FIA Quality Assurance Program: Evaluation of a Tree Matching Algorithm for Paired Forest Inventory Data

James E. Pollard[1], **James A. Westfall**[2], **Paul A. Patterson**[3], **and David L. Gartner**[4]

Abstract.—The quality of Forest Inventory and Analysis inventory data can be documented by having quality assurance crews remeasure plots originally measured by field crews within 2 to 3 weeks of the initial measurement, and assessing the difference between the original and remeasured data. Estimates of measurement uncertainty for the data are generated using paired data statistical analyses. Because plot remeasurements are taken at different, but similar, times by different crews, it can be difficult to match the remeasured trees with the original tree measurements. In the past, this process required a laborious exercise of manual review and assignment of matching codes for the paired tree measurements. An automated process for matching tree data was developed and tested using a previously hand-matched data set. Results of the two matching processes were compared. More than 95 percent of the individual trees could be reliably matched using the automated matching program. The effects of unmatched data being excluded from the uncertainty analysis was minimal.

Introduction

The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture (USDA) Forest Service provides information needed to assess the status and trends of environmental quality in the Nation's forests. The FIA program works to continually improve monitoring and assessment activities by controlling, identifying, and documenting errors and sources of variability that could be detrimental to the quality of FIA inventory results. The quality assurance (QA) program within FIA involves the overall system of management activities designed to assure that quality data are collected. This program can be further divided into quality control and quality evaluation activities. Quality control within the program encompasses the operational techniques and activities that control the data acquisition process such as use of standardized field protocols. Quality evaluation activities involve application of statistical tools to determine if the uncertainty in the data will support programmatic decisions.

A large portion of the QA effort in the FIA program is focused on error control during the field measurement and data collection processes. One key element is provided through crew training and certification with specific national standards. Another key element of quality control in the program is development and annual updating of standardized field protocols that are documented in National Field Manuals (USDA 2003). In addition, the possibility of data entry error is reduced through use of portable field data recorders by inventory crew members. This onsite data recording reduces the chances of transcription-type data entry errors that are common problems in paper transfers. Finally, a variety of field check protocols provide immediate feedback to the crews and provide data to score crew performance.

In addition to extensive quality control activities discussed above, data quality is assessed and documented using performance measurements and post-survey assessments. These assessments identify areas of the data collection process that need improvements or refinements to meet the quality objectives of the program. Specific measurement quality objectives (MQOs) have been developed for the program and are presented in detail in the field methods guides. These quality standards were developed from extensive knowledge of measurement processes in forestry and to meet the program needs of FIA. Evaluation of data quality is accomplished by analysis of plot remeasurement data and comparison of the results to the MQO.

[1] FIA Quality Assurance Advisor, University of Nevada, Las Vegas, 4505 Maryland Parkway, Las Vegas, NV 89154.
[2] Research Forester, U.S. Department of Agriculture (USDA) Forest Service, Northeast Research Station, 11 Campus Blvd. Suite 200, Newtown Square, PA 19073.
[3] Mathematical Statistician, USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401–2394.
[4] Mathematical Statistician, USDA Forest Service, Southern Research Station, 4700 Old Kingston Pike, Knoxville, TN 37919.

Methods

Description of the Problem

An ongoing problem encountered when analyzing QA data is assuring that observations of individual trees are matched for paired statistical analysis. When plots are measured by two independent crews, it is not unusual for the crews to number or identify the trees slightly differently. This creates two data sets that may not be matched by tree number for a variety of reasons. For example, crews began numbering trees at different places on the plot, or crews missed a tree on the plot, setting the numbering sequence off. In addition, crews can number trees using a different spatial rule that can alter the numbering sequence for trees in a data file. Assuring that data are properly matched, and evaluating the consequences of mismatched trees in an inventory data set, is the subject of the current article. This study evaluates two different methods of assuring tree matching prior to data analysis.

The remeasurement process used to generate QA data sets in the FIA program is known as a blind check. This process involves a full reinstallation of a production inventory plot, performed by a qualified inspection crew, without access to the crew data. This results in two data sets that are independent of one another and can be subjected to paired data statistical analyses to obtain an unbiased estimate of the measurement uncertainty associated with crew performance. To analyze the quality of the two independent crews' data it is essential to have the data paired tree-to-tree so any error in the measurement process can be attributed to crew measurement error rather than data management or other nonmeasurement process errors.

The quality of FIA data has been evaluated in the past using blind check data (Pollard and Smith 1999, Pollard and Smith 2000). These data have been incorporated into a national forest health inventory report to document the basic data quality associated with these inventories (Conklin *et al.*, in press). However, to produce these assessments, it was necessary to obtain unbiased remeasurement data that was representative of the FIA program both operationally, temporally, and regionally. Once regional data sets were obtained we began a laborious process of preparing the data for analysis. This included normalizing regional differences in naming conventions and variables measured, as well as matching paired observations to the greatest extent possible.

The most time-consuming aspect of data preparation was assuring that paired observations of tree level variables were correctly matched. As increasing amounts of QA data are generated in the FIA program, and additional States are added to the national inventory, it becomes highly desirable to automate this tree matching process to the fullest extent possible.

Development of the Matching Process

Experience gained in analysis of QA data from inventories from 1998 through 2001 led to development of a hand-matching process for pairing tree-level data. The following steps were involved in this process:

- Two independently measured data files from a given inventory plot were obtained and identified as the crew data and the QA data. Each plot file was composed of four subplots of tree-level data that needed to be tree matched by subplot.
- Each tree in a given file was assigned a number within a subplot, which may or may not match the corresponding tree in the paired file depending on how the sequence was assigned (see discussion above).
- Tree-level variables were renamed in the QA data file and both files were sorted by subplot number, species of tree, horizontal distance of the tree from plot center, azimuth of tree measured at plot center, and diameter of tree at breast height.
- The data from both plots were printed with crew and QA data side by side and the data were visually compared for closeness of all matching parameters including the assigned tree number.
- If the tree numbers were not identical for sorted crew and QA observations within a subplot, then the numbering sequence was adjusted in the QA data file to match the crew data file.
- The decision to adjust the tree number was based on visual inspection for closeness of all matching parameters for a given tree as well as the total number of trees within a given subplot. For example, if crew data tree numbering started at 1 and the QA data had an extra tree in the subplot, then the numbering sequence would be off by one. In this case the tree numbers in the QA data file were adjusted to match the tree numbers in the crew data file. Then an extra number was assigned for the extra tree in the QA data file.

Once the tree numbering sequences in both crew and QA data files were matched, then differences between crew and QA crew observations could be calculated using the subplot and tree number as the identification key for a given tree.

This matching process can be very labor intensive, depending on the type of numbering discrepancy in the data files. For example, it was simple to identify an extra tree in a file and adjust the tree numbering sequences accordingly. However, if the files contained the same number of trees of the same species and the numbering sequence for more than two trees of the same species were transposed, it was much more difficult to identify which tree was the corresponding tree for a given number in the sequence. Occasionally the data for matching parameters for a number of trees in a given file were so close that it was necessary to align tree numbering sequences as a "best judgment" call. This hand-matching process was applied to a large set of Phase 3 FIA blind check data collected between 1998 and 2001 and required person months of effort totaling more than 3 years.

Refinement and Automation of the Matching Process

Refinement of the hand- matching process was initiated as a cooperative effort of three regional statisticians and the FIA Quality Assurance Coordinator. Automation of the process was developed in the SAS programming language and involved the following steps:

- QA variables were renamed in the QA data file and crew and QA data files were merged by State, county, plot number, and subplot number.
- A "distance" was computed for each QA tree to each crew tree using a function based on horizontal distance, azimuth, and diameter of the trees.
- Each QA tree was matched to the crew tree with the smallest distance. Pairs of trees were removed from the matched list because either multiple QA trees were matched to the same crew tree and only the QA with the shortest distance was matched or the distance was too great, or other technical reasons.
- A decision rule was incorporated in the matching algorithm that rejected potential matches having relatively large computed distances. This distance criteria was established to provide a conservative tree matching basis for this exercise. This distance matching criteria can be adjusted in the program if desired.
- The first iteration of matches was saved in a list file.
- Unsuitable matches were removed using similar standards as were used after the first iteration.
- A second iteration of distance functions were computed for those trees not matched in the first iteration.
- The two iterations of matched trees were combined and outputted into a matched tree list.
- The unmatched trees and/or extra/missed trees were separated into subfiles for manual examination to determine any remaining matches and to determine any missed and extra trees.

Description of the Data Files

The data were composed of inventory measurements from approximately 100 Phase 3 inventory plots measured between 1998 and 2001. Data were aggregated from the five FIA regions, for all years, which resulted in a national data set with reasonable representation from all FIA regions. The combined data set contained a total of 4,269 tree records in the QA file and 4,138 tree records in the crew file. The records in one file included trees that had corresponding matches in the other file, as well as additional trees that were unique to one or the other file. These "missed" or "extra" trees were screened from the combined data set resulting in 3,981 pairs of matched tree data that were assigned tree numbers based on best judgment of the analyst.

Results

Automated Matching Process

Application of the automated matching process to the national QA data set produced 3,576 pairs of matched trees after two iterations. Additional matched pairs of data could have been added to this data set by examination of the unmatched tree file and performing a hand-matching process. However, for the purposes of this exercise, it was decided to only use the trees matched by the fully automated process. Once the programming was complete, the actual matching process required less than one day's effort that included multiple runs of the program to verify comparability of the results of the two matching processes.

Uncertainty Analysis

The two data sets (hand matched and automated with two passes) were subjected to an analysis of mean differences between crews and estimates of MQO compliance. Simple MQO values were used to evaluate the robustness of the data sets. The tree level variables chosen for analysis represented characteristics of tree diameter, height, and crowns. The variables analyzed were diameter at breast height (DBH), diameter at root collar (DRC), total length of the tree (Total Length), actual length of the tree (Actual Length), foliar transparency (Transparency), foliar dieback (Dieback), and foliar density (Density) of the crown, as well as the crown class.

Mean Differences Between Crews

One estimate of measurement uncertainty that can be easily calculated is the average or mean difference between crew and QA measurements. Ideally we would expect the mean differences between the two crews to be zero, which would indicate that the two estimates for a given variable were not biased.

In addition to the central tendency of the differences the dispersion of these differences is an indicator of the overall reproducibility of the data set. The Means Procedure in SAS calculates the mean, standard error of the mean, and the minimum and maximum differences. This procedure also allows the mean differences to be tested to determine if they were significantly different from zero (biased) using a Student's t test (Probability | t | Value).

The results of these calculations for both matching processes showed that the hand-matched and automated matched data sets provided very similar estimates of data uncertainty (table 1). The mean differences between investigators were very similar with some variables having slightly larger differences for the QA crews and some having slightly smaller differences for the QA crews. The pattern of probability that the mean differences were not zero was also very similar. There was a tendency for the range of differences to be somewhat larger for the hand-matched data set than for the automated matching process. This would make sense because the automated matching process set

Table 1.—*Mean differences between investigators for diameter, crown, and length variables computed from the hand-matched data set (A) and automated matching data set (B).*

A. Hand-matching process						
Variable	**N**	**Mean**	**Standard error**	**Probability |t| value**	**Minimum**	**Maximum**
DBH	3,573	–0.03	0.02	0.0684	–15.9	22.7
DRC	408	–0.25	0.08	0.0019	–26.2	5.5
Transparency	2,884	0.03	0.14	0.8236	–60	79
Crown Class	1,410	0.10	0.02	<.0001	–4	4
Die Back	2,884	0.08	0.11	0.4536	–80	94
Density	2,884	–1.21	0.21	<.0001	–60	50
Total Length	1,529	0.47	0.22	0.0356	–66	63
Actual Length	1,590	–0.17	0.24	0.4912	–116	92
B. Automated matching process						
Variable	**N**	**Mean**	**Standard error**	**Probability |t| value**	**Minimum**	**Maximum**
DBH	3,250	–0.03	0.00	<.0001	–6.3	4.8
DRC	326	–0.17	0.09	0.044	–25.9	4.0
Transparency	2,594	0.09	0.14	0.5416	–60	79
Crown Class	1,341	0.10	0.02	<.0001	–4	4
Die Back	2,594	0.06	0.12	0.5912	–80	94
Density	2,594	–1.37	0.22	<.0001	–60	45
Total Length	1,443	0.3	0.19	0.1162	–30	57
Actual Length	1,498	–0.05	0.17	0.7749	–67	57

aside 405 sets of measurements for manual inspection based on the matching criteria provided in the program.

It is of interest to note that, with one exception (total length), the variables that had significant bias at < 10 percent probability were the same in both data sets. However, with the exception of density, the mean differences were very small, which would make the significance of the biases somewhat irrelevant.

Measurement Quality Objective Achievement

Analyzing the field crews' performance against the program assigned MQOs can be complex. For example, the MQO for DBH is ± 0.1 inch for every 20 inches of diameter. During this exercise, simplified MQO were assigned to variables as follows to allow easy interpretation of the efficacy of the matching processes (table 2).

To compare MQO compliance between hand-matching and the automated processes, cumulative frequency distributions were computed and the percentage of the differences noted for four levels of differences: zero differences; differences within the MQO; differences within two times the MQO; and differences within three times the MQO (table 3).

As with the results for mean differences, MQO compliance was very similar in both data sets. There was a slight tendency for the automated process to produce slightly improved MQO compliance although the improvement was rarely greater than a 2 percent improvement. It is likely that addition of the hand matched trees at the end of the automated process would result in virtually identical results.

Table 2.—*Simplified measurement quality objectives.*

Variable	Measurement quality objective
DHH	± 0.2 feet 95% of the time
DRC	± 0.4 feet 95% of the time
Transparency	± 10% Class 90% of the time
Crown Class	no errors 85% of the time
Crown Die Back	± 10% Class 90% of the time
Crown Density	± 10% Class 90% of the time
Total Length	± 5 feet 90% of the time
Actual Length	± 5 feet 90% of the time

Table 3.—*Cumulative percentage of the data set with zero differences between crews and one times, two times, and three times the simplified MQO for the hand-matching process(A) and the automated process (B).*

A. Hand-matching process					
Variable	**N**	**Percent zero differences**	**Percent 1X MQO**	**Percent 2X MQO**	**Percent 3X MQO**
DBH	3,573	50	90	94	95
DRC	408	25	74	85	91
Transparency	2,884	38	94	99	100
Crown Class	1,410	68	97	100	100
Die Back	2,884	61	98	99	99
Density	2,884	22	76	95	99
Total Length	1,529	22	74	88	95
Actual Length	1,590	23	75	89	95
B. Automated matching process					
Variable	**N**	**Percent zero differences**	**Percent 1X MQO**	**Percent 2X MQO**	**Percent 3X MQO**
DBH	3,250	54	94	97	98
DRC	326	29	70	82	86
Transparency	2,594	38	94	99	100
Crown Class	1,341	69	97	100	100
Die Back	2,594	61	98	99	99
Density	2,594	23	77	96	99
Total Length	1,443	22	76	90	96
Actual Length	1,498	23	77	91	97

Summary and Conclusions

Development of an automated tree matching shows much promise for time saving and simplification of data base manipulations within the FIA program for the following reasons:

- Hand-matching trees in an inventory data set produced more tree matches but required much more office labor.
- The mean differences between crews (bias) were similar for both matching methods.
- MQO compliance was similar for the two tree-matching procedures although the automated procedure tended to provide slightly better MQO compliance. It is likely that addition of hand matched trees from the list generated by the automated process would have generated very similar MQO compliance.

One needs to consider the size of the data set used in this study, however. With a sample size of thousands of trees, an automated tree-matching algorithm provided estimates of uncertainty and MQO compliance comparable to the laborious hand-matching data screening. However, if regional data sets or data sets for a given State are analyzed, the exclusion of unmatched trees from the data set may have a significant impact on the uncertainty analysis. Additional analyses are needed to evaluate this technique with smaller, regionally representative data sets. In addition, the matching program provides a list of unmatched trees. Using this much-reduced data set, hand screening of the unmatched trees becomes feasible, which should allow application of this process to much smaller data sets than were used in this study.

Acknowledgments

The authors would like to thank Susan Wright from the Northeastern Research Station for providing excellent technical editing of the manuscript. Her attention to detail and timely response is greatly appreciated. In addition, we wish to thank Mark Hansen from the North Central Research Station who provided many helpful suggestions and observations during the development of the automated tree-matching program.

Literature Cited

Conkling, B.L.; Coulston, J.W.; Ambrose, M.J., eds. [In press.] Forest health monitoring national technical report for 2001. Asheville, NC: U.S. Department of Agriculture, Forest Service.

Pollard, J.E.; Smith, W.D. 1999. Forest health monitoring 1998 plot component quality assurance report, vol. I. Research Triangle Park, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station.

Pollard, J.E.; Smith, W.D. 2001. Forest health monitoring 1999 plot component quality assurance report. Research Triangle Park, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station.

SAS/STAT. 1999. Users guide, ver. 8, vol. 2. Cary, NC: SAS Institute Inc. 2552 p.

Steel, R.G.D.; Torrie, J.H.; Dickey, D.A. 1997. Principles and procedures of statistics: a biometrical approach, 3rd edition. New York: McGraw-Hill. 666 p.