Comparison of Imputation Procedures for Replacing Denied-access Plots

Susan L. King[1]

Abstract.—In forest inventories, missing plots are caused by hazardous terrain, inaccessible locations, or denied access. Maryland had a large number of denied-access plots in the latest periodic inventory conducted by the Northeastern Forest Inventory and Analysis unit. The denial pattern, which can introduce error into the estimates, was investigated by dropping the 1999 denied-access plots in the 1986 periodic inventory. The denied-access plots represented the population in terms of percentages of forest and nonforest, ownership, land use, and cubic-foot volume. Board-foot volume was less representative. Several single imputation group means—"Euclidean type" distance measures, multiple regression imputation, and listwise deletion with the adjustment of the stratum weights—are compared for estimating the missing cubic- and board-foot volume on forest land. Information on the forested condition of the denied-access plot can be found only through photointerpretation (PI) or satellite imagery such as Multi-Resolution Land Characteristics, (MRLC). Results were inconclusive following an examination of the standard and sampling errors for the state or the root mean square errors for the denied-access plots. As a result, 2 to 12 percent of the data in increments of 2 percent were dropped randomly in a simulation study; the missing plot attributes estimated using each technique. The best simulation study procedure for PI-based forest/nonforest stratification is PI stratum classification. The best simulation study procedure for satellite-based forest/nonforest stratification is the listwise deletion alternative.

In any forest inventory across multiple ownerships, forest crews are occasionally denied access to private and public land. Plots can also be inaccessible due to hazardous terrain and environmental conditions. One solution is to replace the denied or inaccessible plot with a new plot. This raises concerns about sampling strategies that might produce biased results when the ownership of the new plot differs from that of the denied plot, i.e., replacement alters the original sample design. A second solution for denied-access plots is to replace the missing values required for statistical tables with imputed values. This would allow the same plots to be kept over time in the event that permission for access is obtained during the next inventory cycle.

Another concern is the character of the denied-access plots. Do they differ from those in the population? Are there more forested denied-access plots than nonforested and are these more prevalent in a particular ownership? The latter concern is important because different ownerships influence how a forest is managed and grows. Error can be introduced into the estimate when a single group is over-represented. Answering these questions not only provides useful insight to the denied-access pattern, but also is necessary to determine whether the denied-access plots are missing at random. Denied-access plots introduce error into the survey estimates to the extent that they differ from the accessed plots. The reduced sample size also increases the variance of the estimates of the mean from the sample. As the proportion of denied-access plots increases, the bias and variance also can increase leaving the results open to questioning.

Maryland was selected for the study due to its high percentage of denied-access plots, 1.83 percent in the 1999 inventory, a large percentage for the Northeast. Various replacement techniques as well as listwise deletion were applied for cubic- and board-foot volume. Mean volume estimates on forest land and their corresponding standard and sampling errors provided comparison criteria, although estimates of forest land also were examined.

[1] U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA 19073. Phone: 610–557–4048; fax: 610–557–4350; e-mail: sking01@fs.fed.us.

Materials and Methods

Data

Both the 1999 and 1986 inventories of Maryland were used in this study. The 1999 inventory was used to identify which denied-access plots to drop when reprocessing the 1986 inventory. The 1986 inventory provided truth and was used to evaluate the imputation procedures and the nonresponse pattern.

To reduce the sampling error, the State population was divided into two pools: forested and nonforested. Since the denied-access plots were assumed to not have been visited, the classes of these plots are unknown. However, through photointerpretation (PI) or satellite imagery, a forest/nonforest classification is made. Satellite imagery will be used exclusively in the future for the Phase I stratification, but PI provides a comparison. Some of the difficulties encountered in using old data (1986 inventory) included having only the coordinate locations for 948 of the original 1,177 plots, having PI information for only 1,106 plots, and having only Landsat 7 satellite bands 2, 4, and 7.

The satellite imagery selected was a forest/nonforest map acquired from a National Land Cover data set (formerly Multi-Resolution Land Characteristics (MRLC)). This vegetation map was made by the U.S. Geological Survey EROS Data Center (Vogelman *et al.* 2001) and is based on 1992 Landsat 7 Thematic Mapper data and various other intermediate-scale spatial data were used as ancillary data.

For the forest/nonforest call, the MRLC was reclassified so that the forest classes as well as woody wetland received a value of 1 and other pixels received a value of 0. The plots were overlaid on the reclassified image to obtain a forest/nonforest call. The call for PI was based on the land-use classification using photos. For these data, PI had an overall accuracy of 94.4 percent compared to field plots, whereas the satellite forest/nonforest classification was only 80.17 percent as accurate as field plots. These percentages are similar to other data sets and indicate that PI is a more accurate forest/nonforest classifier.

A "5x5 sum" filter was passed over the reclassified forest/nonforest MRLC image. In the resulting image, each pixel represented the count of forested pixels inside a 5x5 window. The plots were overlaid on the filtered image to obtain a forested pixel call for each plot. The forested pixel count ranged from 0 to 25 and was divided into four classes (MRLC5 strata) based on a study using Connecticut data (Hoppus *et al.* 2001). The PI strata are the PI classifications.

One of the imputation procedures substitutes the plot with the nearest Euclidean or spectral distance. For the spectral data, plots were overlaid on bands 2, 4, and 7 from a satellite image created from early 1990s imagery that was composited and radiance- and terrain-corrected by the Earthsat Corporation. Earthsat distributed a band 7-4-2 subset of this original image via the National Aeronautics and Space Administration. The image was from Landsat 7 and had a resolution of 28.5 m. No filter was applied to this image because the original image had been geocorrected to an extremely high positional accuracy.

Imputation Techniques

1. Listwise Deletion: The 21 denied-access plots were dropped and the stratum weights adjusted to reflect the reduced sample sizes in affected strata. The sample size is reduced by the number of denied-access plots and the stratum weights are adjusted to reflect the reduced number of plots in the affected strata.
2. Group Mean Replacement: The denied-access plots were replaced with the State, County, PI, or MRLC5 stratum average or the stratum classification average. The denied-access plots were replaced with the group mean average for cubic- and board-foot volume from their respective forest/nonforest pool. The plot was assigned the forest/nonforest pool classification for both PI and satellite imagery. The sample size was not altered.
3. Replaced with Euclidean (PI) or spectral distance (satellite): On-the-ground classification for forest/nonforest, timberland, cubic-foot volume, and board-foot volume from the nearest plot was substituted for the missing values on the denied-access plot. Again, the sample size was not altered.
4. Statistical Multiple Imputation: The data have a monotone missing pattern because all the plot attributes are either present or missing. Consequently, either a parametric regression procedure, which assumes normality, or a non-parametric method, which uses propensity scores, is appropriate (Rubin 1987, SAS Institute Inc. 2001). The regression approach produced lower sampling errors, so the results for this procedure are reported. As with the group

mean replacement, the plot was assigned the forest and timberland value of the pool. The estimates were imputed separately both by forest/nonforest and by cubic- and board-foot volume. This allows maximum and minimum parameters to be set more accurately in PROC MI (SAS Institute Inc. 2001). The PI classes and the three satellite bands also were used as imputation variables. The previous methods are single-imputation procedures. In multiple imputation, the missing data are filled in *m* times, generating *m* complete data sets. The *m* complete data sets are evaluated by using standard statistical analysis and the *m* complete data sets are combined to produce a final result.

Double Sampling for Stratification

The sampling design used in 1986 was double sampling for stratification. There were four stratum classes in Maryland: forested and nonforested plots inventoried in 1976, and first-time inventoried forested plots and nonforested plots in the 1986 inventory. Each class had 18 land areas consisting of a county or a grouping of adjacent counties. Thus, there were 36 forested and 36 nonforested strata. The variable of interest is total or mean cubic-foot volume, board-foot volume, or forestland area. The variance equation was equation 12.12 (from Cochran 1963).

Multiple Imputation Equations

After *m* multiple imputations or *k* simulated complete data sets were generated and each was analyzed as if it were a complete data set, means and variances were combined to form a single estimate. The means can be averaged to form a single estimate but the variances must account for both the within-imputation variance (average of the complete data estimates) and between-imputation variance (see Allison 2002, Gartner and Reams 2002, Reams and McCollum 1998, and SAS Institute Inc. 2001).

Results

Similarity

One goal was to investigate whether denied-access plots differ from those in the general population. Of 1,177 plots in the 1986 inventory, 21 were denied-access plots. The population is 59.98 percent forested and 66.67 percent (14 plots) of the denied-access plots were forested. The denied-access plots are distributed across Maryland rather than a particular region of the State (fig. 1).

As mentioned earlier, different ownerships influence how a forest is managed and grows so the estimates can be biased when one group is overrepresented. With only 14 forested denied-access plots and more than three categories, the minimum cell count of five observations is not met and a Chi-square goodness-of-fit test cannot be performed. Not all the ownership categories are represented in the denied-access plots (fig. 2a). The largest category, Other Private Individuals, is 61 percent of the population and 57 percent of the denied-access sample; class percentages match closely given that there are only 14 plots.

Not all forest types are represented in the denied-access plots (fig. 2b), but the percentages are nearly identical because, again, there are only 14 denied-access plots. For both cubic-foot (fig. 2c) and board-foot (fig. 2d) volume, the extreme classes are not represented in the denied-access plots. The percentages of denied-access plots are more similar to those of the sample population for cubic-foot than for board-foot volume. Consequently, greater bias would be expected in the estimates.

Actual Results for Denied-access Plots

Table 1 shows the State results for PI and MRLC satellite imagery. Under the column for forest land, zero indicates the

Figure 1.—*This map shows both the 21 denied- and all nondenied-access plots in Maryland. The background is a forest/nonforest map and the denied-access plots are in all parts of the State on forest and nonforest land. The western portion of the State is more mountainous, thus it has more plots and more denied-access plots than the other regions.*

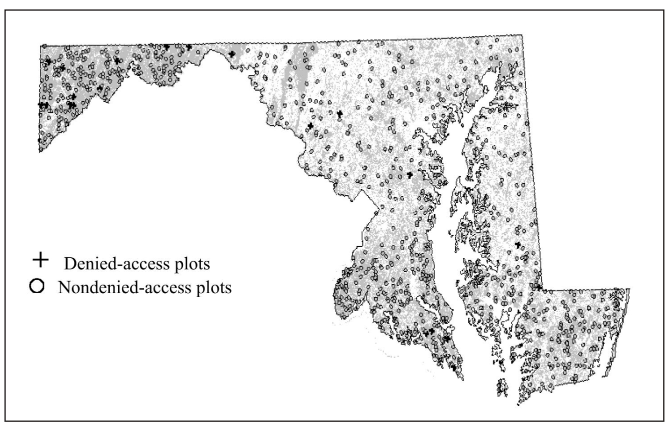

Figure 2.—*Comparison of sampled and denied-access populations by (a) ownership, (b) forest type, (c) cubic-foot volume, and (d) board-foot volume. The sampled population includes both denied and nondenied-access populations. The sampled and denied-access populations are less similar for board-foot volume.*

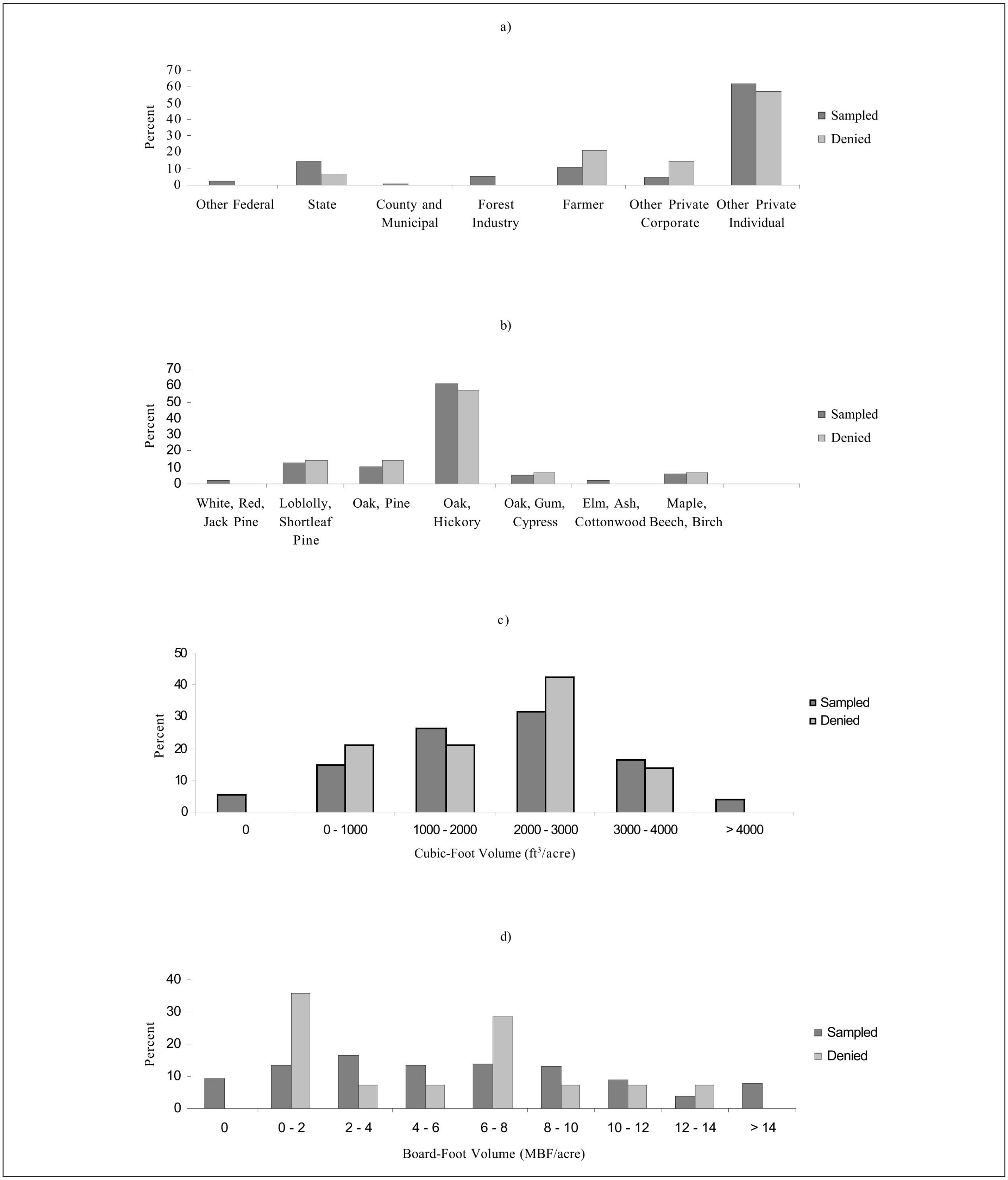

Table 1.—*State results for PI and MRLC satellite*

Procedure	PI cubic-foot volume				Satellite cubic-foot volume			PI board-foot volume			Satellite board-foot volume		
	Forest land	Mean volume	Standard error mean volume	Sampling error mean volume	Mean volume	Standard error mean volume	Sampling error mean volume	Mean volume	Standard error mean volume	Sampling error mean volume	Mean volume	Standard error mean volume	Sampling error mean volume
Complete data	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	727.60	20.40	2.80	727.60	20.40	2.80	2181.68	82.80	3.80	2181.68	82.80	3.80
Listwise deletion	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	726.62	20.47	2.82	726.62	20.47	2.82	2177.17	83.04	3.81	2177.17	83.04	3.81
State average	0	0.07	0.04	55.19	1.25	0.64	50.7	20.21	0.12	55.19	3.69	1.87	50.72
	1	728.08	20.41	2.80	725.26	20.43	2.82	2183.19	82.72	3.79	2173.64	82.72	3.81
County average	0	0.01	0.01	67.11	1.23	0.76	61.76	0.05	0.031	67.11	3.40	2.18	64.24
	1	727.54	20.42	2.81	726.03	20.43	2.81	2183.68	82.72	3.79	2176.38	82.70	3.80
PI or MRLC5 stratum average	0	0.05	0.04	86.65	1.44	0.67	46.53	0.15	0.13	86.65	4.23	1.96	46.40
	1	726.26	20.37	2.80	726.53	20.43	2.81	2175.72	82.62	3.80	2176.75	82.72	3.80
Stratum class-ification average	0	0.04	0.02	67.11	1.65	1.52	91.89	0.12	0.08	67.11	5.26	4.83	91.75
	1	727.68	20.42	2.81	726.37	20.43	2.81	2179.74	82.69	3.79	2177.31	82.71	3.80
Euclidean or spectral distance	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	727.60	20.40	2.80	727.79	20.49	2.82	2197.18	83.56	3.80	2180.16	82.91	3.80
Regression imputatioin	0	0.02	0.13	81.68	0.17	0.17	102.92	0.37	0.28	76.43	0.35	0.42	120.04
	1	729.84	20.54	2.82	726.38	20.54	2.83	2194.33	83.50	3.81	2184.45	82.94	3.80

results for nonforest land and one indicates results for forest land. Euclidean or spectral distances replace the denied-access plot with attributes from the nearest plot, increasing the likelihood that the forest/nonforest classification is correct. A correct classification would result in zero cubic- and board-foot volumes on nonforested land. The standard and sampling errors are very close for both PI and MRLC satellite imagery. Except for the Euclidean distance and regression imputation for estimating mean board-foot volume, mean volumes are close to the mean volume for the complete data. There is no best procedure. In the next step to determine the best procedure, root-mean square errors (RMSE) were calculated for the 21 denied-access plots by procedure. From table 2, the PI stratum average is the best procedure for both cubic- and board-foot volume; there is no superior procedure for the satellite imagery. The MRLC5 stratum average has the smallest RMSE, but the State, county, and stratum classification averages are close. Euclidean or spectral distance and regression imputation had the highest RMSE.

Results of Simulation Study

The next step was a simulation study to determine a winning procedure at the 2-percent denial rate and whether that best procedure might be different at higher rates. The denials were determined randomly without considering maintaining a forest-to-nonforest ratio of 60:40. Only plots with both coordinates and satellite bands 2, 4, and 7 were used so that all procedures at the same denial percentage eliminated the same plots. A different random number seed was used for each percentage for 100 simulations. The squared, absolute, and standard errors of the forested estimates were calculated for each procedure and then averaged by the stratum classification. For each stratum, the averaged results were sorted in ascending order by procedure. The procedure with the smallest comparison statistic was assigned as the best procedure for that stratum. The tabulated RMSE results for the PI and satellite-based simulation at different deletion percentages are shown in tables 3 and 4. For PI, replacement by the PI stratum average was the best procedure. For the satellite-based simulation, listwise deletion was the

Table 2.—*RMSE for 21 denied-access*

Plots Procedure	Cubic-foot volume PI	 Satellite	Board-foot volume PI	 Satellite
State average	931.46	1054.73	3460.85	3845.50
County average	909.32	1052.48	3455.14	3835.21
PI or MRLC5 stratum average	609.32	1033.88	2745.57	3753.63
Stratum classification average	918.85	1065.12	3495.95	3948.84
Euclidean or spectral distance	1326.85	1241.91	5918.58	4479.65
Regression imputation	1056.59	1100.59	4149.88	3974.99

Table 3.—*Number of times the RMSE for the PI-based simulation was best at different deletion percentages*

	Cubic-foot volume						Board-foot volume					
Procedure	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**
Listwise deletion	8	6	7	3	8	7	6	6	8	4	6	6
State average	3	7	6	6	5	9	2	6	6	7	5	6
County average	2	3	2	4	3	1	4	5	4	3	3	3
PI stratum average	13	13	16	16	13	16	15	12	12	13	16	16
Stratum classification average	4	4	3	2	1	2	4	2	2	2	2	2
Euclidean Distance	4	3	2	2	3	1	3	3	2	1	1	1
Regression Imputation	2	0	0	3	3	0	2	2	2	6	3	2

Table 4.— *Number of times the RMSE for the satellite-based simulation was best at different deletion*

	Cubic-foot volume						Board-foot volume					
Procedure	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**
Listwise deletion	20	21	22	21	21	23	20	20	23	24	24	21
State average	0	0	0	1	2	0	2	1	1	2	1	2
County average	6	4	5	5	4	4	5	6	4	2	3	3
PI stratum average	6	7	7	6	9	8	8	6	5	5	7	10
Stratum classification average	0	1	1	1	0	1	1	1	2	2	0	0
Euclidean Distance	1	0	0	0	0	0	0	0	0	0	0	0
Regression Imputation	3	3	1	2	0	0	0	2	1	1	1	0

best procedure with the lowest RMSE in more than 20 of the 36 strata. The best procedures were the same for the other two comparison statistics: absolute and standard errors.

Another comparison entails tabulating by imputation procedure the number of estimates for mean volume (of 3,600) that exceed the 95-percent confidence interval of the mean volume of the total sample estimate. The results for PI and satellite-based forest/nonforest classification are shown in tables 5 and 6. The best procedure has the lowest number of entries. For PI, the best procedure is the PI strata average. However, listwise deletion gave the same results for board-foot volume. Listwise deletion is the best procedure for satellite imagery.

Conclusion

The denied-access population in Maryland was similar to the sample population in terms of the amount of forested land,

Table 5.— *Number of times an estimate from the simulation fell outside the 95-percent confidence interval for the PI-based simulation at different deletion percentages; 180 estimates would be expected to fall outside the confidence interval*

	Cubic-foot volume						Board-foot volume					
Procedure	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**
Listwise deletion	0	0	1	0	3	0	0	0	0	0	1	0
State average	0	0	0	1	0	4	0	0	0	1	1	1
County average	0	0	0	1	1	3	0	0	0	1	0	0
PI stratum average	0	0	0	0	0	2	0	0	0	0	1	0
Stratum classification average	0	0	0	2	0	3	0	0	0	1	1	0
Euclidean Distance	1	5	11	12	8	14	0	1	7	6	10	17
Regression Imputation	0	1	1	3	1	6	0	1	0	5	3	4

Table 6.— *Number of times an estimate from the simulation fell outside the 95-percent confidence interval for the satellite-based simulation at different deletion percentages; 180 estimates would be expected to fall outside the confidence interval*

	Cubic-foot volume						Board-foot volume					
Procedure	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**
Listwise deletion	0	0	1	0	3	0	0	0	0	0	1	0
State average	1	6	7	16	28	36	0	0	1	4	8	6
County average	2	4	16	29	36	38	1	2	6	15	15	20
PI stratum average	0	1	1	7	10	15	0	0	0	1	3	2
Stratum classification average	0	2	3	16	14	15	1	5	5	13	17	15
Euclidean Distance	22	46	61	104	124	164	20	37	53	82	109	131
Regression Imputation	3	13	26	41	44	51	2	6	11	21	29	40

geography, ownership, forest type, and cubic-foot volume. The two groups were less similar for board-foot volume. Results for estimating forest land are not presented, but they were the least biased since the denied-access forested vs. nonforested land closely matched the sample population.

The State results are inconclusive and the results must be analyzed at the stratum level or by examining the denied-access plots to differentiate between procedures. PI-based forest/nonforest classification is superior to satellite-based classification as evidenced by lower standard and sampling errors as well as lower RMSE. This is not surprising since PI-based forest/nonforest classification more nearly agrees with plot-based estimates than satellite-based classification.

Regression imputation and Euclidean or spectral distance had the poorest performance. The effectiveness of regression imputation depends on the relationship between the available independent variables and the variable to be imputed. The PI classes and satellite bands did not capture the relationship with either cubic- or board-foot volume. The plots are 2.7 km apart and land use and history between nearest neighbor plots can vary greatly. As a result, the Euclidean distance technique was ineffective. Surprisingly, the spectral distance also failed to capture the relationship with cubic- or board-foot volume. Also, the best procedure did not vary by deletion percentages but depended on whether PI or satellite-based classification was used to classify the forested/nonforested condition of the plots. For PI-based stratification, the winning procedure is the group mean replacement, PI stratum classification, whereas for satellite-based stratification, the winning procedure is the listwise deletion alternative with the adjustment of the stratum weights.

Literature Cited

Allison, Paul D. 2002. Missing data. Thousand Oaks, CA: Sage University Papers on Quantitative Applications in the Social Sciences, 07-136. Sage Publications. 93 p.

Cochran, William G. 1963. Sampling techniques. 2d ed. New York, NY: John Wiley and Sons. 428 p.

Gartner, D.; Reams, G.A. 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, 2d 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.

Hoppus, M.; Arner, S.; Lister, A. 2002. Stratifying FIA ground plots using a 3-year old MRLC forest cover map and current TM delivered variables selected by “decision tree” classification. In: Reams, Gregory A.; McRoberts, Ronald E.; Van Deusen, Paul C., eds. Proceedings, 2d 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: 19–24.

Reams, G.A.; McCollum, J.M. 1998. The use of multiple imputation in the Southern annual forest inventory system. In: Hansen, Mark; Burk, Tom, eds. Integrated tools for natural resources inventories in the 21st century; 1998 August 16–20; Boise, ID. Gen. Tech. Rep. NC-212. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 228–233.

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

SAS Institute Inc. 2001. SAS/STAT® software: changes and enhancements, Release 8.2. Cary, NC: SAS Institute, Inc. 324 p.

Vogelman, J.E.; Howard, S.M.; Yang, L.; *et.al.* 2001. Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering and Remote Sensing. 67(3): 650–661.