The Use of Multiple Imputation in the Southern Annual Forest Inventory System

Gregory A. Reams and Joseph M. McCollum

Abstract.—The Southern Research Station is currently implementing an annual forest survey in 7 of the 13 states that it is responsible for surveying. The Southern Annual Forest Inventory System (SAFIS) sampling design is a systematic sample of five interpenetrating grids, whereby an equal number of plots are measured each year. The area representative and time series nature of the SAFIS plot design offers increased flexibility in analyzing the data for both large- and small-domain means. Users of FIA information are often interested in the estimates of small-domain means, at the multi-county or FIA survey unit level. Restricting analyses to the most recently measured annual panel results in many missing cells in standard inventory tables. Rather than treat the four unmeasured panels as missing, imputed values are used to update plots in all panels. An initial set of rules and solutions for imputing are provided for SAFIS.

The Forest Inventory and Analysis (FIA) unit of the USDA Forest Service's Southern Research Station is responsible for providing inventory estimates for 13 southern states. Over the last 50 years, the states relied on an approximate 10-year periodic survey, and this system worked well in the past, when resource assessments were less dynamic than today. However, in today's world, the periodic 10-year survey is perceived to be accurate for several years but increasingly unreliable thereafter.

In response to the need for more timely and accurate inventory data, the Southern Annual Forest Inventory System (SAFIS) was initiated in 1997. The SAFIS sampling design is a modification of the periodic design; instead of measuring all plots in 1 to 2 years, an equal number of plots is measured each year, with a plot remeasurement cycle of 5 years. This results in a systematic sample of five interpenetrating grids, and each annual grid is in statistical terms a complete annual panel because the same sample elements (plots) are measured every 5 years (fig. 1).

The expansion factor for each SAFIS plot is 5,760 acres over the 5-year period, and 28,800 acres on an annual basis. This creates some interesting alternatives when analyzing the data, especially when estimates of small domains are desired. There are many examples of subregional analysis of FIA data, and the Southern Station has often suggested to users a minimum area rule of thumb of 1 million acres. Over the full plot cycle of 5 years, this results in an approximate sample size of 173 plots or about 35 plots per year per million acres. If the intent is to estimate inventory on an annual basis, some thought should be given to increasing the information available beyond that of plots measured in the current year. Imputation can provide a cost-effective solution for increasing the information available for any given year.

Imputation is a technique that replaces each missing or deficient value with more acceptable values representing a distribution of possibilities (Rubin 1987). For this study, plot measurements in the four unmeasured panels are considered missing. Imputation methods are seemingly new to forest inventory, because there are few publications that formally address the topic. However, upon further inspection, it is clear that the profession has practiced imputation for several decades, most notably in Scandinavia (Poso 1978, Holm *et al.* 1979). In the United States, many inventory systems have used imputation, although under the label of modeling rather than imputation. The data that are modeled or imputed are treated as actual, and the business of producing inventory estimates proceeds.

Historically, inventories employing imputation have used different methods. For example, the Southern Research Station has at times used a plot matching procedure (Cost, personal communication). The need for imputation has typically resulted from access problems in remote and roadless areas in coastal swamps and wetlands. In this situation, the inventory has relied on a matching donor plot routine that is conceptually similar to the Census Bureau's hot-deck procedure (Sande 1983), while the North Central Station has modeled (projected) plots using STEMS (Belcher *et al.* 1982), an individual tree projection model. The modeled plots are used along with measured plots to produce inventory estimates (Leatherberry *et al.* 1995). The use of models within STEMS to update or "project" plots is an example of

Head of Inventory Techniques and Remote Sensing and Mathematical Statistician, respectively, Forest Inventory and Analysis, USDA Forest Service, Southern Research Station, Asheville, NC, USA.

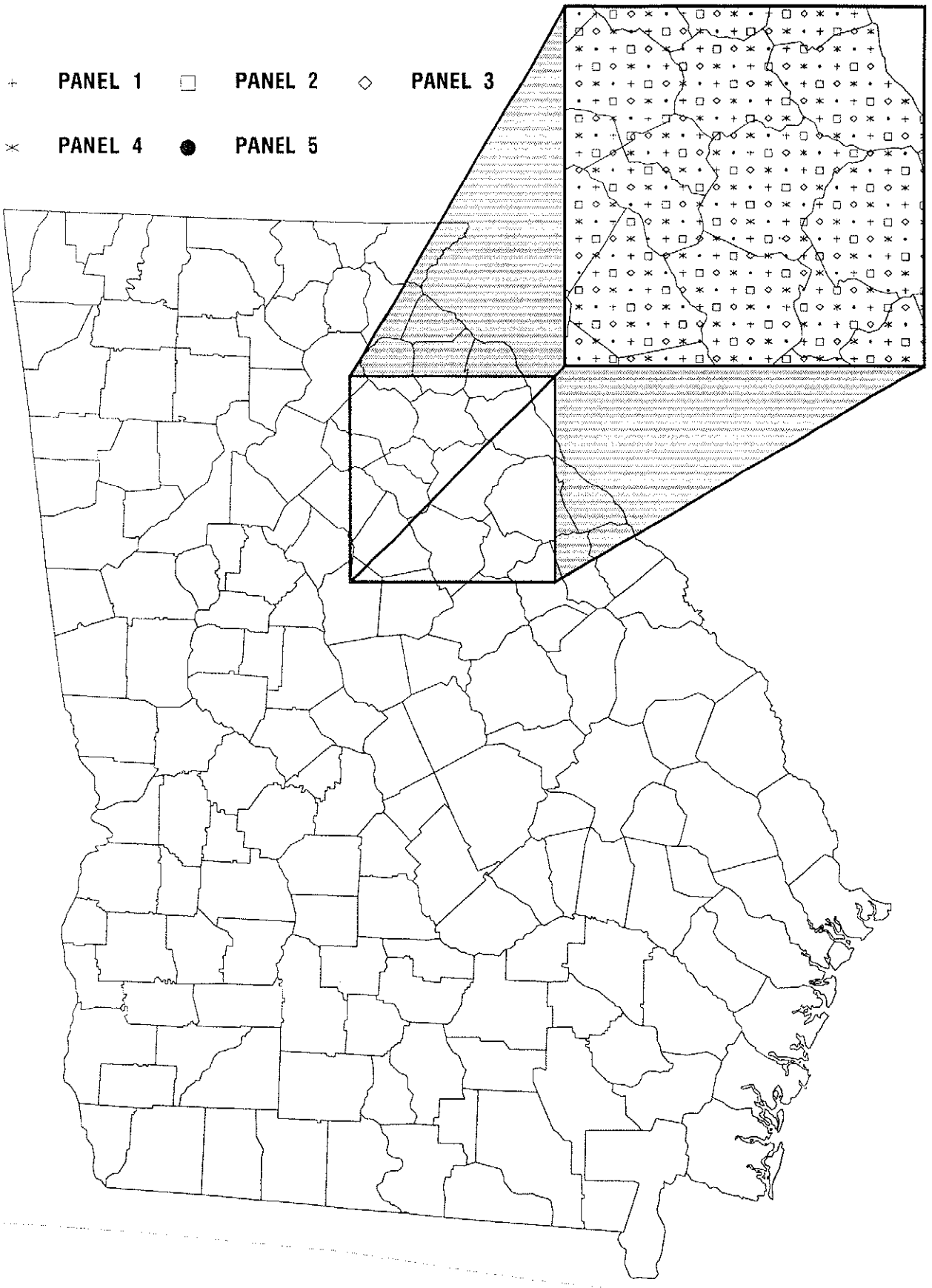

Figure 1.—*The interpenetrating panel design of SAFIS in Georgia. All plots in panel 1 are measured in year 1, panel 2 plots in year 2,..., panel 5 plots in year 5.*

mean imputation as defined by Rubin (1987). Obviously imputation works well, because a number of inventories use the concept.

Under SAFIS, if an annual estimate is desired, the simplest solution is to estimate current inventory based on the most recently measured panel. This is an intuitively appealing idea, but there are at least two practical concerns for users. The first concern is that with only one-fifth of the data, there will invariably be missing cells for many of the standard FIA core tables. To illustrate this, a stand table was generated using 20 percent (panel 1 of 5) of the data for survey unit 3 in Georgia (table 1). When compared to the complete set of plots (panels 1 through 5), the panel 1 stand table contains 24 missing cells for softwood species and 37 missing cells for hardwood species. The second concern is that even for commonly occurring forest types, the number of plots measured in a year can be small. Imputation or modeling can increase the available inventory information for any given year.

Given the obvious information gaps that occur when using only current-year data, a reasonable alternative is to retain all the data across the five annual panels and act as if all data are current. In this case, some users will disagree about whether the older panels are deficient.

For example, suppose a destructive hurricane similar to Hugo has occurred. Hurricane Hugo damaged more than 4.5 million acres and reduced softwood volume in affected areas by 21 percent (Sheffield and Thompson 1992). In this situation, use of prior panels is obviously dated given the recent catastrophe. For this and other situations, such as estimation for small domains, an approach where imputation procedures are used to update deficient plot data can prove useful. The following section outlines an imputation procedure for application to small domains.

METHODS

Four adjoining counties in central Georgia (Bibb, Crawford, Monroe, and Jones) were chosen as a small-domain case study. The total area of these counties is slightly less than 1 million acres. The plots in each county were assigned to panels 1 through 5, and sorted by forest type, physiographic class, stand size, stand age, and disturbance history. Except for the most commonly occurring forest types, this resulted in too few observations being available for imputation, and a coarser matching procedure based on forest type and physiographic class was implemented (table 2).

Table 1.—*Number of live trees on commercial forest land, by species and diameter class. Lower case x indicates that trees were observed in panel 1 (20 percent of the full survey), 0 indicates not observed in the full survey, and - indicates observed in full survey but not in panel 1.*

Species	All classes	5.0-6.9	7.0-8.9	9.0-10.9	11.0-12.9	13.0-14.9	15.0-16.9	17.0-18.9	19.0-20.9	21.0 28.9	29.0 and larger
	- - - - - - - - - *Number of trees* - - - - - - - - -										
SOFTWOOD:											
Longleaf pine	x	x	x	x	x	x	x	x	x	x	0
Slash pine	x	x	x	x	x	x	x	x	-	0	0
Shortleaf pine	x	x	x	x	x	x	x	x	x	x	0
Loblolly pine	x	x	x	x	x	x	x	x	x	x	-
Pond pine	x	-	-	-	-	x	-	0	-	0	0
Virginia pine	-	-	0	0	0	0	0	0	0	0	0
Pitch pine	0	0	0	0	0	0	0	0	0	0	0
Table-Moun pine	0	0	0	0	0	0	0	0	0	0	0
Spruce pine	x	-	x	-	0	0	0	-	-	-	0
Sand pine	x	x	x	-	0	0	0	0	0	0	0
E. white pine	0	0	0	0	0	0	0	0	0	0	0
E. hemlock	0	0	0	0	0	0	0	0	0	0	0
Spruce and fir	0	0	0	0	0	0	0	0	0	0	0
Baldcypress	x	x	x	x	x	x	-	-	-	x	-
Pond cypress	x	x	x	x	-	x	x	x	-	-	0
At. w.-cedar	0	0	0	0	0	0	0	0	0	0	0
E. Redcedar	x	x	x	x	x	x	-	0	0	0	0
Total softwoods	X	X	X	X	X	X	X	X	X	X	-

Table 2.—*Number of FIA plots by panel (P1-P5), forest type, and physiographic class for the counties of Bibb, Crawford, Monroe, and Jones in Central (survey unit 3) Georgia. T represents the total number of plots (P1+P2+P3+P4+P5) by forest type and physiographic class within the four counties. The needed grouping of plots by panel lists the number of current-year plots needed (imputed) to create a current-year data set for all 5 panels. The available number of plots listing is for all plots in survey unit 3 (excluding Bibb, Crawford, Monroe and Jones counties) by panel, type, and physiographic class. All imputed plots come from the available pool.*

Forest	Physio	Observed						Needed					Available				
type	class	P1	P2	P3	P4	P5	T	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
Longleaf/Slash	Xeric	1	.	.	.	.	1	.	.	.	.	.	2	.	1	1	.
	Mesic	1	.	.	.	.	1	.	.	.	.	.	4	6	6	5	6
Longleaf	Xeric	.	.	.	1	.	1	.	.	.	.	.	1	2	.	.	1
	Mesic	.	.	.	.	.	.	.	.	.	.	.	8	12	10	13	8
Loblolly-Shortleaf	Xeric	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
	Mesic	13	13	10	10	15	61	48	48	51	51	46	88	78	96	69	87
Loblolly	Xeric	1	1	1	.	.	3	.	.	.	.	.	3	3	1	3	2
	Mesic	4	3	5	9	11	32	28	29	27	23	21	63	52	56	58	52
	Hydric	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
Oak-Pine	Xeric	.	.	.	1	.	1	.	.	.	.	.	2	2	1	3	1
	Mesic	6	9	5	5	3	28	22	19	23	23	25	47	48	44	53	38
	Hydric	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1
Oak-Hickory	Xeric	.	1	.	.	.	1	.	.	.	.	.	2	4	2	5	3
	Mesic	10	7	11	12	7	47	37	40	36	35	40	102	104	86	97	111
	Hydric	.	.	.	.	.	.	.	.	.	.	.	1	.	2	2	.
Oak-Gum-Cypress	Mesic	2	1	4	.	2	9	.	.	.	9	.	13	19	24	17	18
	Hydric	.	.	.	1	.	1	.	.	.	.	.	21	17	22	29	17
Elm-Ash-Cottonwood	Mesic	.	1	.	1	.	2	.	.	.	.	.	8	8	8	4	2
	Hydric	.	.	.	.	.	.	.	.	.	.	.	3	3	2	.	.

For SAFIS, implementing an imputation scheme is conceptually easy but a bit operationally tedious, especially if all but the current-year's data are imputed. For example, the number of plots by forest type (eight classes), physiographic class (three classes), and panel that must be imputed to replace the four out-year panels with current-year data are listed in table 2 under the needed columns. To further explain, 61 plots across all five annual panels occur on mesic sites and are of loblolly-shortleaf forest type. Replacing the four out-year panels with current-year data for panel 1 requires 48 imputed plots, 48 imputed plots for panel 2, and 46 plots to complete panel 5.

The donor plots for this study are defined as similar plots; the matching is based on same forest type and physiographic class within the same survey unit and annual panel. The numbers of available (donor) plots are listed in table 2. The available pool of plots is composed of all plots in the survey unit except for those in the four counties. The reason for excluding plots within the four counties is that the same sample values would be repeated too often, and the desired goal of imputation is to base the estimates on a full range of plots that more reasonably represent the distribution of possibilities.

When the same sample elements are imputed repeatedly, the situation becomes similar to mean imputation via regression models. Mean imputation is a form of single imputation and results in an underestimate of the variance; the underestimate is directly attributable to the fact that there is no variability in the predicted (imputed) value given the same set of predictants (Ek *et al.* 1997). The underestimate is less pronounced with plot matching procedures because the variability of imputed values is greater since there are multiple donors that meet the

match criteria (Rubin 1987). For similar reasons, Sande (1982) recommends adding a random error term to regression-based imputations.

Regression-based imputation as suggested by Sande (1982) is defined as,

$$Y_{imp} = \hat{Y} + \hat{e},$$

where $\hat{Y}$ is the predicted value obtained from the fitted model, which is based on the complete observations, and $\hat{e}$ is the estimated residual that may be obtained by hot deck from the actual residuals of the fitted values or randomly generated using the estimated distribution of the residuals.

For forest inventory, there are at least two advantages of using multiple imputation over single imputation: first, when imputations are randomly drawn in an attempt to represent the distribution of the data, multiple imputation increases the efficiency of estimation. Second, when the multiple imputations represent repeated random draws under a model for nonresponse, valid inferences can be obtained by combining complete-data inferences in a straightforward manner. Because multiple imputation maintains the diversity that is inherent to the data, inventory users and specialists can reach valid inferences using familiar complete-data tools (Van Deusen 1996). The basics of estimation with multiple imputation follow.

MULTIPLE IMPUTATION BY EXAMPLE

Let Q be the quantity of interest in the survey. It could, for example, be $\bar{y}$ or some other parameter that is of interest. For the example presented here, merchantable cubic foot volume per acre of loblolly-shortleaf pine sawtimber stands will be Q. Q is a k-dimensional row vector, and assume inferences for Q based on the assumption that

$$(Q-\hat{Q}) \sim N(O, U)$$

where Q hat is an estimate of Q. After generating *m* simulated-completed data sets and analyzing each of them as if they were genuine complete data sets, we now have *m* estimates for Q and U, i.e., $\hat{Q}_{*1},\ldots,\hat{Q}_{*m}$ and $\hat{U}_{*1},\ldots,\hat{U}_{*m}$.

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

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

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

$$T_m = \bar{U}_m + (1 + m^{-1})\, B_m \qquad (2)$$

where

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

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

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

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

The results of a hot-deck imputation (m = 3) for merchantable cubic foot volume per acre of loblolly-shortleaf pine stands on mesic sites in the study area (Bibb, Crawford, Monroe, and Jones counties) are presented in table 3. The columns m = 1, m = 2, and m = 3 list the mean and variance for each of three imputed data sets by panel. Each panel estimate (mean and variance) is composed of 20 percent current-year plot measurements and 80 percent imputed data. The imputed data come from current-year plots within the survey unit (excluding the four counties) that match by forest type (loblolly-shortleaf), physiographic class (mesic), and stand size (sawtimber).

In general, the more refined the matching the more precise the imputations should be. Coarse matching will lead to an increase in both the within- and among-variance components of T_m. The question of how many imputations are necessary has been addressed by case studies and simulation studies. In a simulation study, Rubin and Schenker (1986) found that m = 2 or m = 3 was adequate for non-response rates of up to 60 percent. The application given in table 3 demonstrates that with m = 3 imputations, the imputed mean volumes per acre are quite reasonable and well within the 2 standard errors that one can expect if using only annually measured plots.

The variance estimates for merchantable cubic foot volume by individual imputation and for the multiple (m = 3) imputation indicate what Meng (1994) has demonstrated. That is, multiple imputation confidence intervals will be conservative. Meng (1994) further elaborates that the multiple imputation intervals are narrower than those from corresponding incomplete-data methods.

Using imputation for small domains has several advantages. First, a more complete and accurate set of FIA core tables can be constructed. Second, users and analysts of FIA data can use standard complete-data analysis methods. Third, annual estimates of inventory can be made each year, and this will largely eliminate the need for implicit imputations by external FIA user groups.

Table 3.—*Hot-deck imputations (m=3) of merchantable cubic foot volume per acre (mcfv) for loblolly-shortleaf pine sawtimber stands on mesic sites.* $\bar{Q}_3$ *(Eq. 1) is the mean of means and* T_3 *(Eq. 2) is the total variance of* $\bar{Q}_3$. *Ninety-five percent confidence intervals for mean mcfv from measured plots by panel are listed in the last column.*

	m = 1		m = 2		m = 3		Mean	Total	Measured plots only
	Mean	Variance	Mean	Variance	Mean	Variance	of means	variance	95% CI
Panel 1	2,145	956,986	1,974	565,976	2,324	796,895	2,147	814,066	2,064 ± 625
Panel 2	2,534	926,783	2,401	1,058,905	2,342	654,787	2,426	893,081	2,378 ± 456
Panel 3	2,191	930,702	2,170	895,794	2,330	916,174	2,230	924,375	2,025 ±1,046
Panel 4	2,359	1,278,874	2,357	1,426,937	2,133	858,295	2,283	1,210,439	2,132 ± 616
Panel 5	2,302	967,298	2,502	1,716,006	2,575	1,560,944	2,460	1,441,316	2,103 ± 663

ACKNOWLEDGMENTS

The following people reviewed this paper: Dr. Francis A. Roesch, USDA Forest Service, Asheville, NC, USA, and Dr. Paul C. Van Deusen, NCASI, Tufts University, Medford, MA, USA.

LITERATURE CITED

Belcher, D.W.; Holdaway, M.R.; Brand, G.J. 1982. A description of STEMS the Stand and Tree Evaluation and Modeling System. Gen. Tech. Rep. NC-79. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station. 18 p.

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.

Holm, S.; Hagglund, B.; Martensson, A. 1979. A method for generalization of sample tree data from the Swedish National Survey. Report No. 25. Umea, Sweden: Swedish University of Agricultural Sciences, Department of Forest Survey.

Leatherberry, E.C.; Spencer, J.S., Jr.; Schmidt, T.L.; Carroll, M.R. 1995. An analysis of Minnesota's fifth forest resources inventory, 1990. Resour. Bull. NC-165. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station. 102 p.

Meng, X. 1994. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 4: 538-573.

Poso, S. 1978. National forest inventory in northern Finland. In: National Forest Inventory, Proceedings of IUFRO Subject Group S4.02 Meeting, Bucharest, Romania.

Rubin, D.B.; Schenker, N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association. 81: 366-374.

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

Sande, I.G. 1982. Imputations in surveys: coping with reality. The American Statistician. 36(3): 145-152.

Sande, I.G. 1983. Hot-deck imputation procedures. In: Madow, W.G.; Olkin, I., eds. Incomplete data in sample surveys, Volume 3, Proceedings of the Symposium. New York: Academic Press: 334-350.

Sheffield, R.M.; Thompson, M.T. 1992. Hurricane Hugo effects on South Carolina's forest resource. Res. Pap. SE-284. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station. 51 p.

Van Deusen, P.C. 1996. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27: 379-384.