

2 FIM

## Evaluating Multiple Imputation Models for the Southern Annual Forest Inventory

Gregory Reams, Joseph McCollum, USDA Forest Service  
Gregory Reams, USDA Forest Service, P.O. Box 2680, Asheville, NC 28802

**Key Words:** Forest inventory, annual survey, interpenetrating panels, imputation.

**Abstract:** The USDA Forest Service's Southern Research Station is implementing an annualized forest survey in thirteen states. The sample design is a systematic sample of five interpenetrating grids (panels), where each panel is measured sequentially. For example, panel one information is collected in year one, and panel five in year five. The area representative and time series nature of the sample design offers increased flexibility in providing estimates of annual growth, mortality, removals and change in forest area. Restricting analyses to the most recently measured panel results in many missing cells in standard forest inventory tables. Rather than treat all unmeasured panels as missing, imputed values are used to update plots in each unmeasured panel. Because it is uncertain what analyses the ultimate users of these public-use data may engage in, we evaluate the effect and consequences of excluding important predictor variables from the imputation models.

### INTRODUCTION

Missing values are a problem in many survey data sets. In forest inventory data sets it has been common practice to update missing or deficient data using a variety of modeling techniques. Deficient data in the forest inventory context most often means old data. "Old" does not necessarily relate to well defined time intervals since observation; it could also relate to certain catastrophic events such as hurricanes or wildfires that have occurred since the completion of data collection. In either case, the representiveness of the data is in question.

The USDA Forest Service's Southern Forest Inventory and Analysis (FIA) program is implementing a new sample design. For those familiar with the longstanding 10-year periodic FIA survey, the new design takes the large periodic survey and divides it into five interpenetrating smaller samples, referred to as panels 1 through 5. Panel 1 plots are measured in year 1, panel 2 plots in year 2, etc., such that all plots are measured by the end of year 5. Within each panel the same sample elements (plots) are measured on each succeeding occasion. Currently panel remeasurement occurs on a five year interval.

The chief advantage of the annually repeated survey over the traditional periodic sample is that the separate annual samples provide information about variations that occur between the periods (Reams and Van Deusen 1999). This provides the ability to estimate annual and secular trends. Also, the sum of repeated surveys over the entire period can lead to better statistical inference than a single, concentrated, one-time survey (Kish 1965).

The annual panel design results in one panel of current information, and four panels of data varying in age from 2 to 4 years. We could treat all data as current, however we could also update the values through the use of models or imputation procedures. Forest inventory experts have commonly used models to update old or missing data. Usually regression models are used and the data are then treated as actual, and the inventory estimates are produced (Reams and Van Deusen 1999). This procedure is also known as single imputation, and the procedure as commonly used results in negatively biased estimates of the variance (Reams and Van Deusen 1999).

Public-use (shared) data bases are analyzed by many ultimate users with varying degrees of statistical expertise and computing power, and with different scientific questions and objectives. For example USDA Forest Service, Forest Inventory & Analysis (FIA) data are used for a variety of purposes. The FIA information is most often used as the official forest inventory statistics for the nation.

Because FIA has been operating for nearly 70 years, the data are often identified as the only long-term inventory and monitoring data available to the public for addressing long-term productivity trends related to natural and anthropogenic factors. As such, the FIA data have been used to address research questions never or infrequently considered by designers and users of the program. Recent research topics include the potential influence of air pollution or acid rain on forest productivity, and the potential influence forests may have on mitigating global climate change. Other studies include the long-term sustainability of southern forests as influenced by competing land-use alternatives and the influence of global climate change on forest productivity.

The reason for discussing the traditional use of FIA data for current status and trends in forest inventory and potential uses of FIA for exploratory analyses revolves around the use of imputation for annualized forest inventories. We will demonstrate two methods of imputation (modeling) for the annual survey program. We will also state that the use of modeled or imputed data to conduct associational and exploratory studies for quantitative assessment of biological and ecological processes should be discouraged.

The objectives of this study are:

1. Compare multiply-imputed estimates of inventory, when imputing 20%, 40%, and 60% of the data. We use merchantable cubic-foot volume per acre as our variable of interest.
2. Compare several methods for executing the imputations. The predominant comparison is between regression model based imputations and implicit categorical X matching models. Another comparison, that of using a model with high  $r^2$  versus a model with low  $r^2$  for imputations is presented.
3. Demonstrate that using imputed data for investigating and developing associational relationships should be avoided if possible.

#### METHODS

We use merchantable cubic-foot volume per acre of natural loblolly pine stands as our inventory variable of interest. It is known that cubic-foot volume per acre is a function of stand density, site and age (Schumacher 1939, Schumacher and Coile 1960). Using FIA data from the most recent survey of central Georgia (Thompson 1998) to model cubic-foot volume per acre (Y) resulted in the regression model  $\hat{Y} = a + b(\text{basal area/acre}) + e_i$  with an  $r^2 = .88$ . We used this model to predict values of cubic-foot volume per acre for regression based imputations. We also fit the model  $\hat{Y} = a + b(\text{site}) + e_i$  with an  $r^2 = .32$ , to predict values of cubic-foot volume per acre to evaluate whether there was a practical difference between inventory estimates when using a model with high predictive capability and one that is relatively low. Site as defined in the above model is a classification of each forest plot in terms of inherent capacity to grow crops of industrial wood.

The most precise matches for imputing cubic-foot volume per acre should be based on matching basal area/acre, stand age, and site productivity. However, we could not use the model  $Y = f(\text{basal area/acre, stand age, and site})$  to compare the explicit (regression) method to implicit model matching procedures. This is because when the observed sample is sampled without replacement for matching donor records to recipients the donor distributions are quickly depleted. We

choose to make the matching coarser by dropping age and site as independent variables. Because cubic-foot volume per acre can be modeled with a simple linear regression on basal area per acre with an  $r^2 = .88$ , little is lost with model performance.

#### EXPLICIT IMPUTATION MODELS WITH UNIVARIATE Y and COVARIATES

A common procedure in forest inventory is to update sample plots at either the individual tree or whole plot level. The most common method of predicting univariate  $Y_i$  from a set of predictors  $X_i$  is the normal linear regression model. Using the notation of Rubin (1987),

$$Y_i \sim N(X_i \beta, \sigma^2)$$

is the specification for  $f(Y_i | X_i, \theta)$ ,  $\theta = (\beta, \log \sigma)$ ,  $\beta$  is a vector of  $q$  components and  $\sigma$  is a scalar.

The posterior distribution of  $\theta$  involves only the units of  $Y_i$  observed. Standard Bayesian calculations as described by Box and Tiao (1973), with the normal linear model result in *a posteriori*,  $\sigma^2$  is  $\hat{\sigma}_1^2 / (n_1 - q)$  divided by a  $\chi^2$  random variable with df  $n_1 - q$ , where  $n_1$  is the number of  $Y_i$  observed. Also  $\beta$  given  $\sigma^2$  is normal with mean  $\hat{\beta}_1$  and variance-covariance matrix  $\sigma^2 V$ , where, in terms of the usual least-squares statistics based on  $n_1$  vectors  $(Y_i, X_i)$ , where  $i$  is an element of the observed sample, results in

$$\hat{\sigma}_1^2 = \sum_{obs} (Y_i - X_i \hat{\beta}_1)^2 / (n_1 - q)$$

$$\hat{\beta}_1 = V \left[ \sum_{obs} X_i^t Y_i \right]$$

where

$$V = \left[ \sum_{obs} X_i^t X_i \right]^{-1}.$$

Since we have described the posterior distribution of  $\theta$  in terms of standard distributions from which we can draw, the estimation task is complete.

The imputation task for this model follows:

1. Draw a  $\chi^2$  random variable with df  $n_1 - q$ , say  $g$ , and let

$$\sigma_*^2 = \hat{\sigma}_1^2 (n_1 - q) / g.$$

2. Draw  $q$  independent  $N(0,1)$  variates to create a  $q$ -component vector  $Z$  and let

$$\beta_* = \beta_1 + \sigma_* [V]^{1/2} Z,$$

where  $[V]^{1/2}$  is a square root of  $V$  such as the triangular square root obtained by Cholesky factorization.

3. Draw the  $n_0$  values of  $Y_{\min}$  as

$$Y_{i*} = X_i \beta_* + z_i \sigma_*,$$

where the  $n_0$  normal deviates  $z_i$  are drawn independently.

A new imputed value for  $Y_{\min}$  is initiated by drawing a new value of the parameter  $\sigma_*^2$ . Thus if  $m$  repeated imputations are desired, these three steps are repeated  $m$  independent times. Donor values for recipient  $X_i$  come from the most recently measured panel and are matched using hot-deck procedures as documented in the following section. A list of donors for each county is developed from all current year plots collected throughout the FIA survey work unit. Plots within the county are excluded from the county donor list.

### IMPLICIT IMPUTATION WITH HOT-DECK

We used a matching procedure similar to the Census Bureau's hot-deck procedure (Sande 1983). The procedure uses categorical  $X$ 's (for our study, natural loblolly pine stands and basal area/ac.) to match donors to recipients and the donors then give their values to the recipients. If more than  $m$  respondents are matches, then a subsample of  $m$  respondents can be drawn without replacement. If less than  $m$  matches are found, then one or more of the  $X$ -variables are made coarser. The choice of matching fields in either sequential or random choice procedures must be made considering likely sources of variation and the number of complete or eligible records available as potential donors. If too many fields are used for matching, there is the risk of a poor match in the imputed records (Sande 1982).

We compared matching 3 and 4 categories of basal area/acre. Classes for the 3 category matching are 0-60 square feet/ acre (sq. ft./ac.), 61-105 sq. ft./ac., and 105+ sq. ft./ac. Classes for the 4 category matching are 0-35 sq. ft./ac., 36-70 sq. ft./ac., 71-105 sq. ft./ac.

and 105+ sq. ft./ac.

For the site productivity class matching model we were restricted to using 2 classes. Site productivity class is defined as a classification of timber land in terms of inherent capacity to grow crops of industrial wood. The class identifies the average potential growth in cubic-feet/acre/year (trees 5 inches d.b.h. or larger to a 4-inch top) and is based on the culmination of mean annual increment of fully stocked natural stands. There are 6 observed classes for site productivity, where class 1 is the most productive, and class 6 the least productive. Because the distribution of observed site productivity classes is clumped into only several classes we were forced to collapse the observed classes into two classes. We defined class 1 as observed classes 1-4 and class 2 as observed classes 5 and 6.

With hot deck methods, the variance of the estimates in simple cases is known to be larger than the variance of the usual expansion estimates of means and totals. However, there may be a reduction in bias. Compared to some other methods of imputation, such as the use of normal linear regression models, hot deck methods should produce imputed data sets that appear more realistic and do a better job of reflecting distributional properties (Sande 1982).

Data sets created by explicit (regression models) and implicit (hot-deck matching) models are now ready for estimation of inventory means and variances. The method for doing this follows.

### IMPUTED MEANS AND VARIANCES

Let  $Q$  be the quantity of interest in the survey, for example, the mean  $Y$  in the population,  $\bar{Y}$ , and also that  $U$  is a statistic providing the variance. After generating  $m$  simulated-complete datasets and analyzing each of them as if they were genuine complete datasets we now have  $m$  estimates for  $Q$  and  $U$ , i.e.  $\hat{Q}_{*1}, \dots, \hat{Q}_{*m}$  and  $\hat{U}_{*1}, \dots, \hat{U}_{*m}$ .

The  $m$  repeated complete-data estimates and associated complete-data variances for  $Q$  is

$$\bar{Q}_m = \sum_{i=1}^m \hat{Q}_{*i} / m$$

which is the mean of means. The total variance of  $\bar{Q}_m$  is estimated by

$$T_m = \bar{U}_m + (1 + m^{-1}) \beta_m$$

where

$$\bar{U}_m = \sum_{i=1}^m \hat{U}_{*i} / m$$

is the average of the  $m$  complete-data variances, and

$$\beta_m = \sum_{i=1}^m (\hat{Q}_{*i} - \bar{Q}_m) (\hat{Q}_{*i} - \bar{Q}_m) / (m-1)$$

is the variance among the  $m$  complete-data estimates.

The results from using regression based and hot-deck based imputations are listed in tables 1-5. Using either 3 or 4 categories of basal area/acre for hot-deck matching produced nearly identical results (tables 1 and 2). The results listed in tables 1-5 indicate that both regression and hot-deck based methods produced nearly identical results for both means and variances. The similarity of the two methods is somewhat expected because the Bayesian regression model explicitly incorporates the variability of the observed sample for each of the predicted values. Use of regression models that do not include the addition of a residual value ( $e_i$ ) for each predicted mean value will result in a biased under-estimate of the variance.

Table 1. Hot-deck imputations of loblolly pine cubic-foot volume per acre for one (20%), two (40%), and three (60%) panels using  $m=3,4$  and 5. Matches are based on four basal classes, 0-35 sq. ft./ac., 36-70 sq. ft./ac., 71-105 sq. ft./ac., and 105+ sq. ft./ac.

	<u>m=3</u>	<u>m=4</u>	<u>m=5</u>
	-----20%-----		
mean	1707.20	1708.59	1711.10
s.e.	647.98	560.52	501.89
	-----40%-----		
mean	1699.59	1700.27	1699.87
s.e.	632.12	548.73	489.94
	-----60%-----		
mean	1712.02	1709.40	1709.10
s.e.	639.99	533.55	497.00

Table 2. Hot-deck imputations of loblolly pine cubic-foot volume per acre for one (20%), two (40%), and three (60%) panels using  $m=3,4$ , and 5. Matches are based on three basal classes, 0-60 sq. ft./ac., 61-105 sq. ft./ac., and 105+ sq. ft./ac.

	<u>m=3</u>	<u>m=4</u>	<u>m=5</u>
	-----20%-----		
mean	1704.12	1702.89	1706.15
s.e.	657.79	567.97	508.37
	-----40%-----		
mean	1708.80	1701.19	1696.26
s.e.	642.35	552.38	494.76
	-----60%-----		
mean	1691.69	1692.33	1695.62
s.e.	653.49	565.96	505.85

Table 3. Hot-deck imputations of loblolly pine cubic-foot volume per acre for one (20%), two (40%), and three (60%) panels using  $m=3,4$ , and 5. Matches are based on two site productivity classes, 20-84 cu. ft./ac./year, and 85+ cu. ft./ac./year.

	<u>m=3</u>	<u>m=4</u>	<u>m=5</u>
	-----20%-----		
mean	1672.26	1680.39	1687.69
s.e.	640.27	558.62	499.66
	-----40%-----		
mean	1719.26	1719.20	1722.56
s.e.	626.61	544.28	487.40
	-----60%-----		
mean	1732.99	1719.38	1716.38
s.e.	628.89	545.08	485.69

Table 4. Regression model based imputations of loblolly pine cubic-foot volume per acre for one (20%), two (40%), and three (60%) panels using  $m=3,4$ , and 5. Matches are based on four basal area classes, 0-35 sq. ft./ac., 36-70 sq. ft./ac., 71-105 sq. ft./ac., and 105+ sq. ft./ac.

	<u>m=3</u>	<u>m=4</u>	<u>m=5</u>
	-----20%-----		
mean	1720.06	1721.99	1723.14
s.e.	638.16	552.15	494.74
	-----40%-----		
mean	1702.58	1704.09	1702.69
s.e.	622.62	537.94	481.22
	-----60%-----		
mean	1732.78	1729.40	1731.05
s.e.	622.71	536.38	480.96

Table 5. Regression model based imputations of loblolly pine cubic-foot volume per acre for one (20%), two (40%), and three (60%) panels using m=3,4, and 5. Matches are based on three basal area classes, 0-60 sq. ft./ac., 61-105 sq. ft./ac., and 105+ sq. ft./ac.

	<u>m=3</u>	<u>m=4</u>	<u>m=5</u>
	-----20%-----		
mean	1697.10	1702.09	1703.29
s.e.	647.88	561.16	502.08
	-----40%-----		
mean	1700.72	1707.66	1707.16
s.e.	626.99	542.45	486.72
	-----60%-----		
mean	1687.41	1692.71	1695.70
s.e.	638.11	552.98	493.44

Results indicate that imputing 20%, 40% or 60% of the data produced minimal changes in the mean and variance. Increasing imputations from m=3 to m=5 resulted in reductions in the estimated variance. Imputations from a poorly fit model based on site productivity class and imputations from a well fit model based on basal area/acre produced similar estimates.

#### USING MODELED OR IMPUTED DATA FOR EMPIRICAL STUDIES

Users of FIA data often develop empirical models of biological processes. For example forest yield is often modeled as a function of stand density, age and site. Under the annual forest inventory there is the desire to create modeled or imputed data bases for unmeasured plots, especially for small area estimates. Modeling and imputation can generate clean-looking data bases that are easily used, and there is the possibility that users would consider using simulated data for empirical modeling (Schreuder and Reich 1998).

Use of modeled data sets for the development of empirical models should be discouraged. To illustrate this assertion, we modeled (imputed) data values for merchantable cubic-foot volume (MCFV) for three of the five annual panels (60% of the survey plots) in central Georgia.

Using the full data set (all five observed panels) we predicted MCFV as a function of basal area per acre for all trees greater than 5 inches diameter at breast height (dbh). Using the full data set values resulted in the following model:

$$MCFV = -221.62 + 26.42 (BA)$$

$$r^2 = 0.88, n = 474$$

where, MCFV is merchantable cubic foot volume per

acre, and BA is basal area per acre. The correlation matrix for MCFV and BA is:

$$\begin{vmatrix} 1 & 0.9396 \\ 0.9396 & 1 \end{vmatrix}$$

We simulated a full data set (all 5 panels) with 60 percent (3 annual panels) of the MCFV values predicted from the above model, and 40 percent observed data. The following model was fit to the 60 percent simulated and 40 percent measured data set:

$$MCFV = -318.08 + 27.49 (BA),$$

$$r^2 = 0.63, n = 474$$

The correlation matrix for MCFV and BA is:

$$\begin{vmatrix} 1 & 0.7944 \\ 0.7944 & 1 \end{vmatrix}$$

These results are as expected with modeled (imputed) data sets. That is, the empirical estimate of the variance-covariance matrix is often not preserved (Ek et al. 1997). When the objective is to produce empirical models for specific target populations it is best to use complete data sets for these studies. A potentially serious weakness of simulated data sets is that under different yet reasonable assumptions, very different data bases and conclusions can be reached by different researchers.

#### ESTIMATES OF FOREST INVENTORY

Comparing the ability of modeling(imputation) procedures to reproduce plot and tree level information on a sample element by element basis is not our goal. In the Southern Annual Forest Inventory System (SAFIS), the goal is to produce valid global statistical inference not optimal point predictions. Using multiply-imputed (MI) data sets for forming estimates of forest inventory is a more practical and defensible application, than is the development of empirical regression models of biological processes.

This is because multiple imputation takes into account the uncertainty of estimated data. This strength has been identified as weakness by some researchers since MI often leads to overly conservative estimates of the variance (Fay 1996). In our potential applications within SAFIS we view the overly conservative estimates of variance as an asset. This is because, even with

improper multiple imputation, confidence intervals generated are still valid because the actual coverage rates are higher than the claimed (nominal) confidence level (Rubin 1996, Judkins 1996).

### CONCLUSIONS

Multiply-imputed data sets created either through explicit regression modeling or by implicit hot-deck models appear to work about equally well for estimating current status of forest inventory variables, such as cubic-foot volume per acre. This statement assumes that the imputed data sets based on regression models have variability added back into the mean estimate for each predicted data point. Our Bayesian model did this, however users of imputed data should be aware that the use of regression models without provisions for injecting variability into predicted mean values will result in negatively biased variance estimates.

Imputed data sets that contained either 20 percent, 40 percent or 60 percent imputed data worked well for estimating cubic-foot volume per acre. Means of the imputed data sets ( $m=3, 4$  and  $5$ ) matched closely with the observed full data set. One standard error of the mean volume/acre is approximately 600-700 cu. ft./ac. for  $m=3, 4$  or  $5$  and for all models whether regression based or by categorical matches on important X variables such as basal area.

Imputed data sets created from explicit regression models should not be used for exploratory data analysis. Typically forest scientists use survey data to develop associational relationships among numerous variables. For example cubic foot volume of wood is often predicted by plot or stand basal area, stand age, and site productivity. We have demonstrated that imputation by an explicit regression model does not preserve the empirical correlation matrix between two independent variables. We recommend that associational relationships be developed from actual measured data.

### REFERENCES

- Box, G.E.P., Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*. Reading, Mass: Addison-Wesley.
- Ek, A.R., Robinson, A.P., Radtke, P.J., Walters, D.K. (1997), "Development and Testing of Regeneration Imputation Models for Forests in Minnesota," *Forest Ecology and Management*, 94,129-140.
- Fay, R.E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91,490-498.
- Judkins, D.R. (1996), "Comment: Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91,507-510.
- Kish, L. (1965), *Survey Sampling*, New York: Wiley.
- Reams, G.A., Van Deusen, P.C. (1999), "The Southern Annual Forest Inventory System", *Journal of Agricultural, Biological and Environmental Statistics*, 4(3),108-122.
- Rubin, D.R. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91,473-489.
- Sande, I.G. (1982), "Imputation in Survey: Coping With Reality," *The American Statistician*, 36(3),145-152.
- Sande, I.G. (1982), "Hot-Deck Imputation Procedures," in *Incomplete Data in Sample Surveys*, eds. W.G. Madow and I. Olkin, New York: Academic Press, pp.334-350.
- Schreuder, H.T., Reich, R.M. (1998), "Data Estimation and Prediction for Natural Resources Public Data," Research Note RMRS-RN-2, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.
- Schumacher, F.X. (1939), "A New Growth Curve and Its Application to Timber-Yield Studies," *Journal of Forestry*, 37,819-820.
- Schumacher, F.X., Coile, T.S. (1960), "Growth and Yield of Natural Stands of the Southern Pines," *T.S. Coile Inc.*, Durham, NC.
- Thompson, M.T. (1998), "Forest Statistics for Central Georgia," Resource Bulletin SRS-26, U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC.