)$$

where i is the index for the table entry, Q is the average predicted sum per imputation run, X is the observed full data set sum, T_i is the variance of the imputed mean found in equation 6, and Var_i is the observed full data set variance of the sum found using the full 419 plots.

An analysis of variance was run on each of these variables, including a Tukey mean separation technique. Both tables have cells with observed values of zero. Since these cells also have variances of zero, they were deleted from the analysis. The results of the analyses of variance appear below.

RESULTS

The ANOVA showed significant differences ($p < 0.0001$) in the performance of the different methods for estimating table 19, the species by diameter class table (table 1). Imputing tree-level variables by matching on predicted values performed best. Imputing the plot-level variables also performed well. Imputing the tree-level variables by matching performed the worst.

Table 1.—*Imputation results for the species by diameter class table*

Model	Mean square error ratio[a]
Tree: modeling on predicted values	1.42 a
Plot: matching on initial values	2.31 ab
Plot: matching on predicted values	2.33 ab
Plot: matching by forest type	2.51 ab
Tree: modeling on initial values	2.91ab
Tree: matching on predicted values	4.36 b
Tree: matching on initial values	4.40 b

[a] Treatments with the same letters are not significantly different from each other.

The ANOVA also showed significant differences ($p < 0.0001$) in the performance of the different methods for estimating table 24, the county by species group table (table 2). Imputing tree-level variables by matching on initial values performed best. Imputing tree-level variables by matching on predicted values also performed very well. Imputing the tree-level variables by matching also performed well. Imputing the plot-level variables by matching performed the worst.

The large values for the ratios for table 24 are due to occasional counties that have very similar stands in the panel data. This causes the observed variances for those counties to be very small. Therefore, any increase in variation caused by the estimation process will cause very large values to appear.

Table 2.—*Imputation results for the county by species group table*

Model	Mean square error ratio[a]
Tree: modeling on initial values	10.96 a
Tree: modeling on predicted values	11.17 a
Tree: matching on predicted values	12.00 a
Tree: matching on initial values	12.42 a
Plot: matching on initial values	213.66 b
Plot: matching on predicted values	224.45 b
Plot: matching by forest type	309.89 b

[a] Treatments with the same letters are not significantly different from each other.

DISCUSSION

The plot-level imputation methods performed well for the species by diameter distribution table, but not for the county by species group table. The tree-level matching imputation methods did not perform as well as the other methods for the species by diameter distribution table. The tree-level modeling using the growth model projections as the independent variable performed best overall.

This study used only three-fifths of one panel of data to predict 1-year changes. The proposed implementation would use the three most recently measured panels of data to predict 5-year changes (to update the two oldest panels). The rankings of the imputation methods may not remain the same for the proposed implementation. Increasing the prediction interval from 1 year to 5 years will probably cause an increase in the variances for all of the imputation methods. However, each of the imputation methods has its own additional sources of variability that will be affected differently by changing from the scenario in this study to the proposed implementation. We suspect that the modeling methods are likely to be affected more by the change to a 5-year interval than the matching methods.

Increasing the time interval to 5 years will increase the number of plots per county. This should decrease the number of counties with very small observed variances. But it is not yet clear how much better these methods will perform for the county tables. Some caution will be needed when trying to analyze small groups of plots like counties or small ownership groups.

LITERATURE CITED

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. Sta. Bull. 603. Auburn, AL: Auburn University, Alabama Agricultural Experiment Station. 51 p.

Gartner, D.; Reams, G. 2001. A comparison of several techniques for estimating the average volume per acre of multipanel data with missing panels. Gen. Tech. Rep. SRS-47. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station: 76-81.

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

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.

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

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