The Effect of Data Quality on Short-term Growth Model Projections

David Gartner[1]

Abstract.—This study was designed to determine the effect of FIA's data quality on short-term growth model projections. The data from Georgia's 1996 statewide survey were used for the Southern variant of the Forest Vegetation Simulator to predict Georgia's first annual panel. The effect of several data error sources on growth modeling prediction errors was determined, including the effect of site index measurement errors. The study suggests that for tree attributes, such as volume by species-diameter class combinations, data quality will be the largest source of prediction error. For plot attributes, site index measurement errors will be the largest source of prediction error.

With the change from a periodic statewide survey to the current rotation panel system, a method of combining the data from several panels into a single estimate is needed. The current official statistic is the moving average. However, the moving average will be biased in the presence of a linear trend. Therefore, an alternative that will reduce this bias is needed. One of the alternatives being considered by the Southern Station is imputation. Previous short-interval studies (Gartner and Reams 2002) have suggested that using growth model projections will improve the imputation results. However, growth model projection errors will be incorporated into the variance of imputation results. This stimulated my interest in growth model projection errors.

Research on the propagation of measurement errors in the input data through the growth projection process has found that site index measurement errors created some of the largest variations in the predicted values (Gertner and Dzialowy 1984, Mowrer and Frayer 1986). Since I did not have much confidence in our site index estimates, I decided to empirically estimate the amount of prediction error due to different measurement errors, including site index measurement errors.

Methods

Data

The data from Georgia's 1996 statewide survey were used to predict Georgia's first annual panel. The site indices from the first panel were used in the growth model. Only plots that were completely within one condition class were used. Plots that had been harvested during the time between measurements and plots that had no trees or saplings were not used. This left 369 plots and over 9,000 trees. Even though the surveys were about 2 years apart, the actual elapsed time between measurements ranged from 0.1 to 3.6 years.

Model

The Southern variant of the Forest Vegetation Simulator (FVS) (Donnelly *et al.* 2001) from the Forest Service's Forest Management Service Center in Fort Collins, Colorado, was used to make the predictions. To incorporate the effects of the different elapsed times between measurements, predictions were made for 1, 2, and 3 years. Then the changes predicted by the growth model were multiplied by the actual elapsed time divided by the number of years in the growth projection. For example, for the plot with 3.6 years of actual elapse time, the growth model projected changes were multiplied by 3.6 and then divided by 3.0.

Effects

The study involved: 1) using the FVS growth model with the site index estimate from the first panel, 2) removing the effects of tree damage on tree growth, 3) eliminating some apparent diameter and height data problems, and 4) rerunning the FVS growth model with a range of site indices to determine which site index minimized the residual sum of squares for individual tree volumes for each plot.

Damaged trees were taken to be outliers in terms of the growth model's behavior. That is, the growth model was designed to predict growth that is uninterrupted by exogenous damage.

[1] Mathematical Statistician, U.S. Department of Agriculture, Forest Service, Southern Research Station, Knoxville, TN 37919. Phone: 865–862–2066; e-mail: dgartner@fs.fed.us.

Damaged trees had their observed values set to their predicted values. About 4 percent of the trees had signs of damage.

To determine whether a tree had questionable data, I created an acceptance region for diameter and height growth that ranged from a maximum growth plus a measurement error to zero growth minus the measurement error. The maximum growth rates were determined by running the North Carolina State University pine plantation growth model for loblolly pine. I used 600 trees per acre and the highest site index permitted by the software, which happened to be 99 feet base at age 25 years. Then I multiplied the maximum growth rates of the quadratic mean diameter and the dominant height by 1.5. This produced a maximum diameter growth rate of 1.135 inches per year and a maximum height growth rate of 6 feet per year. I took the diameter measurement error to be 0.5 inches and the height measurement error to be 15 feet. Trees with growth data outside this region had their observed values set to the predicted value. Less than 2 percent of the trees fell outside the acceptance region

To determine the amount of growth model prediction error due to site index measurement error, I searched potential site index values to determine the site index that minimized the sum squared error for tree volume estimates for each plot. Because of the difficulty in adapting a growth model for use in standard optimization routines, I resorted to a grid search. The site indices were varied in 1-percent increments from 55 percent to 200 percent of their panel 1 values. Not all plots reached their minimum in this range, but the residual sum of squares for these plots was less than 2 percent of the total residual sum of squares for the "optimal" site indices.

The sum squared differences between the values for the 1996 statewide survey and the first panel were calculated for diameter growth, height growth, individual tree volume growth, basal area growth, plot volume growth, and plot mortality. For the tree variables, only the surviving trees were used. These sums of squares were not corrected for the means. The sizes of the effects were measured as the percent reduction in the sum of squares.

Results

Tree Variables

The growth model did a better job at predicting diameter growth than height growth or volume growth (fig 1). The effect of site index measurement errors was only 5 percent of the total growth sum of squares for diameters, about 12 percent for height, and about 9 percent for volume.

The reduction in the residual sum of squares caused by editing out probable diameter and height data errors ranged from 10 to 18 percent. For diameters and tree volumes, the diameter and height errors contributed twice the sum of squares of the site index measurement errors. However, the height data errors contributed less to the height error sum of squares than the site index measurement errors.

Plot Variables

The growth model predicts basal area growth and plot volume growth well, around 48 percent of the sum of squares for each, but not for mortality (fig. 2). The growth model predictions

Figure 1.—*Contributions of the growth model, damaged trees, data errors, and site index to the percent growth sum of squares for diameter, height, and individual tree volume.*

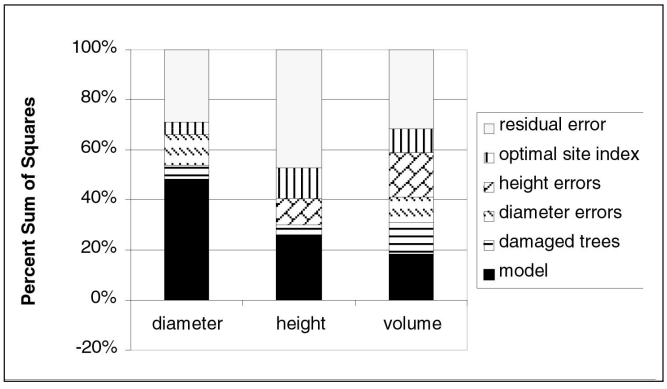
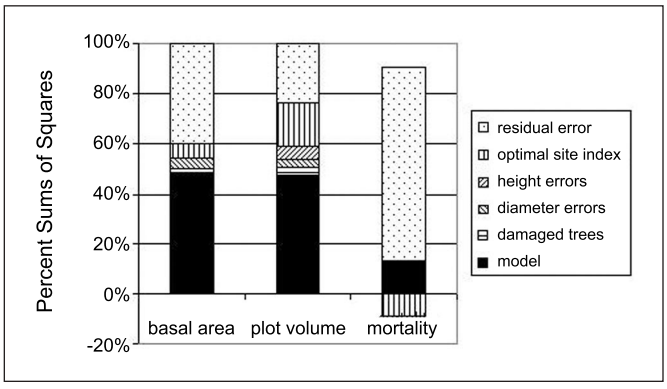


Figure 2.—*Contributions of the growth model, damaged trees, data errors, and site index to the percent growth sum of squares for basal area per acre, volume per acre, and mortality.*



using the standard site index estimates reduced the mortality sums of squares by about 16 percent. But using the optimal site index caused an 11-percent increase in the sums of squares.

The effect of site index measurement errors reaches almost 17 percent for plot volume. The effects of the diameter and height data errors are much smaller for the plot variables than for the tree variables.

Discussion

The site index that minimizes the sum squared error for individual tree volume is a function of not only the true site index, but also the growth model. Therefore, the optimal site index in this study may not be the site index as measured in the field. So the effects of the optimal site index should be considered a maximum attainable result. Decreasing the site index measurement errors by increasing the number of trees used to estimate the site index may not reduce the sum squared error shown here.

Also, the data used averaged only 2 years apart. This study may need to be repeated when data 5 years apart become available.

Dave Hyink[2] noted that other versions of the FVS model have given unusual mortality predictions with the default parameters. His experience showed that resetting the maximum-potential-basal-area parameter greatly improved the mortality predictions. This suggests that the mortality prediction function can be easily improved.

The data used for this study were measured before some of the new national data standards were implemented. One of these standards in particular requires field crews to electronically flag any observations of trees that lose more than 0.5 inches in diameter. If this new standard can prevent accidental recordings of reductions in diameter, then most sums of squares for questionable diameter data will become part of the model sum of squares. The Southern Station FIA unit's data acquisition team is implementing a similar data error check for height measurements.

Growth prediction errors are only some of the errors that will contribute to the variance of imputation results. Roesch's (1999) simulation study suggests that the greatest source of additional variation will be associated with predicting harvesting and conversions from forest to nonforest. We currently don't have any good models for predicting harvesting rates and intensities.

The long-term goal is to determine the different sources of error that contribute to the variances of imputation results and to determine the tradeoffs available to reduce these sources of variance. This study suggests that for tree attributes, such as volume by species-diameter class combinations, data quality will be the largest source of prediction error. The data acquisition band is already working on this problem. For plot attributes, site index measurement errors will be the largest source of prediction error. For combinations of plots, predicting harvesting rates and intensities will become the largest source of prediction error. This study is a small first step in determining the different sources of error that contribute the variances of imputation results, and the tradeoffs available to reduce these sources of variance.

Literature Cited

Donnelly, Dennis; Lilly, Barry; Smith, Erin. 2001. The Southern variant of the forest vegetation simulator. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Forest Management Service Center. 61 p.

Gartner, David; Reams, Gregory. 2002. A comparison of several techniques for estimating the average volume per acre for multipanel data with missing panels. In: Reams, Gregory A.; McRoberts, Ronald E.; Van Deusen, Paul C., eds. Proceedings, 2nd Annual Forest Inventory and Analysis symposium; 2000 October 17–18; Salt Lake City, UT. Gen. Tech. Rep. SRS-47. Asheville, NC: U.S. Department of Agriculture. Forest Service, Southern Research Station. 76–81.

Gertner, George Z.; Dzialowy, Paul J. 1984. Effects of measurement errors on an individual tree-based growth projection system. Canadian Journal of Forest Research. 14: 311–316.

Mowrer, H.T.; Frayer, W.E. 1986. Variance propagation in growth and yield projections. Canadian Journal of Forest Research. 16: 1196–1200.

Roesch, Francis. 1999. Mixed estimation for a forest survey sample design. In: Proceedings of the Section on statistics and the environment of the American Statistical Association.

[2] David Hyink, personal communication, Senior Scientific Specialist, Forest Resources Research and Engineering, Weyerhaeuser Company, Tacoma, WA 98477.