United States
Department of
Agriculture

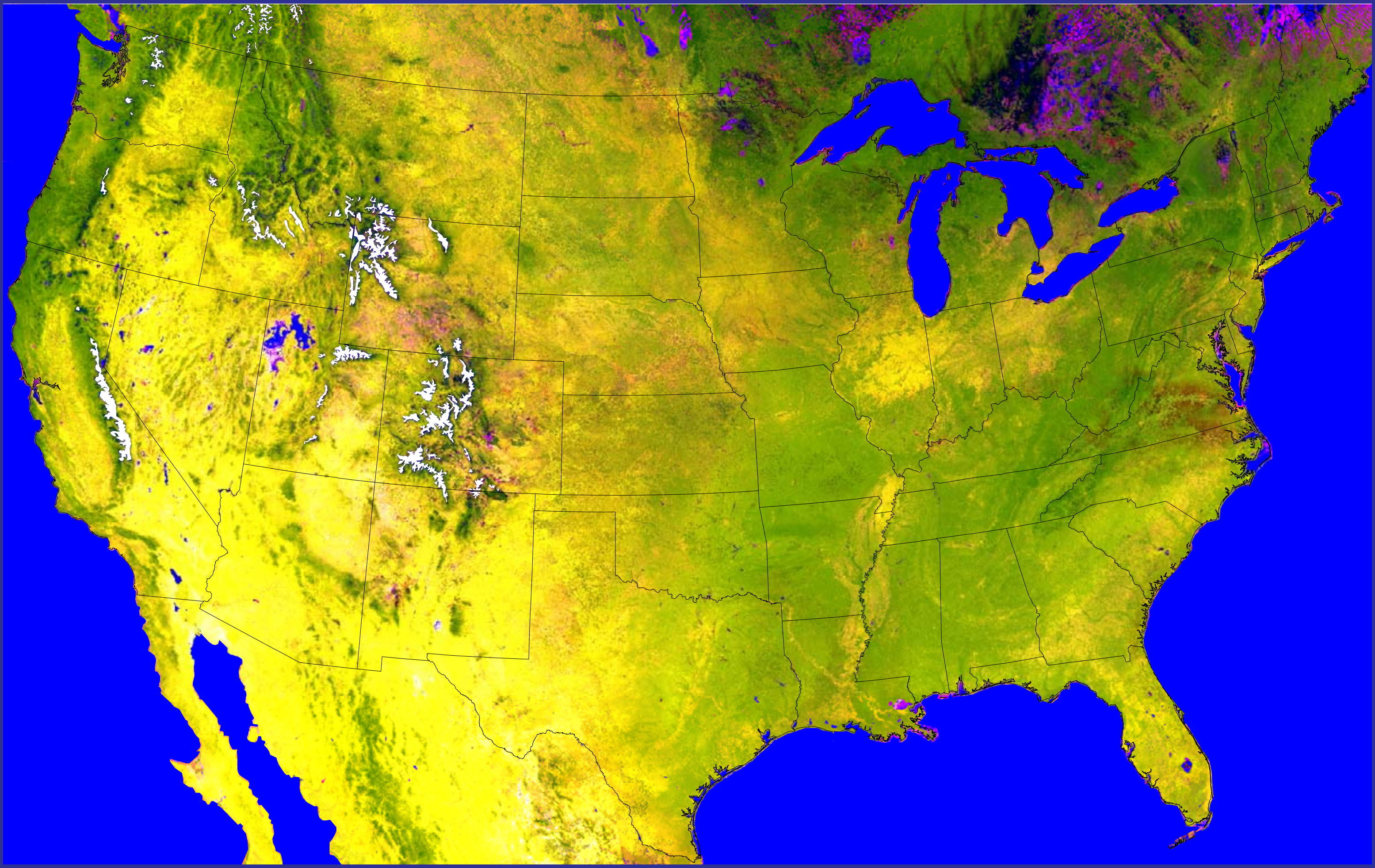
Forest Service

**Southern
Research Station**

General Technical
Report SRS–47

# Proceedings of the Second Annual Forest Inventory and Analysis Symposium

Salt Lake City, Utah
October 17–18, 2000



Cover: Forest area of the United States based on modeling of Advanced Very High Resolution Radiometer (AVHRR) data.

Map by Willam H. Cooke, with assistance from Lyn Thornhill.

## DISCLAIMER

The use or trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Department of Agriculture of any product or service.

Papers published in these proceedings were submitted by authors in electronic media. Editing was done to ensure a consistent format. Authors are responsible for content and accuracy of their individual papers.

December 2001

Southern Research Station
P.O. Box 2680
Asheville, NC 28802

# Proceedings of the Second Annual Forest and Inventory Symposium

***Edited by***
Gregory A. Reams
Ronald E. McRoberts
Paul C. Van Deusen

Salt Lake City, Utah
October 17–18, 2000

***Hosted by***
USDA Forest Service, Southern Research Station

***Sponsored by***
USDA Forest Service, Southern Research Station
USDA Forest Service, North Central Research Station
USDA Forest Service, Rocky Mountain Research Station
National Council for Air and Stream Improvement, Inc.

***Published by***
USDA Forest Service
Southern Research Station
Asheville, North Carolina
December 2001

# PREFACE

On October 17-18, 2000, the USDA Forest Service Southern Research Station sponsored a second annual Forest Inventory and Analysis (FIA) symposium in Salt Lake City, Utah. Symposium organizers successfully sought to document regional and national progress in implementing an enhanced FIA program. By the end of calendar year 2000, all Forest Service research stations will have begun annual inventories in at least one State using protocols developed for a system that is national in both scope and scale.

Papers included in this publication have been sorted into a number of general topic areas. Those areas include remote sensing and Geographic Inventory Systems (GIS), statistical estimation and modeling, information management, analysis and reporting, and a special-studies session. We thank all presenters and convey special thanks to those who so copiously documented their work and submitted papers for this publication.

Gregory A. Reams
Ronald E. McRoberts
Paul C. Van Deusen
Asheville, North Carolina

# CONTENTS

## Remote Sensing and GIS

*Moderator:*

**GREG REAMS**
USDA Forest Service
Southern Research Station

# ANALYZING REGIONAL FIA DATA IN THE ARCVIEW™ GEOGRAPHIC INFORMATION SYSTEM[1]

David Reed, Scott Pugh, Pat Miles, and Kurt Pregitzer[2]

**Abstract—**The ArcView™ Geographic Information System (GIS) is probably the GIS that is most widely used by federal and state natural resource management agencies, industries, and non-profit institutions. As such, there is a great deal of expertise and comfort with this package within the FIA user community. The North Central Research Station and Michigan Technological University have recently cooperated in the release of ArcView GIS projects for each of the eleven North Central states. These projects facilitate analyses of current volume, net growth, mortality, and removals for any geographic region within a state, whether determined from ancillary polygon data (such as soil or watershed boundaries) or defined on-screen by the user. Current efforts are focused on developing a single project that will allow similar analyses that are not limited by state boundaries within the eleven-state North Central region.

## INTRODUCTION

The users of USDA Forest Service Forest Inventory and Analysis (FIA) data include state and federal agencies, forest industry, other corporate entities, non-profit organizations, and private citizens. Many of these users utilize the ArcView™ (ESRI, Inc.) Geographic Information System (GIS) to facilitate their analyses of natural resource data. Many of the users of FIA data, therefore, have in-house expertise in the use of ArcView. They also typically have many other types of natural resource information, such as soil maps, watershed boundary delineations, political boundaries, transportation networks, etc., in ArcView format. To date, it has been difficult to link the FIA data with these other resource data, and it has been difficult for users to conduct their own flexible spatial analyses of FIA information.

The USDA Forest Service North Central Research Station and Michigan Technological University have recently cooperated in the release of ArcView projects for each of the eleven North Central States. In ArcView, a project is a file for organizing work consisting of views, tables, charts, layouts, and scripts (ESRI, Inc. 1998). These projects utilize FIA data as presented in the Eastwide database (Hansen and others 1992); the data files undergo a pre-processing routine to increase the speed of the ArcView Avenue scripts that execute the analyses. The projects are available at no cost from either organization, and allow users to conduct analyses of acreage, current volume, net growth, mortality, and removals by species or forest type for any geographically defined region within a state's boundaries. Users can subset plots within a defined region using any of the descriptive FIA variables such as landowner, forest type, age class, site index, and so on. Users can do multiple sorts so they can, for example, examine volume, productivity, and utilization for all of the 31-40 year old aspen plots within a large watershed.

The FIA ArcView GIS projects promise to be of great utility to users. It is currently awkward to conduct analyses across state boundaries, with users being required to define the areas of interest in each state, conduct the analyses of interest in each state, and combine the results outside of the GIS system. Current efforts are focused on developing a single ArcView project for the entire eleven-state North Central region. This will allow users to conduct analyses for any geographic area within the region, with the GIS seamlessly combining information across state boundaries to produce results in a single step.

## EXAMPLE ANALYSES

The incorporation of multi-state functionality further increases the utility of the existing routines to users. In many instances, areas of interest do not follow state or other political boundaries. Watersheds and other ecological units cross state boundaries, many mills procure wood from more than one state, and so on. The first example illustrates an analysis utilizing FIA plots located within 25 miles of the point where the boundaries of Illinois, Iowa, and Wisconsin intersect (fig. 1). This selection utilizes FIA plots from the 1985 Illinois, 1990 Iowa, and 1996 Wisconsin surveys. Table 1 shows the volume, growth, mortality, and removal estimates for selected species from the 274,943 acres of timberland within this area. Table 2 provides detailed information on the volume and change of the northern red oak resources within this area by two-inch diameter class.

The second example illustrates the use of a polygon defined by another resource layer. This analysis utilizes the FIA plots occurring within a polygon defined by a particular class of alluvial soils in the southern portion of the region (fig. 2). This area contains FIA plots from the following surveys: 1990 Iowa, 1985 Illinois, 1998 Indiana, 1994 Kansas, 1994 Nebraska, and 1989 Missouri. The acreage of some of the most common forest types in this area is shown in table 3.

Tables 2 and 3 illustrate just some of the analyses available from the software. In addition, composite volumes, net growth, mortality, and removals (and their associated

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Professor and Assistant Research Scientist, School of Forestry and Wood Products, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931; Forester, USDA Forest Service, North Central Research Station, 1992 Folwell Ave., St. Paul, MN 55108; and Professor, School of Forestry and Wood Products, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, respectively.

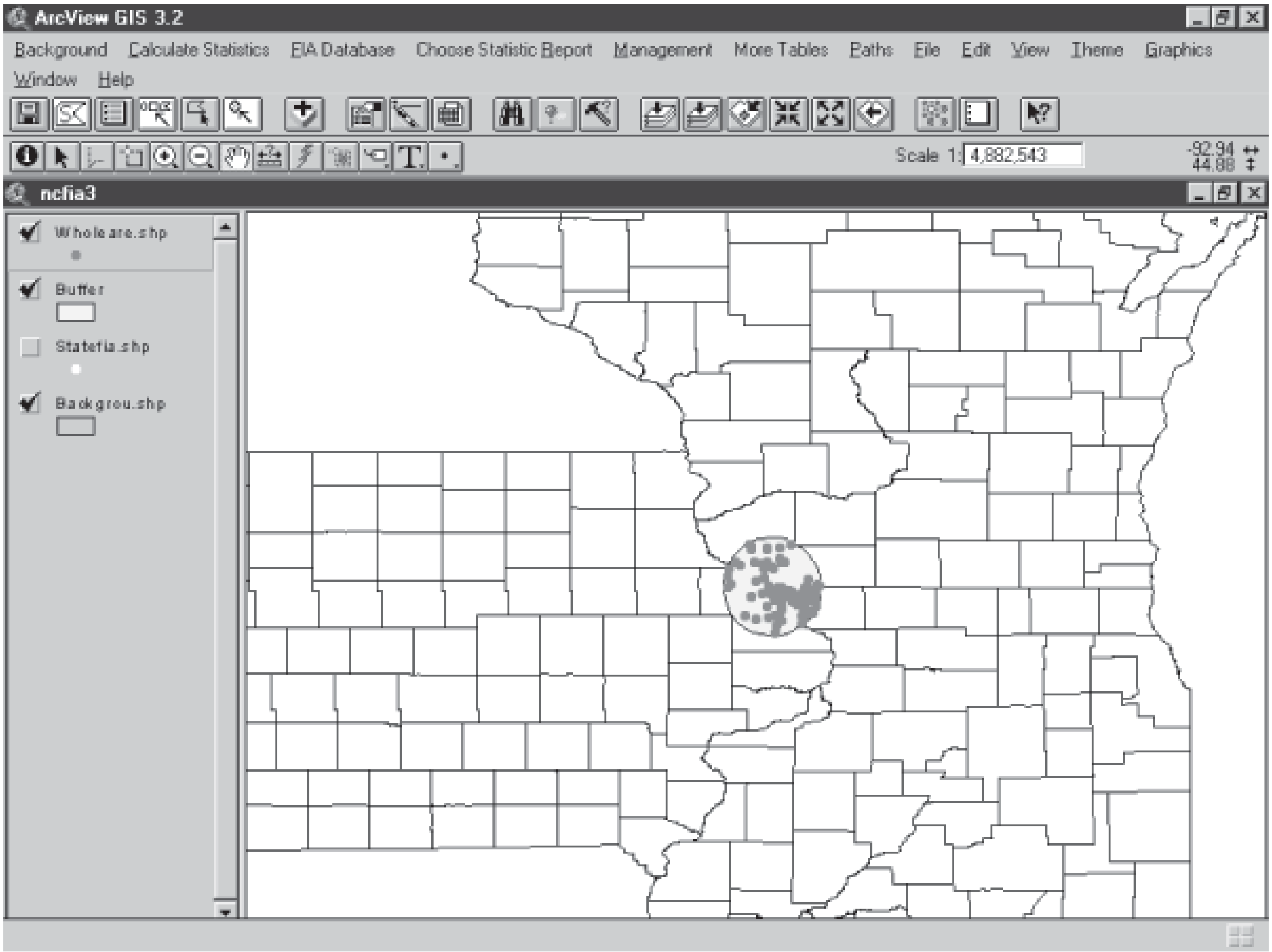


Figure 1—FIA plots located within 25 miles of the point where the boundaries of Illinois, Iowa, and Wisconsin intersect, obtained using the ArcView buffering function to define the area of interest, which is then overlain with the FIA plot location information.

**Table 1—Current volume, growth, mortality, and removal estimates for selected species on the 274,943 acres of timberland within 25 miles of the intersection of Illinois, Iowa, and Wisconsin (fig. 1)**

| Species | Volume | Growth | Mortality | Net growth | Removals[a] |
|---|---|---|---|---|---|
| | *1000 ft$^3$* | - - - - - - - - - - - - - - - - - - | *-1000 ft$^3$/yr-* | - - - - - - - - - - - - - - - - - - | |
| Basswood | 23,246.4 | 584.5 | 229.0 | 355.4 | 45.3 |
| Elm | 21,271.9 | 1,298.6 | 503.3 | 795.3 | 601.6 |
| White oak | 14,301.7 | 156.2 | 77.0 | 79.2 | 313.7 |
| Northern red oak | 12,921.7 | 460.4 | 91.6 | 368.8 | 2,257.9 |
| Sugar maple | 12,009.8 | 225.6 | 90.1 | 135.5 | 461.1 |

[a] Removals for Illinois 1985 are not available in the FIA inventory data or in the GIS-FIA Model. An estimate is available from a single year survey of mills (Hahn 1997).

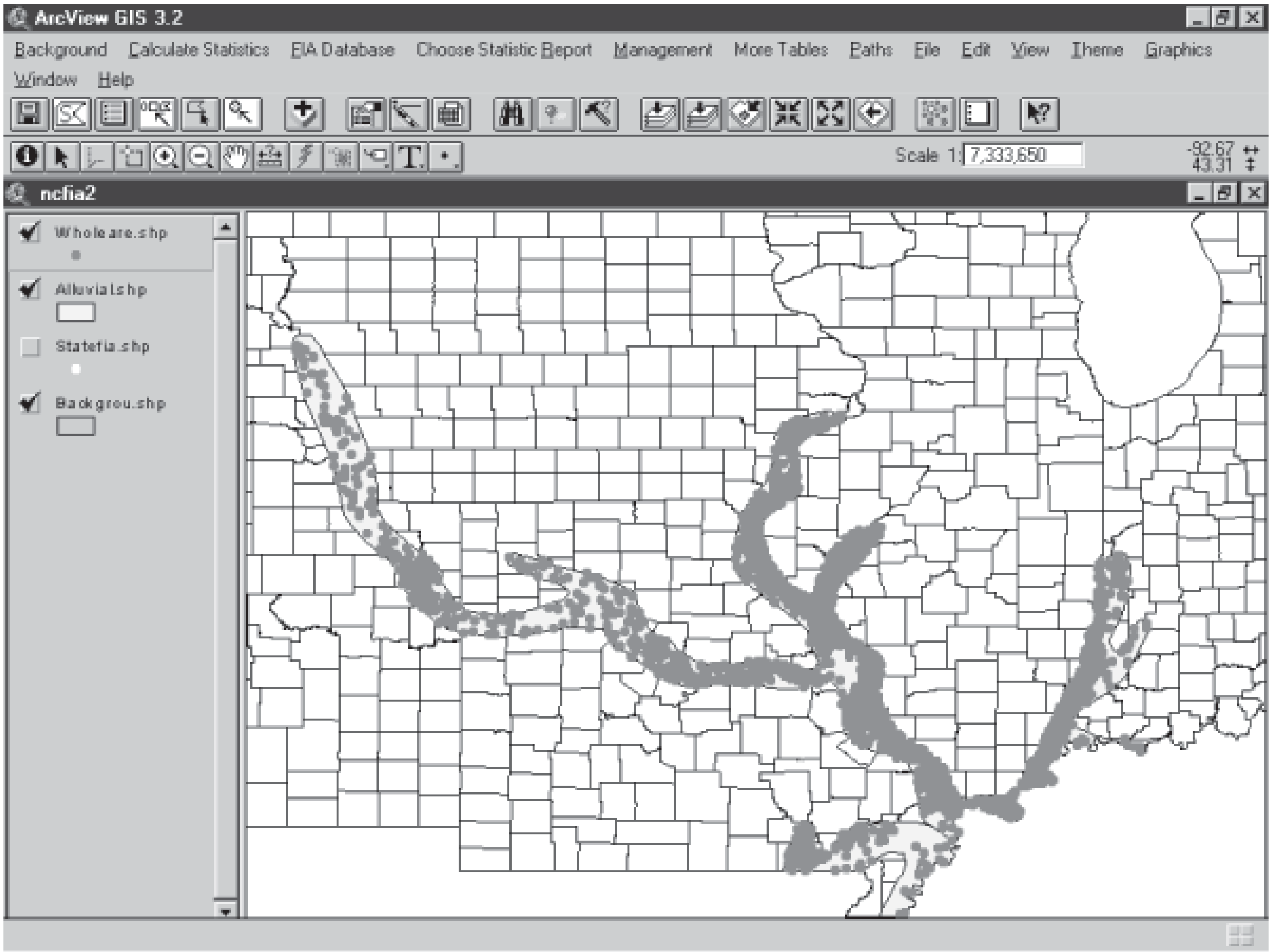


Figure 2—FIA plots occurring within a polygon defined by a particular class of alluvial soils in the southern portion of the North Central region, created by overlaying a general soils polygon coverage with the FIA plot location information.

**Table 2—Volume and change of the northern red oak resources within 25 miles of the intersection of Illinois, Iowa, and Wisconsin (fig. 1) by two-inch diameter class**

| Diameter | Volume | Growth | Mortality | Net growth | Removals[a] |
|---|---|---|---|---|---|
| *In.* | *1000 ft$^3$* | | *- - - - - - - - -1000 ft$^3$/yr- - - - - - - - -* | | |
| 4 | 0.0 | 96.6 | 0.0 | 96.6 | 0.0 |
| 6 | 150.1 | 7.2 | 8.9 | -1.7 | 0.0 |
| 8 | 185.1 | 8.8 | 3.2 | 5.6 | 0.0 |
| 10 | 702.3 | 31.4 | 3.5 | 27.9 | 0.0 |
| 12 | 853.2 | 31.7 | 2.9 | 28.7 | 14.3 |
| 14 | 1,129.4 | 22.1 | 4.0 | 18.1 | 32.7 |
| 16 | 1,497.5 | 48.0 | 7.2 | 40.9 | 207.0 |
| 18 | 2,793.7 | 108.2 | 27.2 | 81.1 | 396.5 |
| 20 | 1,205.4 | 42.5 | 30.2 | 12.2 | 525.0 |
| 22 | 1,454.3 | 2.6 | 1.8 | 0.7 | 247.0 |
| 24 | 1,822.1 | 38.1 | 0.0 | 38.1 | 369.6 |
| 26 | 1,128.5 | 23.2 | 2.6 | 20.6 | 465.9 |
| Total | 12,921.7 | 460.4 | 91.6 | 368.8 | 2,257.9 |

[a] Removals for Illinois 1985 are not available in the FIA inventory data or in the GIS-FIA Model. An estimate is available from a single year survey of mills (Hahn 1997).

**Table 3—The timberland acreage of some of the most common forest types in the area defined in figure 2**

| Forest type | Acres |
|---|---|
| Maple-beech-birch | 503,100 |
| Oak-hickory | 473,900 |
| White oak-red oak-hickory | 454,579 |
| White oak | 408,115 |
| Elm-ash-cottonwood | 377,100 |

standard errors) are available for all selected plots. A composite stand table giving average trees per acre by species and two-inch diameter class is also available in the single state projects. In general, all of these tables are available for the 50 most common species occurring in the region.

## DISCUSSION

A concern with the analyses in figures 1 and 2 is that the data from the different states were taken at different times. These are the most recent publicly available data for the different states, but the result is that the data from the 1985 Illinois survey, which are now fifteen years old, are combined with the two-year old 1998 Indiana data. Without a commonly accepted stand projection system, these inventories cannot be updated to a common date. With the advent of the annual surveys, though, this problem with the timing of different state surveys will be greatly reduced. The ArcView projects can easily be updated when annual data from the various states are released, resulting in an annual regional project. This will provide users with up-to-date resource information in a GIS format with which they are familiar and are already comfortable using.

The latitude-longitude location information publicly released for the FIA plots does not represent exact plot locations. The coordinates are rounded off or randomly altered to mask exact plot locations; the method for doing this varies across the US. When combined in a GIS, this has the effect of including some plots within a polygon boundary that should not be there, and excluding some plots from a polygon that should really be included. In Michigan, some forested plot locations appear in Lake Superior, for example. For large areas, the effect of these distortions is minimal. For small areas, though, a rather large proportion of the total forest area can be involved. An analysis of the Michigan 1993 data indicates that county-level volumes estimated in the GIS by overlaying county boundaries with the FIA plot locations are almost always within 10 percent of the volumes estimated using the county code recorded in the Eastwide database (Hansen and others 1992). The imprecise plot location effect is minimized for geographic areas with low perimeter:area ratios, and increases as the amount of boundary area increases relative to the interior area. Circular areas, such as in figure 1, have low perimeter:area ratios and minimal errors introduced by imprecise plot locations. The introduced errors are still proportionally greater for smaller circles than they are for larger ones. Long, linear areas, such as riparian zones, have large perimeter:area ratios and may be expected to have relatively large introduced errors due to imprecise plot locations. It makes little sense to try to examine corridors less than two or more miles in width, for example, if the plot locations are only known to the nearest mile. Users can obtain some guidance from the estimated standard errors produced during the analyses; these are underestimates, however, because they do not fully consider the imprecise plot locations. Users need to check these precision estimates, and make sure that the number of plots or size of area selected results in a greater indicated level of precision than is really required.

The ArcView projects described here were taken directly from the data in the Eastwide database. The FIA data are pre-processed to speed the ArcView analyses, but summaries are based on the formulae given by Hansen and others (1992). At the state level, results from the ArcView projects match almost exactly with those from the respective state publications. The ArcView projects calculate standard errors for composite volume estimates, net growth, and removals; these are calculated using the ratio formulae given in the respective state reports. When data from different states are combined, a simple weighting procedure is used to estimate the standard errors for the total defined area.

Execution time is an issue with any software designed to manipulate large amounts of data. With tens of thousands of FIA plots and hundreds of thousands of tree records in the North Central region, a great deal of programming effort has gone into increasing the efficiency of the analyses illustrated here. All of the analyses illustrated in figures 1 and 2 and the results summarized in tables 1-3 can be obtained in about 20 minutes on a 233 MHz laptop computer provided the soils layer is available at the beginning as an ArcView theme.

## SUMMARY

ArcView projects containing the FIA data for each of the eleven North Central states and associated data processing procedures are currently available from the US Forest Service North Central Research Station or Michigan Technological University. Many natural resource organizations use the ArcView GIS software and have in-house expertise for conducting analyses using this software. The individual state projects provide the capability for FIA data users to link FIA data to other natural resource data, and to conduct in-house analyses of natural resource information for any defined geographic region within a state. The regional model under development will extend these capabilities across state lines within the eleven-state North Central region. Both the individual state and the regional models will provide users with a vehicle to utilize annual inventory data, when available, to conduct quick analyses addressing a variety of resource issues.

## ACKNOWLEDGMENTS

Michigan Technological University and the US Forest Service North Central Research Station, through Agreement 00-CR-11231300-040, each provided partial support for this work. Tom Wardle, Deputy State Forester for the State of Nebraska, and Michael Hyslop, GIS Analyst in the School of

Forestry and Wood Products, Michigan Technological University, provided very helpful comments during the preparation of this manuscript. For information on ArcView projects for the eastern states outside the North Central region, please contact one of the authors at Michigan Technological University.

## REFERENCES

**ESRI, Inc.** 1998. Introduction to ArcView GIS. Environmental Systems Research Institute, Inc. Redlands, CA.

**Hahn, Jerold T.** 1997. Illinois forest statistics, 1985. Resource Bulletin. NC–103. St. Paul, MN: U.S. Department Agriculture, Forest Service, North Central Forest Experiment Station. 101 p.

**Hansen, M.H.; Frieswyk, T.; Glover, J.F.; Kelly, J.F.** 1992. The Eastwide forest inventory data base: users manual. Gen. Tech. Rep. NC–151. St. Paul: U.S. Department of Agriculture, Forest Service, North Central Research Station. 48 p.

# FOREST/NONFOREST CLASSIFICATION OF LANDSAT TM DATA FOR ANNUAL INVENTORY PHASE ONE STRATIFICATION[1]

Jim Rack[2]

**Abstract**—Launch of Landsat 7 creates the opportunity to use relatively inexpensive and regularly acquired land cover data as an alternative to high altitude aerial photography. Creating a forest/nonforest mask from satellite imagery may offer a cost-effective alternative to interpretation of aerial photography for Phase One stratification of annual inventory plots. This paper describes the procedures: they include image rectification, registration, and spatial filtering to allow accurate co-location with field plots and attempt to compensate for minor plot location errors. Identification of clouds and their removal from further analysis is outlined. Image alarms are described as a coarse filter for arriving at a forest/nonforest mask, with unsupervised classification as the fine filter. Accuracy assessment results for single-date, dual-date, filtered and unfiltered combinations are reported, as well as cost estimates.

## INTRODUCTION

Phase One stratification for Forest Inventory and Analysis (FIA) survey purposes has traditionally been accomplished through detailed photo interpretation of the most current high altitude photography. Conversion from periodic to annual inventories necessitates acquisition of current imagery more often than high altitude photo programs can provide at traditional funding levels. Satellite image analysis offers a cost effective alternative.

One advantage of satellite imagery is the ability to machine-process large areas in a relatively short time. Another is that imagery is acquired on a regular cycle. Costs of imagery and analyses are lower. Disadvantages relate to accuracy: resolution is not as fine as aerial photography, and the human ability to interpret context, shape, and texture are lost.

The following methodologies were developed and tested in an attempt to apply well-established, simple techniques that could be quickly and easily implemented across a large program such as FIA.

## IMAGE ACQUISITION AND PREPROCESSING

Midsummer imagery may offer the best opportunity for discriminating forest from nonforest if only a single date is used in the analysis. Addition of imagery from another season may improve accuracies by incorporating phenological differences.

Although the choice of sensors is expanding, driven mostly by a desire for increased spatial resolution, Landsat Thematic Mapper (TM) remains one of the better choices when considering classification needs of a forest target. Its spatial resolution is somewhat smaller than the size of an FIA plot, yet not so small as to overwhelm storage and processing capacities when dealing with large land areas. Its spectral resolution is greater than most other commercial systems and offers better classification potential.

Two dates of imagery were used in this study. The first was a July 24, 1999 Landsat 5 scene from Path 27, Row 26 shifted 70 percent south. The second was an October 13, 1999 leaf-off Landsat 7 scene from Path 26, Row 27. Both were rectified to the Minnesota Department of Natural Resources (DNR) standard of extended zone 15 UTM projection, NAD83 datum, using the MN Department of Transportation ARC/INFO roads coverage for ground control points.

The October scene needed some cloud removal. A Normalized Difference Cloud Index, (TM5 - TM6) / (TM5 + TM6) was calculated and added to the image as an additional band. An unsupervised classification of 70 classes was then created and cloud and cloud-shadow classes identified and masked out. Another unsupervised classification of 150 classes was run on the cloud-free images to identify obvious water and remove it from further analysis.

The remaining unmasked imagery contained forest and nonforest land pixels that can be roughly separated using an image alarm available in many types of image analysis software. The analyst roams the imagery digitizing a variety of coniferous and deciduous stands, trying to include the range of variability for each. I selected 30 of each and merged the 30 separate polygons into one coniferous and one deciduous signature. The image alarm allows the analyst to edit parallelepiped limits for bands 3, 4, and 5. The analyst interactively edits until the pixels alarmed reasonably represent the labeled class. If two dates of imagery are used, there are six bands to edit. The pixels identified with the alarm are separated into a preliminary “forest” area of interest (AOI) that will be further classified. The remaining pixels are identified as a preliminary “nonforest” AOI.

## CLASSIFICATION REFINEMENT

Each of the AOIs are further classified using a 35-class unsupervised technique to build signatures, followed by a maximum likelihood supervised classification of the pixels.

[1] Paper presented at the Second Annual Forest Inventory and Anaylsis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Image Analyst, MN Department of Natural Resources, Division of Forestry—Resource Assessment, 413 SE 13th Street, Grand Rapids, MN 55744.

This classification will do a better job than the alarm of distinguishing fine differences between forest and nonforest. These 35 classes are labeled using 3 or more sets of the best available aerial photography. Photos should be well distributed across the image.

Some signatures from this classification will be confused and require further refinement. Each is identified as an AOI, and another unsupervised classification is performed using some number of subclasses, which are then relabeled using the photographs. Eventually, most ambiguity should be removed, and a model can be written to recode all pixels to either forest or nonforest classes.

Two spatial concerns related to FIA plots need to be addressed. One is accuracy of plot locations. A small test of plot location accuracy conducted in our office on 1990 era plots revealed a root mean square error of approximately 50 meters, even after plot locations had been extensively edited. The second refers to the difference in area between a pixel and an FIA plot cluster. One pixel is a square 30 meters on a side. It would take approximately 9 pixels in a 3 by 3 (3x3) matrix to cover the same area represented by the new 4 subplot cluster or the old 10 subplot cluster. To compensate somewhat for the inaccurate locations of the FIA plots and the difference in size between a pixel and a plot, a 3x3 majority filter was used to assign the majority makeup of a 9-pixel area to the center pixel. This 3x3 filter matrix passes over each pixel in the classification and outputs a new filtered classification. Results are reported for both the single-pixel classifications and the filtered classifications.

## ADDITIONAL STEPS FOR TWO-DATE CLASSIFICATIONS

Using two dates of imagery requires some additional work in preprocessing. Registration between the two images must be checked: if pixel locations are not coincident, the images will require additional registration work. An overlap area must be identified so that analyses are restricted to pixels that contain data from all bands of both input images. Layers from the two images will have to be “stacked” to create one image for analysis, containing as many as 12 bands of data.

## RESULTS AND DISCUSSION

Classifications were compared to 106 actual FIA plots field measured in 1998 and 1999. Four classification combinations were compared to various combinations of FIA ground land use (GLU) and history. Histories of “clearcut”, “natural significant disturbance”, and “man-caused significant disturbance” on forested plots were considered to be in a nonforested state for accuracy assessment purposes. Plots were also checked visually on the imagery to see if current condition matched attributes in the database: for example, if a forested plot was clearcut after the field visit and before the image date. An additional 44 nonforested plots were added to compensate for the small number of field-visited nonforest plots; these were identified by “PI_LAND_USE” and “GLU” codes of nonforest from annual inventory plots selected in 1994-1996. Only the overlap area of the two images was included in the accuracy assessment.

The single-date unfiltered classification accuracy assessment matrix, table 1, shows a certain bias on the part of the analyst towards an aggressive classification of “forest”. This bias offers the advantage of insuring that all or most of the actual forested plots are selected for field visits, but has a negative effect on estimates of forest area.

Results (table 2) from filtering the single-date classification indicate a slight improvement in overall accuracy from 85 percent to 88 percent. The filtering caused six plots classified as “forest” to change to a “nonforest” classification and one plot classified as “nonforest” to change to “forest”. Visual inspection of the plot locations changed by filtering confirmed the neighborhood of pixels to be a generally better representation of conditions at the plot than the single pixel at “plot center”.

Using a second image in the classification also improved the classification accuracy, as shown in table 3, from 85 percent to 88 percent. However, filtering this classification slightly reduced the accuracy, from 88 percent to 87 percent (table 4). In this case, filtering caused the classification of five plots to change from forested to nonforested and nine plots from nonforested to forested.

Visual inspection of the errors from these matrices indicates that plot location is a chief contributor. Plots near borders of

**Table 1—Single-date classification, 7/24/99, unfiltered**

| | FIA forest | FIA nonforest | Total | User's accuracy |
|---|---|---|---|---|
| | | | | *Percent* |
| Classified forest | 81 | 23 | 104 | 78 |
| Classified nonforest | 0 | 46 | 46 | 100 |
| Total | 81 | 69 | 150 | |
| | | | | Overall acccuracy |
| Producer's accuracy (%) | 100 | 67 | | 85 |

**Table 2—Single-date classification, 7/24/99, unfiltered with 3x3 majority filter**

| | FIA forest | FIA nonforest | Total | User's accuracy |
|---|---|---|---|---|
| | | | | *Percent* |
| Classified forest | 81 | 18 | 99 | 82 |
| Classified nonforest | 0 | 51 | 51 | 100 |
| Total | 81 | 69 | 150 | |
| | | | | Overall acccuracy |
| Producer's accuracy (%) | 100 | 74 | | 85 |

**Table 3—Two-date classification, 7/24/99, unfiltered and 10/13/99, unfiltered**

|  | FIA forest | FIA nonforest | Total | User's accuracy |
|---|---|---|---|---|
|  |  |  |  | *Percent* |
| Classified forest | 64 | 13 | 77 | 83 |
| Classified nonforest | 3 | 56 | 59 | 95 |
| Total | 67 | 69 | 136 |  |
|  |  |  |  | Overall acccuracy |
| Producer's accuracy (%) | 96 | 81 |  | 88 |

**Table 4—Two-date classification, 7/24/99, unfiltered and 10/13/99, filtered with 3x3 majority filter**

|  | FIA forest | FIA nonforest | Total | User's accuracy |
|---|---|---|---|---|
|  |  |  |  | *Percent* |
| Classified forest | 65 | 16 | 81 | 80 |
| Classified nonforest | 2 | 53 | 55 | 96 |
| Total | 67 | 69 | 136 |  |
|  |  |  |  | Overall acccuracy |
| Producer's accuracy (%) | 97 | 77 |  | 87 |

forest/nonforest conditions have a higher probability of being labeled incorrectly, especially if plot locations are imprecise. As plot locations are updated with high-accuracy Global Positioning Systems during field visits, classification accuracies should also improve. Other conditions that contributed to errors were plots labeled "marsh without trees" and "right of way," which tended to be misclassified as forest.

The general conclusion one may draw from this small test is that  simple image processing techniques of satellite imagery can offer almost 90 percent accuracies of forest/ nonforest discrimination. Whether this is adequate for Phase One stratification of FIA, or repeatable on other landscapes, is yet to be determined.

## COSTS

Image costs will vary depending on the source of the data. Landsat 7 scenes carry about a $600-$800 price dependent on the level of processing. The image analyses will take between 7 and 10 days per scene center depending on whether it is single date or dual date and the presence of clouds. Twenty scenes of Landsat 7 for Minnesota would cost about $13,000 and image analysis would add $40,000-$50,000. Round numbers would show a cost of about $53,000 for the 53 million acres in Minnesota or $0.001/acre ($0.64/square mile). These are estimates only and not to be considered universal for all users. Individual circumstances could alter these numbers substantially.

# PHASE I FOREST AREA ESTIMATION USING LANDSAT TM AND ITERATIVE GUIDED SPECTRAL CLASS REJECTION: ASSESSMENT OF POSSIBLE TRAINING DATA PROTOCOLS[1]

John A. Scrivani, Randolph H. Wynne, Christine E. Blinn, and Rebecca F. Musy[2]

**Abstract**—Two methods of training data collection for automated image classification were tested in Virginia as part of a larger effort to develop an objective, repeatable, and low-cost method to provide forest area classification from satellite imagery. The derived forest area estimates were compared to estimates derived from a traditional photo-interpreted, double sample. One method used maplets digitized from ancillary imagery. Seed pixels, the other approach, used only available ground plot data and the image to be classified. Both methods of training data collection resulted in classification accuracy approaching 89 percent, and area estimation precision surpassing the FIA standard of 3 percent per million acres of timberland. However, the precision estimate was met in large part from the additional ground truth data collected supplemental to the national standard sample frame of one plot per 6,000 ac. The seed pixel approach is recommended over maplets, because it does not require ancillary imagery and is less costly in analyst time.

## INTRODUCTION

The Agricultural Research, Extension and Education Reform Act of 1998 called upon the Forest Service to develop and implement a strategy to improve the performance of the Forest Inventory Analysis program. A study by the RAND Corporation recommended that FIA explore utilization of Landsat Thematic Mapper (TM) data for area measurements on a national scale (Peterson and others 1999). Rich Guldin, Director of Science Policy, Planning, Inventory, and Information, USDA Forest Service, recently set the goal of completing " . . . the transition from reliance on aerial photography to use of remotely sensed satellite imagery by the end of FY 2003" (Guldin 2000).

The first phase of this transition to operational satellite image utilization will be its use to produce forest area estimates and provide Phase I stratification for the ground sample.

If classification of raw TM data is to be used on an operational basis for FIA area estimation, two things are required:

1. Image analysis techniques that are low-cost, fast, objective, and repeatable.
2. Standard protocols for the collection of training and validation reference data.

If reference data protocols can be based upon existing field protocols, with little or no modification, this transition can be smooth and cost-effective.

This paper reports on work done in Virginia using an automated classification procedure, Landsat TM imagery, and training data collected from the FIA data sampling frame. The objectives of this work were

1. To further develop an objective, repeatable, and low cost process to obtain forest/nonforest stratification (classification) with TM imagery.
2. To use this stratification in conjunction with Phase II and III ground truth to provide adjusted forest land estimates.
3. To develop more objective, low-cost and effective methods of obtaining training data for use in the classification process.
4. To compare the precision of forest area estimates obtained using classified imagery with those obtained through traditional photo-interpretation methods and double sampling.

## DATA

A Landsat 7 Enhanced Thematic Mapper Plus (ETM+) scene covering eastern Virginia, WRS Path 15, Row 34, acquired on March 3, 2000, was used for this analysis. The scene was geo-rectified using 30 ground control points and a first-order polynomial model. The root mean square error (RMSE) for the geo-rectification model was 11.8 m. For a sample of 10 independent ground control checkpoints, the RMSE was 11.9 m. Spectral bands 1–5 and 7 were used for the analysis.

Ground reference data came from annual forest inventory field measurements made in the years 1997-2000 in Virginia. At the time of analysis, 978 Phase II and 24 Phase III ground plots were available, representing slightly > than three of the five panels of the 5-year annual sampling frame. Also available were land use classifications from 285 deleted plots (ground plots for the last survey that had been dropped from the five-panel system but remeasured by Virginia crews). Precise coordinates from differentially corrected GPS observations were collected for all of these points, with an estimated accuracy of better than 10 m. Also available were 753 intensification plots, where aerial photo-interpreted land use points were verified on the ground by field crews. Coordinates for the intensification plots were digitized using 10 m SPOT panchromatic imagery, dated 1993 to 1994. In total, 2,040 land use ground truth points were available. For collections of training data, 430 Phase II

[1] Paper presented at the Second Annual Forest and Inventory (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, Virginia Department of Forestry, 900 Natural Resources Drive, Charlottesville, VA, and Assistant Professor and Graduate Research Assistants, College of Natural Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA, respectively.

plots and 24 Phase III plots were used. Of the remaining plots, 131 could not be used for validation due to clouds or bad ETM+ data, leaving 1,455 plots for validation. The resulting validation sampling intensity was one point per 4,600 ac. Table 1 summarizes the training and validation ground truth data.

The entire scene was used in the classification process; however, a 30 county subset of the image was used for stratification and land use estimation, because the county is the smallest unit for each that estimates are traditionally derived. The 30 county subset contains 5.2 million ac of land, and was 67.1 percent forested, with 3.38 million ac of forest land in 1992, the date of last survey. The entire scene covers approximately 7.6 million ac, with a similar proportion of forest land.

Available for comparison are preliminary, county-level, forest area estimates obtained using the traditional double-sampling technique with a large sample of photo-interpreted points, corrected with ground truth from both Phase II and intensification points. For the 30 county subset, the large sample of photo-interpreted points was 41,275 in size. The standard errors of these estimates were estimated using the formulae of Li and others (1992).

## ITERATIVE GUIDED SPECTRAL CLASS REJECTION

Iterative Guided Spectral Class Rejection (IGSCR) is a hybrid classification method that builds and labels spectral classes for use in supervised approaches such as the maximum likelihood classifier (Wayman and others 2000). The IGSCR algorithm is, in essence, an objective and guided “cluster busting” (Jensen and others 1987, Rutchey and Vilchek 1994) approach that uses specific rejection criteria and large numbers of training pixels.

The IGSCR method accepts and labels a spectral class when it meets the desired inclusion threshold and rejects it if it does not. In this case, the inclusion threshold required at least 90 percent homogeneity within spectral classes and a minimum of at least 20 training data pixels per class. All pixels in spectral classes meeting the 90 percent homogeneity/minimum pixel test are labeled and removed from the original raw image. The unlabeled pixels from the raw image are then clustered into new spectral classes and the next iteration begins. Each of the iterations increases the number of pixels (and spectral classes) with known identity and decreases the number of unclassified training pixels. Once the iterations are complete (based on user-defined parameters such as the percentage of pixels classified or the classification of all training pixels), the known spectral classes are combined into a single signature file. The pure spectral classes are then used with the maximum likelihood decision rule to classify the image.

## TRAINING DATA

### Maplets

Classifications of forest and nonforest land use, termed maplets, were created via heads-up digitizing for relatively small landscape areas within the scene. The image backdrops used were digital orthophoto quarter quadrangles (DOQQ), obtained from the U.S. Geological Survey. DOQQs with image acquisition dates of 1994-1996 were available for 24 of the existing 26 Phase III plots in the pilot study area. Twenty-four 1 km x 5 km maplets were created, approximately centered on each of the 24 Phase III plots. Three categories (forest, nonforest, and uncertain) were used. Any natural or cultural feature as large as a TM pixel (or that dominates the spectral response of a TM pixel) was digitized. Visual inspection of the Landsat 7 ETM+ imagery (panchromatic, multi-spectral, and pan-sharpened) was conducted to determine whether the area mapped had changed since the date of the DOQQ. Areas that changed were edited.

Table 2 summarizes the amount of training data generated, expressed as percent of the image. Water was not sufficiently represented in the maplet sample so additional training data for water were collected visually from the TM image.

Advantages to the maplet process include (1) their potential utility for the FIA program for other uses and (2) the ability to accurately map areas that are traditionally problematic in TM forest/nonforest classifications such as low-density residential areas and recent harvests. The primary disadvantages are (1) the subjective, analyst intensive nature of the process and (2) the lack of national availability of DOQQs or equivalent imagery. These disadvantages led us to examine a process that has high potential to be objective, repeatable, and highly automated. This process uses a subset of FIA plot centers as “seed pixels” to segment areas of the image into training data.

### Seed Pixels

The seed pixel approach started with a random selection of 500 Phase II plots to be used in obtaining training data.

**Table 1—Ground plot location numbers for validation and training data from Phase II, Phase III, and intensification plots**

| Source | Total | Training | Unusable | Validation |
|---|---|---|---|---|
| Phase II plots | 978 | 255 | 69 | 654 |
| Phase III plots | 24 | 23 | 1 | — |
| Deleted Phase II | 285 | 152 | 18 | 115 |
| Intensification | 753 | — | 67 | 686 |
| Total | 2,040 | 430 | 155 | 1,455 |

**Table 2—Training data amounts for IGSCR forest/nonforest classification of Landsat 7 Scene 15/34, March 3, 2000**

| | Maplet | Seed pixels |
|---|---|---|
| Total (percent of image) | 0.52 | 0.88 |
| Composition (percent of total) | | |
| Forest | 50.1 | 69.6 |
| Nonforest | 15.3 | 10.1 |
| Water | 34.6 | 20.3 |

Seventy of these had to be dropped due to plot location in apparent mixed-land-use pixels, locations under clouds in the imagery, or location on some bad data lines in the image, leaving 430 plots. The sampling intensity for seed pixel initiation was one point per 13,200 ac.

At each of the FIA plots, the analyst located the plot center on the image, visually confirmed the land use call, and initiated a seed pixel. ERDAS Imagine software's seed pixel function works by appending adjacent pixels that are within a specified spectral distance of the mean of the pixels already within the cluster. The only analyst input required is the decision as to what maximum spectral distance should be used. The analyst varied the spectral distance parameter in order to create as large a cluster of pixels that appeared to be clearly within the same land use condition as possible. Development of a more objective and automated approach to seed pixel reference data generation, would help speed this process.

Table 2 summarizes the amount of training data generated by the seed pixel methods as well. Again, water was not adequately represented and the same ancillary water reference data were used as in the maplet training approach.

## AREA ESTIMATION

For the photo-interpreted double sample, forest land percentage estimates and standard errors were computed with the formulae of Li and others (1992). Since the estimates obtained from image classification are "wall-to-wall", or a complete enumeration of the landscape, the double sampling estimates are not appropriate for estimating forest area. Instead, we used the approach for adjusted map marginals formulated by Card (1982).

## RESULTS

Several classifications with different starting parameters, specifically the number of ISODATA classes allowed at the first iteration, were tried and all achieved very similar results. Hence, we will report here only the results of 4 classifications, those starting with up to 300 ISODATA classes. Two such classifications were made of the entire scene, and 2 were made for just the 30 county subset. Within each image extent, one used maplets for training data and the other used seed pixel training data. Whole and subset results were very similar. For simplicity and ease of comparison with double-sample methods, we report only the subset results starting with 300 ISODATA classes.

Table 3 presents the Kappa statistics, overall map accuracy, and producers and user's accuracy for the forest and nonforest classes. Overall map accuracy ranged from 88.5 to 88.8 percent.

Table 4 presents the unadjusted and adjusted map marginal estimates of percent forest land, with standard errors. Also presented are results from the traditional photo-interpreted double-sampling estimation (PI). The PI estimate for the 30 county subset was 64.51 percent forest, with a standard error of 0.82 percent. On a per million acres of forest land basis, the standard error is 1.51 percent, well under the national FIA standard of 3 percent per million acres.

**Table 3—Classification accuracy for forest/nonforest by photo-interpretation, and various IGSCR classifications of Landsat 7 Scene 15/34, March 3, 2000**

| | | Image Subset | |
|---|---|---|---|
| Accuracy | Photo-interpretation | IGSCR maplet | IGSCR seed pixel |
| | - - - - - - - - - - - *Percent* - - - - - - - - - - - | | |
| Overall | 93.8 | 88.8 | 88.9 |
| User's | | | |
| Forest | 94.5 | 87.9 | 87.3 |
| Nonforest | — | 90.8 | 92.6 |
| Producer's | | | |
| Forest | — | 95.3 | 96.4 |
| Nonforest | — | 77.8 | 76.3 |
| Kappa statistic | — | 0.7534 | 0.7541 |

**Table 4—Estimates and standard errors of forest land for 30-county subset of Scene 15/34, from photo-interpreted double sampling and IGSCR image classifications**

| | Photo-interpreted double sample | IGSCR maplet | IGSCR seed pixel |
|---|---|---|---|
| | - - - - - - - - - - *-Percent-* - - - - - - - - - - | | |
| Unadjusted Map marginals | 65.26 | 71.57 | 70.36 |
| Adjusted Map marginals | 64.51 | 65.43 | 64.51 |
| Standard error | 0.82 | 1.06 | 1.05 |
| Standard error (per million acres) | 1.51 | 1.95 | 1.93 |

For the IGSCR classifications, the adjusted map marginal estimates are very close to the PI estimates: 65.43 percent (maplet) and 64.51 (seed pixel). As expected, the standard errors are higher: 1.06 percent (maplet) and 1.05 percent (seed pixel). On a per million acres of forest land basis, at 1.95 percent (maplet) and 1.93 percent (seed pixel), the estimates still surpass the FIA precision goal of 3 percent.

Note the 5 to 6 percent overestimate of forest land by the unadjusted map marginals (table 4). This suggests that the IGSCR method is overclassifying forest. Knowledge of the area and visual inspection of the image suggest that the major problems are urban and suburban areas with tree cover similar to areas of forest land use. This suggests that masking of known urban/suburban areas could improve the accuracy of the IGSCR classifications.

## DISCUSSION

The IGSCR classification method performed well in estimating forest land area using adjusted map marginals. The precision of the estimates exceeded the FIA national standard of 3 percent per million acres of forest land. The method compared favorably to photo-interpreted double sampling, although with some loss of precision.

Within the limits of this case study, the IGSCR method proved to be objective and repeatable. Since this work was still developmental, operational costs were not estimated. However, we feel that its cost should be comparable to, if not considerably less than, photo-interpretation and supervised image classification approaches.

Two different protocols for collection of training data were examined and both performed equally well. Further work on the IGSCR method could possibly improve the current classification accuracy (89-90 percent) to approach that of the photo-interpretation methods (93-94 percent).

The amount of training data collected by either method was less than one percent of the image. Previous IGSCR development work (Wayman and others 2000) used 3 to 6 times the amount of training data, however, the classifier's performance was not any better. This result implies that possibly even less training data could be used. Furthermore, the maplet and seed pixel approaches had significantly different proportions of forest and nonforest training data (see table 2), but that seemed to make no difference in accuracy of classification, either overall or by class.

This work has shown that either maplets or seed pixels can work well as training data. Given the extra work involved in creating maplets, we do not recommend this approach unless the maplets are desired for other reasons, such as examination of landscape patterns. The seed pixel approach requires no additional imagery or ground truth other than a portion of the Phase II plots. Higher precision can be gained by either better classification accuracy or more ground validation points. Focusing on problem classes, e.g. suburban areas and recent harvests, could narrow the accuracy gap.

The IGSCR classifications, and resulting forest land estimates, meet the FIA precision standard of 3 percent per million acres. Much of the credit for reaching this goal, however, should be placed on the additional ground truth provided beyond Phase II and Phase III ground plots. In this study, deleted plots and intensification plots were also used. If this same project had been limited to one ground truth plot per 6,000 ac, the estimated standard errors per million acres of forest land would have been approximately 2.7 percent for the PI double sample and 3.5 percent for the IGSCR classifications. FIA programs should consider continuing, or initiating, land use intensification samples to achieve land use precision goals.

## REFERENCES

**Card, D.H.** 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing. 48(3): 431–439.

**Guldin, R.W.** 2000. Remote sensing at the dawn of a new millenium: a Washington, DC, perspective. Keynote remarks presented at the RS2000 conference, April 13, 2000, Albuquerque, NM.

**Jenson, J.R.; Ramsey, E.W.; Mackey, H.E. Jr.; Christensen, E.J.; Shartz, R.R.** 1987. Inland wetland change detection using aircraft MSS data. Photogrammetric Engineering and Remote Sensing. 53(5): 521–529.

**Li, H.G.; Schreuder, H.T.; Van Hooser, D.D.; Brink, G.E.** 1992. Estimating strata means in double sampling with corrections based on second-phase sampling. Biometrics. 48: 189–199.

**Peterson, D.J.; Resetar, S.; Brower, J.; Diver, R.** 1999. Forest monitoring and remote sensing: A survey of accomplishments and opportunities for the future. Report No. MR–1111.0–OSTP. Science and Technology Policy Institute, RAND Corporation. 92 p.

**Rutchey, K.; Vilchek, L.** 1994. Development of an everglades vegetation map using a SPOT image and the Global Positioning System. Photogrammetric Engineering and Remote Sensing. 60(6): 767–775.

**Wayman, J.P.; Wynne, R.H.; Scrivani, J.A.** 2000. Satellite-assisted forest cover mapping in the southeastern United States using iterative guided spectral class rejection. Proceedings, Second international conference on geospatial information in agriculture and forestry; January 10–12; Orlando, FL: 355–362. Vol. 2.

# LANDSAT TM CLASSIFICATIONS FOR SAFIS USING FIA FIELD PLOTS[1]

**William H. Cooke III and Andrew J. Hartsell**[2]

**Abstract**—Wall-to-wall Landsat Thematic Mapper (TM) classification efforts in Georgia require field validation. We developed a new crown modeling procedure based on Forest Health Monitoring (FHM) data to test Forest Inventory and Analysis (FIA) data. These models simulate the proportion of tree crowns that reflect light on a FIA subplot basis. We averaged subplot crown proportions and compared them to Landsat TM classifications for validation. Resolution differences between field data and Landsat TM data make comparisons challenging. We recorded positive correlations between the two types of data for four of the five FIA plots tested. We attribute differences on the fifth plot to misregistration of the two data sources or misclassification of the TM imagery.

## BACKGROUND

The 1974 Forest and Rangeland Renewable Resources Planning Act (RPA) requires the United States Department of Agriculture Forest Service (USDA FS) to provide Congress with statistics on current forest land and rangeland conditions. The Southern Research Station, Forest Inventory and Analysis Program (SRS-FIA) conducts forest inventories for all Southern States from Virginia to Texas. Except for sparsely forested regions in west Texas and west Oklahoma, forested land in the South has several cycles of field inventories in recent history. SRS-FIA employs a systematic grid of permanent remeasurement plots to help meet these inventory requirements. From these plot measurements sample statistics for numerous variables provide the basis for estimating forest/nonforest conditions. Necessary for expanding plot estimates to county, unit, and State levels is an accurate estimate of forest and nonforest area by county. Currently, we use dot grids with National Aerial Photography Program (NAPP) photos to calculate the proportion of forested land. Multiplying this proportion by the estimate of total land area from Bureau of Census records yields an estimate of the land area in forest and in nonforest condition. Correction factors derived from field plots and from assessments of "intensification" plots improves Phase I estimates of forest area.

FIA wants to reduce the frequency of NAPP photo acquisition or eliminate them entirely. Replacing NAPP photography with the pixel-based approach of Landsat Thematic Mapper (TM) data could achieve similar precision and provide State cooperators with land cover maps. FIA plots may provide a critical link between TM data and actual ground conditions. FIA plots yield more detailed and specific information than can be derived from TM data. This study verifies TM data classifications.

## METHODOLOGY

Field inventories in support of the Southern Annual Forest Inventory System (SAFIS) are currently underway in Georgia. Using hand-held Global Positioning System (GPS) receivers, we connect FIA plot data to "real-world" coordinates and then locate field plots on the TM imagery. A county map of Georgia in figure 1 shows Brantley County, the study site for this methodology.

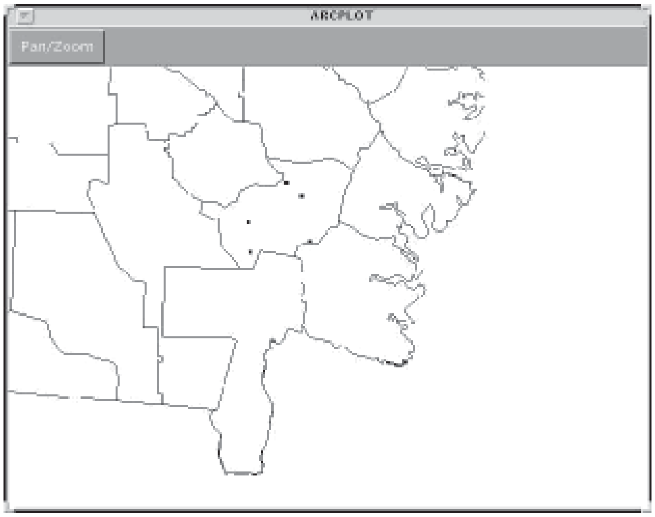


Figure 1—Plots in study site, Brantley County, GA.

Two critical questions arise when we consider FIA plots for remote sensing purposes:

1. How accurately can we locate the FIA plots on the ground and on the TM imagery?  This is a coregistration problem.
2. Which characteristics of the FIA plot data are useful for remote sensing purposes? This is a crown modeling problem.

### Coregistration

Question 1 requires an examination of two sources of registration error—the imagery and the GPS reading on the plot. Problems with accurate coregistration of plots and

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 16–17, 2000.

[2] Research Foresters, USDA Forest Service, Southern Research Station, 201 Lincoln Green, Starkville, MS 39759.

satellite data result from locational errors of the satellite imagery during rectification procedures and errors of the GPS coordinate reading. Figure 2 illustrates the cumulative effect of these error sources. If the error sources are cumulative, FIA subplot 1 (plot center) could be as much as two pixels away from its real-world location.

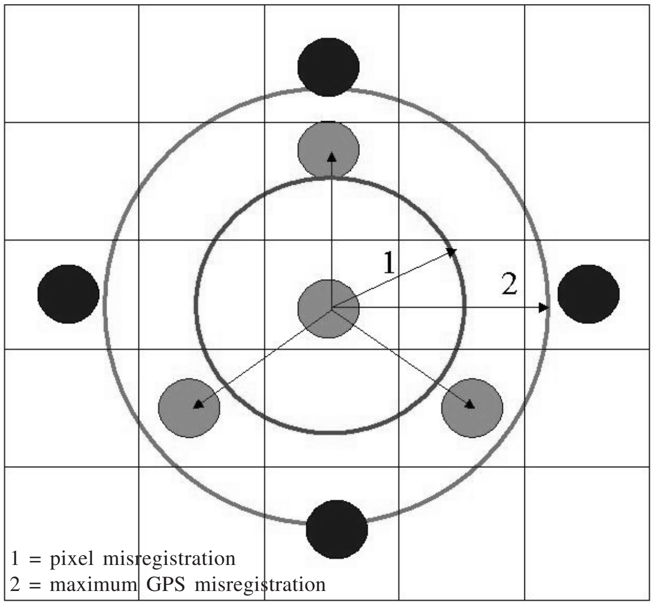


Figure 2—Sources of locational error.

## Crown Modeling

Question 2 presents a challenging problem. The pixel resolution (28.5 m) of TM data restricts the useful level of detail of plot information. Within forested stands, the satellite sensor most likely images dominant, codominant and intermediate trees. More detailed information collected during field sampling [diameter at breast height (d.b.h.), height, etc.] is less useful. Holmgren and Thuresson (1998) point out that satellite images seldom contain enough information to support the decision process in applied forestry. To address information utility, we developed a methodology to compare the individual tree data from FIA field plots with estimates of forest area by a 25-pixel TM window, which is large enough to allow for some of the uncertainty of misregistration.

Based on 304 measurements of trees in New Zealand, Avery (1975) documents a strong linear relationship between d.b.h. and crown diameter for *Pinus radiata* D. Don. This concept was originally designed to predict diameter of trees whose crowns could be measured on aerial photographs. For our study, we developed relationships between measured crown diameter and d.b.h. to enable prediction of crown diameter from d.b.h.

FIA field crews recorded distance and azimuth from each subplot center to each tallied tree. We used this information in a GIS system to provide a geographic reference point for a mechanical reconstruction of the tree crowns on each subplot.

## Data Preparation

We reformatted raw (unedited) plot data from Georgia from ASCII files to a relational database format. We queried individual tree data for these attributes:

1. Crown class (dominant, codominant, intermediate);
2. Species (pine, hardwood);
3. Nonmapped forested plots (edge conditions);
4. No evidence of disturbance; and
5. Live trees with d.b.h. ≥5 in.

Other data preparation included:

1. Assigning pine/hardwood species codes;
2. Computation of each tree location referenced to Universal Transverse Mercator (UTM) coordinates on each subplot based on field measurements of distance and azimuth; and
3. Modeling crown diameter from diameter using Forest Health Monitoring (FHM) data to derive regression coefficients.

We downloaded FHM data from the St. Paul field office site of the Forest Resources Management and Forest Health Protection Web site (http://www.na.fs.fed.us/spfo/fhm/). As the basis for simple linear regressions, these data enabled prediction of crown diameters from d.b.h. For modeling pine crown diameter and hardwood crown diameter, we used 350 observations each. R-square values were .82 and .63 for pine and hardwood prediction models, respectively:

Pine Model: $\text{dbh} * .531225 + 0.0094$
Hardwood Model: $\text{dbh} * .245801 + 2.4555$

We drew crowns at the real-world location of each tallied live tree with d.b.h. ≥5 in. When a tree crown extruded beyond a subplot radius, we terminated that crown at the plot perimeter. Conversely, crowns of trees that intruded on the subplot radius are nontallied trees. We assumed that truncation of extrusive crowns and nontally of intrusive crowns represents a compensating error situation. We ignored crown overlap from a reflectance perspective and performed GIS union operations on overlapping crowns (fig. 3). This ensures that calculation of crown area per plot is a value between 0 and 1. We averaged crown proportion estimates for each subplot for the four subplots to yield crown proportion indices. Resolution differences between the Landsat data and the field data make comparisons difficult.

Figure 4 illustrates the unique problem of comparing field data to image data. To facilitate comparisons, we compared plot index values to 5 by 5 pixel windows on classified Landsat data acquired on December 17, 1996. We calculated proportions for the 5 by 5-pixel window that was most closely centered on the field plot (table 1).

## DISCUSSION

Tables 1 and 2 compare plot and TM. Table 3 references complete breakdowns of crown proportion by subplot.

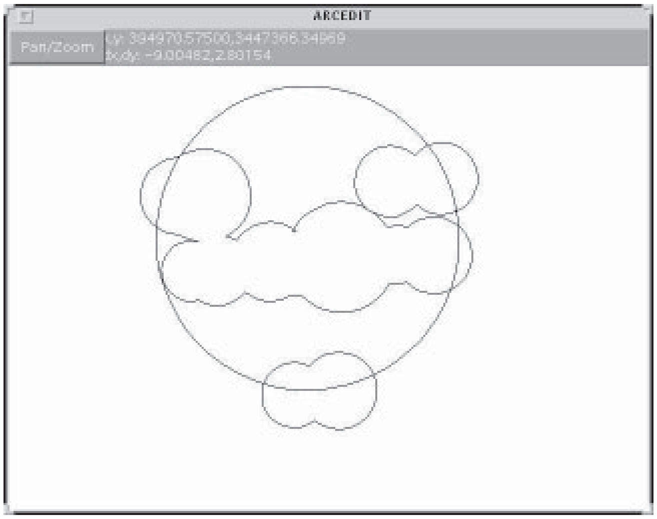


Figure 3—GIS union operation to merge crowns prior to calculating crown proportions.

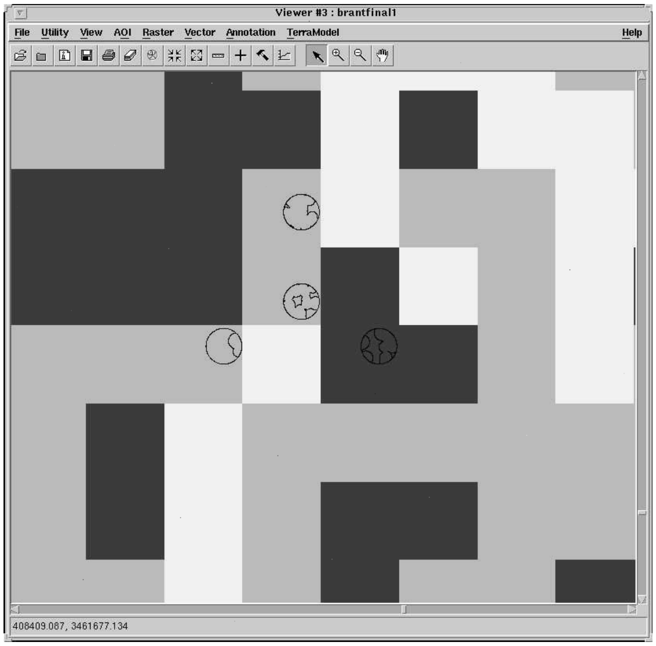


Figure 4—Resolution differences between field plots and Landsat TM imagery.

**Table 1—Comparison of Landsat Thematic Mapper (TM) classification with Forest Inventory and Analysis (FIA) plot data**

| | Plot 1 | | Plot 2 | | Plot 3 | | Plot 4 | | Plot 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Forest type | TM | FIA | TM | FIA | TM | FIA | TM | FIA | TM | FIA |
| | | | | | *Percent* | | | | | |
| Pine ≥ 5 in. d.b.h. | 68 | 100 | 35 | 0 | 80 | 100 | 100 | 100 | 42 | 100 |
| Hardwood ≥ 5 in. d.b.h. | 32 | 0 | 65 | 100 | 20 | 0 | 0 | 0 | 58 | 0 |
| Crown (FIA) | | 54 | | 49 | | 71 | | 41 | | 65 |

**Table 2—Count of trees with d.b.h. < 5 inches**

| Forest type | Plot 1 | Plot 2 | Plot 3 | Plot 4 | Plot 5 |
|---|---|---|---|---|---|
| Pine | 3 | 0 | 1 | 2 | 0 |
| Hardwood | 0 | 14 | 0 | 1 | 12 |

**Table 3—Breakdown of crown proportion by subplot**

| | Plot 1 | | Plot 2 | | Plot 3 | | Plot 4 | | Plot 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Subplot | CA | CP | CA | CP | CA | CP | CA | CP | CA | CP |
| 1 | 81.78 | .49 | 120.24 | .72 | 99.10 | .59 | 73.19 | .44 | 137.89 | .82 |
| 2 | 84.49 | .50 | 30.26 | .18 | 122.03 | .73 | 71.59 | .43 | 147.92 | .88 |
| 3 | 95.49 | .57 | 50.39 | .30 | 121.29 | .72 | 84.64 | .50 | 111.34 | .66 |
| 4 | 104.25 | .62 | 127.11 | .76 | 134.33 | .80 | 48.15 | .29 | 35.88 | .21 |
| Mean CP/plot | .54 | | .49 | | .71 | | .41 | | .64 | |

CA = Crown area per subplot in m2; CP = Crown proportion per subplot calculated by CA/plot area (168.11 m2 ).

### Plot 1

FIA data indicated 100 percent of all trees ≥5 in. d.b.h. were pines. Classified TM data from the 25-pixel window resulted in 68 percent pine and 32 percent hardwood. The mean crown proportion for plot 1 was .54. Table 2 results indicate a fairly even distribution of crowns over the four subplots.

### Plot 2

FIA data indicated 100 percent of all trees ≥5 in. d.b.h. were hardwoods. Classified TM data from the 25-pixel window resulted in 35 percent pine and 65 percent hardwood. The mean crown proportion for plot 2 was .49. Table 2 results show an uneven distribution of crowns over the four subplots. Subplots 1 and 4 are have more than 70 percent crown saturation and subplots 2 and 3 have less than 30 percent crown saturation. Table 2 records 14 hardwoods <5 in. d.b.h., which indicates possible hardwood reflectance from untallied trees on this plot.

### Plot 3

FIA data indicated 100 percent of all trees ≥5 in. d.b.h. were pines. Classified TM data from the 25-pixel window resulted in 80 percent pine and 20 percent hardwood. The mean crown proportion for this plot was .71. Subplots 2, 3, and 4 have more than 70 percent crown saturation and subplot 1 has more than 60 percent crown saturation. Subplot 1 is relatively homogeneous, and the TM results are in agreement with a homogeneous land cover situation.

### Plot 4

FIA data indicated 100 percent of all trees ≥5 in. d.b.h. were pines. Classified TM data from the 25-pixel window resulted in 100 percent pine. The mean crown proportion for plot 4 was .41. Distribution of crown saturation across the subplots is fairly consistent except for subplot 4, which has less than 30 percent crown saturation. Table 2 indicates that there are only two pines and one hardwood with unmodeled crowns on this plot. Since crown saturation is low, it would be interesting to know what features of the landscape are causing pure pine classification results.

### Plot 5

FIA data indicated 100 percent of all trees ≥5 in. d.b.h. were pines. Classified TM data from the 25-pixel window resulted in 42 percent pine and 58 percent hardwood. The mean crown proportion for plot 5 was .64. Subplots 1 and 2 had more than 80 percent crown saturation. Subplot 3 had more than 60 percent crown saturation and subplot 4 had roughly 20 percent crown saturation. Two possible reasons for the nonagreement between FIA and TM results are pixel/plot misregistration or incorrect classification results. Examination of the classified imagery reveals that a one-pixel shift to the northwest would result in 60 percent pine and 40 percent hardwood. High pine crown proportions in subplots 1 and 2 further strengthen the argument for misregistration. Results shown in Table 2 strengthen the argument for incorrect classification results. Twelve hardwood trees <5 in. d.b.h. that were not modeled for canopy proportion estimates and the location and diameter of these stems/crowns should have been modeled. If the majority of these trees are growing beneath the overstory, misregistration is likely. If the majority of these trees are growing in dominant canopy positions, misclassification is likely.

## CONCLUSIONS AND RECOMMENDATIONS

Resolution differences between the FIA field data and the TM data show that we are attempting to “compare apples and oranges.” On the basis of our limited study, there appears to be good correlation between the results of the modeled canopies and the TM classification. However, misregistration and misclassification errors are difficult to quantify. Excluding stems <5 in. d.b.h. from the crown modeling process was a mistake. In future modeling efforts, if tallied stems <5 in. d.b.h. are overtopped, we will not model them on the basis of the canopy position constraint. If stems <5 in. d.b.h. are in a dominant, co-dominant, or intermediate crown position we will model them. This methodological change should provide useful information on plot surface reflectance. We could bridge resolution problems between the two data sources by using LIDAR data or large-scale aerial photography.

This is a preliminary study designed to test the usefulness of FIA plot data for verifying Landsat TM classifications. Now that methodologies are established and automated, numerous plots will be tested.

Finally, new canopy prediction models being tested include species, age, density, crown class, landscape position, and other variables as possible predictors of crown size. These models may improve quantification of crown proportion estimates by subplot.

## REFERENCES

**Avery, T.E.** 1975. Natural resources mesurements. New York: McGraw–Hill. 339 p.

**Holmgren, P.; Thuresson, T.** 1998. Satellite remote sensing for forestry planning—a review. Scandinavian Journal of Forest Research. 13: 90–110.

# STRATIFYING FIA GROUND PLOTS USING A 3-YEAR OLD MRLC FOREST COVER MAP AND CURRENT TM DERIVED VARIABLES SELECTED BY "DECISION TREE" CLASSIFICATION[1]

**Michael Hoppus, Stan Arner, and Andrew Lister[2]**

**Abstract**—A reduction in variance for estimates of forest area and volume in the state of Connecticut was accomplished by stratifying FIA ground plots using raw, transformed and classified Landsat Thematic Mapper (TM) imagery. A US Geological Survey (USGS) Multi-Resolution Landscape Characterization (MRLC) vegetation cover map for Connecticut was used to produce a forest/non-forest map derived from a classified 1993 TM image. A 1996 TM image was used to provide spectral reflectance variables for each pixel, including the values for all 6 raw TM bands and several transformed layers: normalized difference vegetation index (NDVI) and Tasseled Cap brightness, greenness, wetness, and "fourth." Each pixel in the map was assigned a value indicating how many surrounding pixels within a 3X3 or 5X5 window were forested. These same windows were used to calculate, for each pixel, a mean, maximum, minimum, and standard deviation of the raw and transformed layers. FIA ground plots (1996) were split into percent timberland classes using a "decision tree" algorithm that recursively determines the most significant variable and the most significant split of that variable. The final set of grouping criteria was used to statistically stratify a set of FIA ground plots. Results were compared with aerial photo based stratification as well as TM derived forest/non-forest stratification.

## BACKGROUND

Forest Inventory and Analysis (FIA), a program of the USDA Forest Service, is responsible for the national forest inventory and monitoring of the United States. Congress mandates, through the Forest and Rangeland Renewable Resources Planning Act of 1974 and the McSweeny-McNary Forest Research Act of 1928, that FIA continuously determine the extent, condition, and volume of timber, growth, and depletions of the Nation's forest land. In the East, FIA inventories must meet specified sampling errors: a three-percent error per one million acres of timberland is the maximum allowable sampling error for area (Hansen and others 1992). Until now, FIA has reached this accuracy in part by statistically stratifying the FIA ground plots using aerial photos. However, the Agricultural Research, Extension, and Education Reform Act of 1998 (PL 105–185) directs all FIA units to change from an inventory frequency of 10–14 years per state to an annual inventory system that ground samples 20 percent of each state per year (Gillespie 1999). This new inventory design requires plot stratification every five years.

## OBJECTIVE

The Northeastern FIA unit, responsible for surveying the 13 northeastern states, uses aerial photos from the National Aerial Photography Program (NAPP) for FIA ground plot stratification. NAPP currently is on a seven year cycle. The high cost of additional qualified photo interpreters necessary to complete aerial photo stratification in all the states on a five year cycle plus the seven year cycle of NAPP has led to investigations of the use of satellite imagery to stratify the ground plots.

The objective of this study was to stratify FIA ground plots into "percent timberland per plot" classes using variables objectively selected from a large pool of potentially effective stratifiers. Selection would be made by a "decision tree" algorithm that recursively determines the most statistically significant variable and the partition of that variable with the highest level of significance. Over one hundred different images and forest cover maps derived from Landsat TM scenes, considered strongly correlated with forested landscapes, were subjected to this decision tree selection method. Additionally, several non-satellite variables that are often highly correlated with forest cover were added to the assemblage of potential stratifiers. An important aspect of our study was the inclusion of two forest cover maps classified from Landsat TM and produced by USGS. The potential cost efficiency of using existing satellite based forest cover maps to stratify the ground plots generated much interest in comparing these maps with other products.

This study begins to explore the hypothesis that the statistically most significant predictor variables for percent timberland may also be used to successfully stratify the plots in order to reduce the variance of estimates of total timberland area and tree volume. The final selection of predictor variables made by the decision tree algorithm to group the FIA plots into percent timberland classes was used to form the strata for a timberland area estimate. Finally, an important objective of this study was to compare sampling errors of state level estimates of percent timberland with other stratification efforts.

[1] Paper presented at the Second Annual Forest Inventory and Analyis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, Supervisory Forester, and Forester, USDA Forest Service, Northeastern Research Station, 11 Campus Blvd., Suite 200, Newtown Square, PA 19073, respectively.

## METHOD

### Study Area

The state of Connecticut was used for this study because all the ground plots were located using Global Positioning System (GPS) and there was a mostly cloud free Landsat TM scene available that was acquired the same year that the plots were measured. Additionally, two USGS forest/non-forest maps are available for the state.

### Selection of Predictor Variables

Four different types of landscape level variables were made available to the decision tree algorithm for use as predictors for plot level variables: Landsat TM satellite imagery, including the raw bands and vegetation index images; Classified forest/non-forest maps derived from Landsat TM imagery; pixel neighborhood texture maps; and non-satellite variables (table 1).

**Table 1—Landsat TM derived and other predictor variables provided to the decision tree algorithm as candidates to predict classes of the response variable, percent timberland per plot**

| Predictor variable | Date of TM scene | Source |
|---|---|---|
| TM raw images and vegetation indexes: | | |
| All six raw TM bands | 8/1996 | USGS |
| *Normalized Difference* Vegetation Index | 8/1996 | USGS (scene) /NE FIA |
| *Tasseled Cap Transformation-* Brightness, Greenness, Wetness, Fourth | 8/1996 | USGS (scene) /NE FIA |
| TM derived forest/non-forest maps: | | |
| *Multi-Resolution Landscape* Characterization forest map | 8/1993 | USGS |
| GAP forest map | 8/1993 | USGS |
| *Normalized Difference* Vegetation Index threshold forest map | 8/1996 | USGS (scene) /NE FIA |
| Moving window filter images: | | |
| 3X3 pixel window algorithms for unclassified variables: minimum, maximum, mean, standard deviation | 8/1996 | USGS (scene) /NE FIA |
| 5X5 pixel window algorithms for unclassified variables: minimum, maximum, mean, standard deviation | 8/1996 | USGS (scene) /NE FIA |
| 3X3 pixel window algorithm for classified variables: total forested pixels | 8/1993&96 | USGS (scene) /NE FIA |
| 5X5 pixel window algorithm for classified variables: total forested pixels | 8/1996&96 | USGS (scene) /NE FIA |
| Non-TM variables: | | |
| Elevation | 30m | USGS 1:24000 DEM |
| Slope | 30m | USGS 1:24000 DEM |
| Precipitation | 4km | PRISM– www.ocs.orst.edu/prism/prism_new.htm |
| Soil permeability | 1km | STARTSGO database |
| Soil bulk density | 1km | STARTSGO database |
| Length of roads <1/2km | 30m | TIGERLINE road file- www.census.gov |
| Length of roads <1km | 30m | TIGERLINE road file- www.census.gov |

Five different images were produced from vegetation index algorithms applied to all or part of the raw TM bands. There are a number of algorithms used to extract information such as; biomass, leaf area index, and percent vegetative ground cover, which are called Vegetation Indexes (VI). These algorithms reduce the multiple bands in a TM image to a single number per pixel that predicts vegetation characteristics (Jensen 1996). The hypothesis is that forest cover falls within a certain well-defined region of a given VI map based on the "brightness" value of pixels. One of the more common VI's used is the NDVI, which makes use of the ratio between reflected near-infrared light and red light (and others 1973, Larsson 1993). Other VI's evaluated were the layers derived from the Tasseled Cap transformation (Crist and Cicone 1984).

One of the three forest/non-forest maps evaluated was produced from the NDVI image. Based on our analysis, the higher the pixel's brightness value the more likely it was to cover a forested area on the ground. The NDVI map was "thresholded" at a certain brightness level whereby all pixels above this level were classified as forest and those pixels below that level were classified as non-forest. The threshold level that provided the most accurate map when compared with aerial photos was selected for the final NDVI threshold map.

Forest/non-forest maps were also acquired from Gap Analysis Program (GAP) and National Land Cover Data (NLCD) (formerly Multi Resolution Landscape Characterization (MRLC)) vegetation cover maps for the Connecticut study area. Both of these products are sponsored and coordinated by the USGS and are designed to provide a map of current land cover types over the U.S (Scott and Jennings 1998, Jennings 1993). These maps are based on TM classification and differ from each other and from other TM images for a variety of possible reasons, including; differing dates and quality of TM imagery used, different classification methods applied, differing minimum mapping unit, and differing definitions of forest land employed.

An evaluation of FIA ground plot geometry and the locational uncertainty of both TM pixels and plots due to image registration errors and GPS errors, respectively, suggests that images which quantify pixel values within a 3X3 or 5X5 pixel window may be highly correlated with percent timberland totals for the four subplots that make up an FIA ground plot (fig. 1). Moving window filters applied to forest/non-forest maps produce images where the value of each pixel is equal to the sum of forested pixels within the local pixel neighborhood. Plots stratified with these "filtered" images result in estimates of timberland with lower variance (Hoppus and others 2000, Riemann and others 2000). Calculated variables for the 3X3 and 5X5 moving window filters of the unclassified images include the minimum, maximum, mean and standard deviation of the window values.

Elevation, slope, precipitation, soil permeability, soil bulk density, and the length of roads within 0.5 km and 1.0 km were also provided as predictor variables for percent timberland per plot. They were compared to the satellite based variables by the decision tree algorithm.

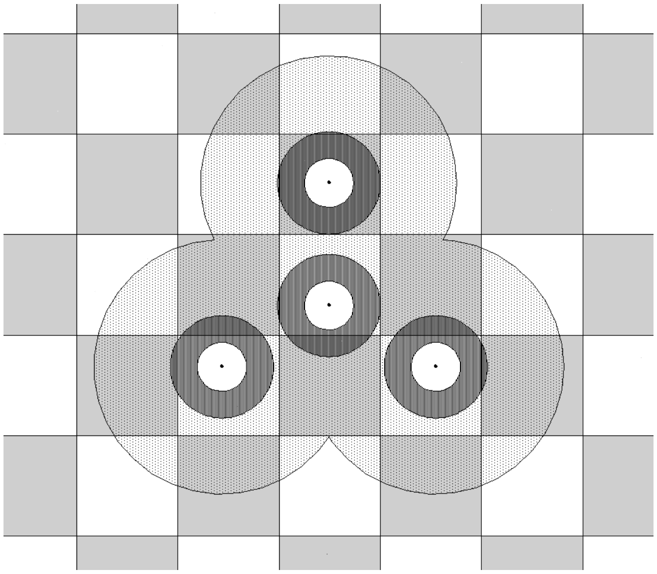


Figure 1—The FIA ground plot geometry versus 30m TM pixels. The plot consists of a cluster of four 0.017 ha subplots. The dark grey circles represent the area of locational error due to GPS errors. The larger grey circles represent the potential locational error due to image registration.

**Defining Classes Using the Decision Tree Algorithm**

FIA ground plots, measured in 1996 throughout the state of Connecticut, were split into percent timberland classes using a "decision tree" algorithm that recursively determines the most significant predictor variables and the most significant splits of each variable based on other predictor variables. The term recursive refers to any mathematical procedure in which any element is computed systematically from the one preceding it. The final set of grouping criteria was used to statistically stratify a somewhat independent set of FIA ground plots.

The software package used to select significant variables is based on statistical procedures described in a paper by Biggs et al 1991. The procedure begins by grouping all observations of the response (or dependent) variable based on each of the predictor (or independent) variables available to the decision tree. Continuous predictable variables are first partitioned into 10 equal-sized intervals. The classes of each predictor variable are then recursively combined by selecting the pair of classes that are most similar based on a *F-test*. The most significant of these combined groupings is then determined. After a Bonferroni adjustment to the significance level to account for the number of classes for each variable, these "best" groupings for all predictor variables are then compared to determine the most significant variable. The population of plots is then split according to the best variable grouping if the significance level $p$ is less than a predetermined value ($p$=0.01 for this study).

For example, the continuous response variable, percent timberland per plot, is first split arbitrarily into 10 classes by consecutive groups of values of one of the predictor variables (fig. 2). Each class has nearly equal numbers of

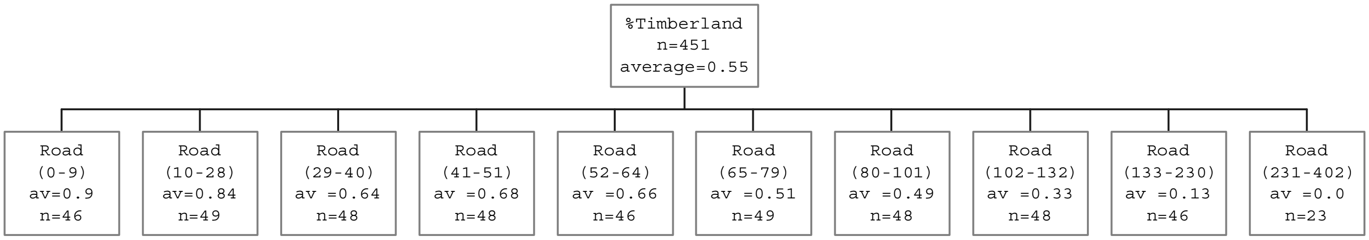

%Timberland
n=451
average=0.55
Road (0-9) av=0.9 n=46
Road (10-28) av=0.84 n=49
Road (29-40) av =0.64 n=48
Road (41-51) av =0.68 n=48
Road (52-64) av =0.66 n=46
Road (65-79) av =0.51 n=49
Road (80-101) av =0.49 n=48
Road (102-132) av =0.33 n=48
Road (133-230) av =0.13 n=46
Road (231-402) av =0.0 n=23

Figure 2—The decision tree algorithm first splits the continuous response variable into 10 classes of approximately equal size by consecutive groups of values of each of one of the predictor variables. Here the predictor variable, roads within 1 km, splits the response variable, percent timberland per plot.

observations. An *F* significance test is applied to each adjacent pair of classes to determine if these groups of the response variable are statistically different based on a selected *p* value of 0.01. Classes found not to be statistically different are merged (fig. 3). An *F* significance test, with a Bonferroni adjustment to account for the number of classes, for each response variable grouping is used to determine the most significant predictor variable. The decision tree selects the predictor variable with the lowest *p* value for each generation of classes.

Each class created by the decision tree based on the most significant predictor variable is also split, if possible, by each of the remaining predictor variables. The decision tree algorithm applies the *F* significance test to this next generation of classes. This process is repeated for each generation of classes until the split results in too few observations (specified by the operator at 10) or the level of significance is reached.

**Building the Stratification Model**

Ten random samples of 50 percent of the FIA ground plots in the state of Connecticut (226 plots) were split into percent timberland classes by predictor variables using the decision tree algorithm. The predictor variable selected as the most significant for the first generation split was noted in each case. The “filtered” image produced from summing the forested pixels in a 5X5 moving window applied to the MRLC forest/non-forest map was the most significant variable for the first split - six out of ten times. The maximum algorithm for a 3X3 moving window filter applied to raw TM band three (red light) was selected twice. The minimum NDVI algorithm for a 3X3 moving window filter was selected once as was the 3X3 window filter for the MRLC forest cover map.

The most common predictor variable selected for each generation of splits from the random samples was chosen for the model. The predictor variable values that defined the timberland classes were determined by taking an average from the samples.

The final model was then applied to all the FIA plots in the state of Connecticut. The chosen predictor variables were used to group the plots into percent timberland classes or strata. The total area of the state defined by each of the strata was calculated and used to weight the plot classes to estimate the total timberland in the state.

## RESULTS

The final decision tree classes of “percent timberland per plot” were created by two generations of predictor variable splits. The MRLC forest/non-forest map, filtered by a 5X5 pixel window that counted total forested pixels, was used for the first generation split. Three of the four classes in the first generation split were in turn split by images created from moving window filters. Two of the classes were split by the brightness values of an image created by applying a “minimum” 5X5 pixel moving window filter to the NDVI image. The timberland class defined by the highest numbers of MRLC forested pixels was split by the image created from applying a “standard deviation” 5X5 pixel moving window filter to the raw TM band three (fig. 4). This combination of predictor variables resulted in an R-squared of 0.61.

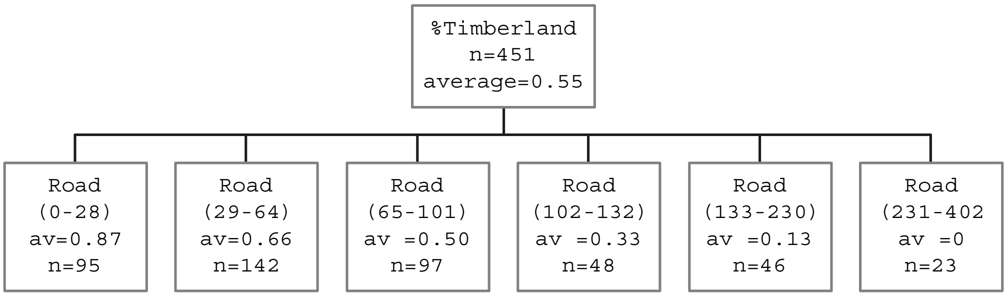

%Timberland
n=451
average=0.55
Road (0-28) av=0.87 n=95
Road (29-64) av=0.66 n=142
Road (65-101) av =0.50 n=97
Road (102-132) av =0.33 n=48
Road (133-230) av =0.13 n=46
Road (231-402 av =0 n=23

Figure 3—Groups of the response variable are merged when they are found to be statistically similar by an *F*-test. Here seven of the original 10 classes of the response variable, percent timberland per plot, have been merged into three, while only three of the original classes remain unchanged.

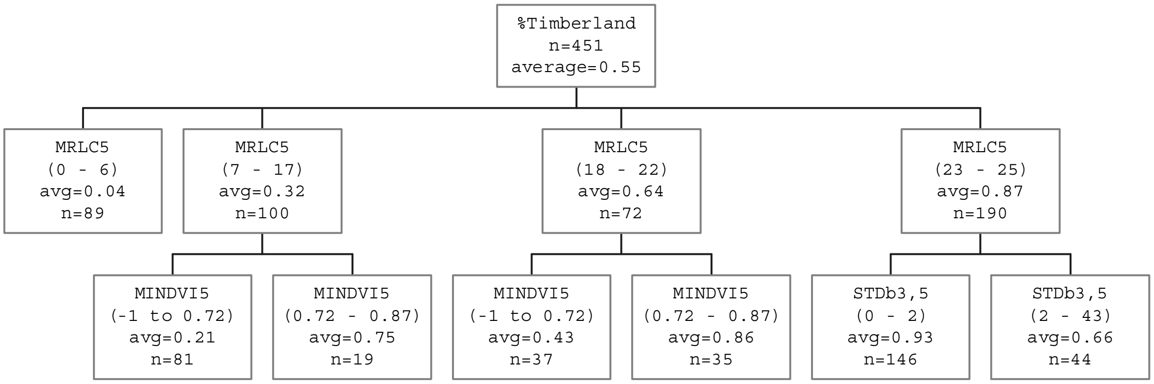


Figure 4—Final decision tree model for classes of percent timberland for all Connecticut ground plots. Each box shows the predictor variable, predictor variable class boundaries, average percent timberland, and the number of plots.

Stratified plot estimates of percent timberland and net cubic foot volume, using the model classes as the strata, had smaller sampling errors than estimates stratified by photo-interpreted timberland/non-timberland plots or any of the unfiltered forest/non-forest maps produced from classified TM satellite imagery. The model did not perform as well as photo plots interpreted for six categories of volume as well as timberland cover. Finally, the model did not stratify the plots such that the sampling error was three percent or less per million acres of timberland: a sampling error of 2.3 percent is required for the approximately 1.6 million acres of timberland in the state of Connecticut (table 2).

The MRLC forest/non-forest map filtered for total forested pixels by a 5X5 moving window was selected as the significant predictor variable for the first generation split of the plots, indicating that the USGS product shows promise as a tool for FIA plot stratification. The fact that nearly all of the significant predictor variables were based on 3X3 or 5X5 filters indicates that the geometry match between the plots and the TM pixels requires a measurement of each pixel's neighborhood for best results.

## CONCLUSION

The decision tree algorithm selected predictor variables that split the response variable, percent timberland per plot, into classes capable of producing stratified estimates of total timberland in the state of Connecticut with less sampling error than any other satellite based strata tried so far. The technique is relatively objective and based on a logical hypothesis that predictor variables that are highly correlated with ground plot variables should provide useful strata for population estimates. In any case, without the decision tree algorithm to look at all the numerous combinations of predictor variables, this particular set of predictor variables and class boundary values would have never been selected.

## REFERENCES

**Biggs D.; DeVille. B.; Suen, E.** 1991. A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics. Vol. 18. No. 1. 49–62.

**Crist, E.; Cicone, C.** 1984. Application of the Tasseled Cap concept to simulated Thematic Mapper data. Photogrammetric Engineering & Remote Sensing. 52(1): 81–86.

**Table 2—Stratified plot estimates and percent sampling error for total timberland area and net cubic-foot volume for the state of Connecticut. The stratifiers include the decision tree model (Model); photo-interpreted timberland/non-timberland plots (PI2); photo-interpreted volume plots (PI7); an unfiltered MRLC forest/non-forest map (MRLC F/NF); the model with just the first split based on the filtered MRLC forest/non-forest map (MRLC5); and estimates based on unstratified ground plots (None)**

| Stratifier | Estimated timberland | Sampling error | Estimated cubic-ft vol | Sampling error |
|---|---|---|---|---|
| | *Thousand acres* | % | *Million* | % |
| Model | 1,667 | 2.59 | 3,158 | 4.01 |
| PI2 | 1,690 | 2.65 | 3,229 | 4.33 |
| PI7 | 1,715 | 2.31 | 3,256 | 3.64 |
| MRLC F/NF | 1,658 | 3.24 | 3,133 | 4.62 |
| MRLC5 | 1,673 | 2.85 | 3,178 | 4.15 |
| None | 1,699 | 3.74 | | |

**Gillespie, A.** 1999. Rational for a national annual forest inventory program. Journal of Forestry. 97(12): 16–20.

**Hansen, M.; Frieswyk, T.; Glover, J.; Kelly, J.** 1992. The eastwide forest inventory data base: users manual. Gen. Tech. Rep. NC–151. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station. 48 p.

**Hoppus, M.H.; Riemann, R.H.; Lister, A.J.** 2000. Remote sensing strategies for forest inventory analysis utilizing the FIA plot database. Eighth forest service biennial remote sensing applications conference, ASPRS, Albuquerque, NM. 2000 April 10–15.

**Jennings, M.** 1993. Natural terrestrial cover classification: assumptions and definitions. GAP Analysis Technical Bulletin 2, Moscow, ID: U.S. Fish and Wildlife Service.

**Jensen, J.** 1996. Introductory digital image processing: a remote sensing perspective. Upper Saddle River, NJ: Prentiss Hall: 179–195.

**Larsson, H.** 1993. Regression for canopy cover estimation in acacia woodlands using Landsat TM, MSS, and SPOT HRV XS data. International Journal of Remote Sensing. 14(11): 2129–2136.

**Riemann, R.H.; Hoppus, M.L.; Lister, A.J.** 2000. Using arrays of small ground sample plots to assess the accuracy of Landsat TM-derived forest-cover maps. Accuracy 2000, 4th International symposium on spatial accuracy assessment in natural resources and environmental sciences. 2000 July 12–14; Amsterdam, The Netherlands: 541–548.

**Rouse, J.; Haas, J.; Schell, J.; Deering, D.** 1973. Monitoring vegetation systems in the Great Plains with ERTS. Proceedings, 3rd ERTS symposium: 48–62. Vol.1.

**Scott, J.; Jennings, M.** 1998. Large-area mapping of biodiversity. Annals of the Missouri Botanical Garden. 85: 34–47.

# HONEYCOMBING THE ICOSAHEDRON AND ICOSAHEDRONING THE SPHERE[1]

**Joseph M. McCollum[2]**

**Abstract**—This paper is an attempt to trace the theoretical foundations of the Forest Inventory and Analysis and Forest Health Monitoring hexagon networks. It is important in case one might desire to alter the intensity of the grid or lay down a new grid in Puerto Rico and the U.S. Virgin Islands, for instance. The network comes from tessellating an icosahedron with hexagons and projecting those hexagons to a sphere. The paper proposes a sample network for Puerto Rico and the U.S. Virgin Islands.

## INTRODUCTION

Pardon the title; it is bad grammatical form to verb a noun. It is also geometrically impossible to square the circle using classical means. There will always be a few chords from the circle left over. Using the method that follows, one will find that it is impossible to completely tessellate a sphere with regular hexagons. There will be twelve pentagons left over. Historically, Forest Inventory and Analysis (FIA) plots in the American South have been laid out on a square grid or no grid at all—that is, haphazardly. Forest Health Monitoring (FHM) plots have been laid out on a hexagon network. Hexagons, squares, and triangles tile the plane (or any study area on earth). Carr and others (1999) list a set of criteria for global grid cells that argue in favor of hexagons; among other things, hexagons provide maximum area for minimum perimeter. The astrophysicist Max Tegmark (1996) listed similar criteria and built a similar grid for the sky. In geometry, squaring the circle means attempting to rearrange a circle to form a square; and in this paper, honeycombing an icosahedron is attempting to fit a honeycomb pattern on top of an icosahedron. To accomplish this task, start with an unfolded icosahedron, as shown in figure 1. Ultimately, this icosahedron shall be projected to the sphere of the earth, as shown in figure 2.

An icosahedron is a geometric solid with 20 faces, all of which are equilateral triangles. Throughout this paper, the "poles" of the icosahedron will be points at the very top and very bottom of the unfolded solid. The "cuts" will be those line segments connected to the poles, and the "ends" of the "cuts" will be the points at which the cuts join. A similar method will work with an octahedron, but the icosahedron approximates a sphere better than any other platonic solid. Tesselate each face with nine triangles. Take six of the triangles from one face to form a hexagon. There will be three triangles on each face that ultimately form parts of pentagons. The result is a solid with 32 faces, 20 of which are hexagons and 12 of which are pentagons. Geometers

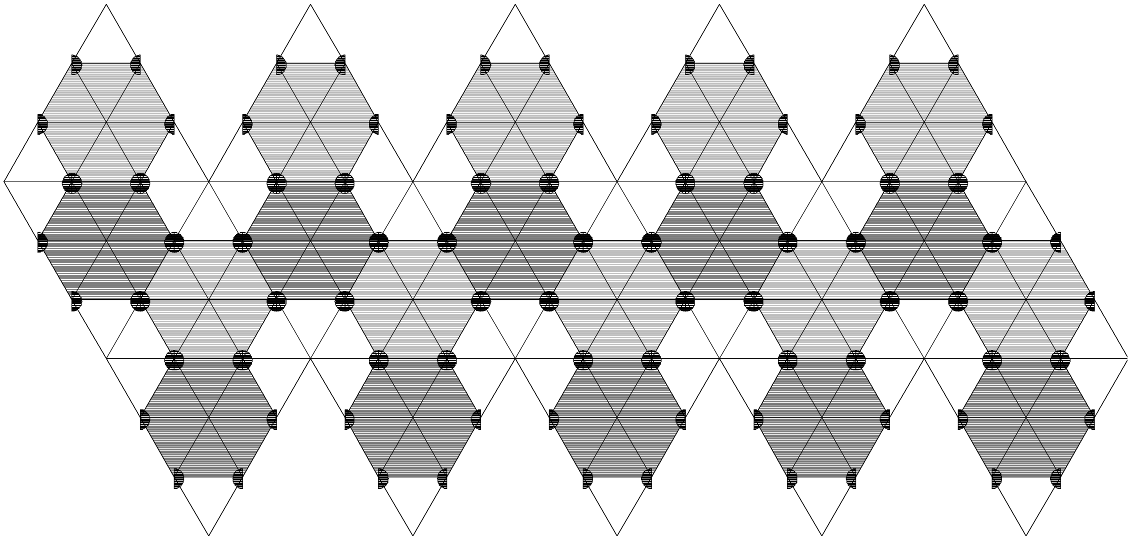


Figure 1—Icosahedron tessellated with hexagons.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Computer Specialist, USDA Forest Service, Forest Inventory and Analysis, Southern Research Station, Asheville, NC 28804.

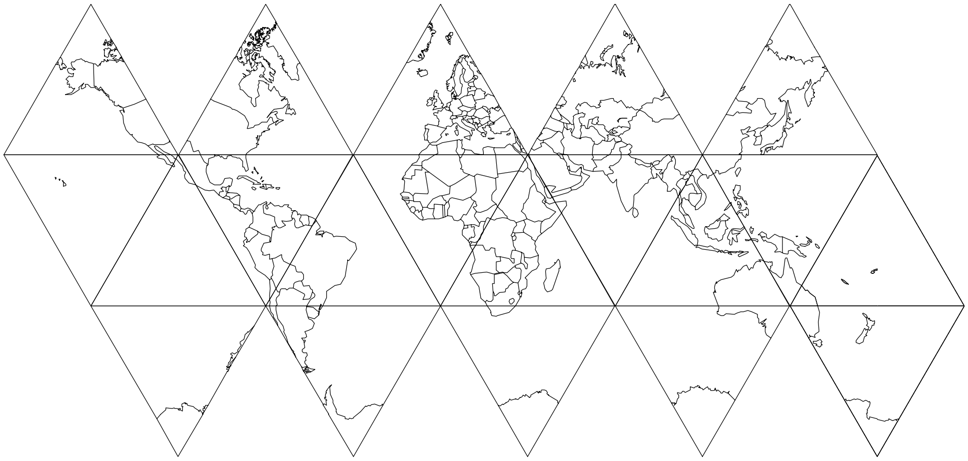


Figure 2—World map projected on an unfolded icosahedron.

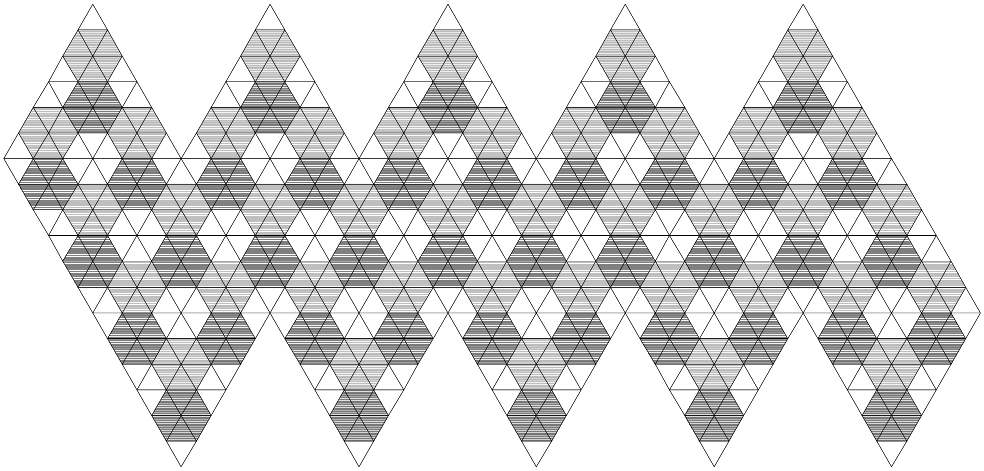


Figure 3—The second-order tessellation of the icosahedron with hexagons using the triangle orientation.

know this solid as a truncated icosahedron, athletes know it as a soccer ball; and although the reader may not be aware, virologists know it as the structure of a virus with triangulation number T = 3 (Johnson and Speir 1997). Figure 1 has sixty vertices marked with bold circles, and chemists know this structure as $C_{60}$ (Maggio 1994).

### What If More Hexagons Are Desired?

As indicated by Carr and others (1999) and Tegmark (1996), hexagons are useful because they form an ideal network by which to divide a study area into smaller areas from which to draw plots. Under the Lambert Azimuthal Equal Area projection [used in the Snyder (1992) model of this method], the default radius of the earth is 6,370,997 m. A perfect sphere of this radius has a surface area equal to $5.10 \times 10^{14}$ m$^2$. This number is about $1.26 \times 10^{11}$ ac. The icosahedron in figure 1 must be subdivided to match the scale of an FIA plot. There are 5,936 ac per FIA plot, which compared to the surface area of the earth is one part in 21 million. Each large FHM hex has 27 FIA hexes and, compared to the surface area of the earth, is 1 part in 786,000. If one wished to construct sixteenfold FIA hexes (for a more intense FHM grid), one would need a cell that is 1 part in 1,327,000. The short answer of how to create more hexagons is to create more triangles, as shown in figure 3.

Instead of the 20 hexagons and 12 pentagons in figure 1, we now have 110 hexagons and 12 pentagons. This method will work elegantly if there are $9n^2$ triangles on each face, for a total of $180n^2$ triangles. Twelve pentagons will consume 60 triangles, leaving enough triangles for $30n^2 - 10$ hexagons. Note that if n = 162, then $30n^2 - 10$ is near 786,000; if n is around 840, then $30n^2 - 10$ is around 21 million. An icosahedron that circumscribes the earth has a surface area about 14.6 percent larger than that of the sphere—one may wish to use this surface area instead of that of the earth. More exact numbers appear in table 1.

**Table 1—Number and size of polygons under the triangle orientation of various resolutions**

| Order | Triangles | Hexagons | + | Pentagons | = | Polygons | Area of triangle | Area of hexagon |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | *m*$^2$ | *m*$^2$ |
| 1 | 180 | 20 | | 12 | | 32 | 2.83 x $10^{12}$ | 1.70 x $10^{13}$ |
| 2 | 720 | 110 | | 12 | | 122 | 7.08 x $10^{11}$ | 4.25 x $10^{12}$ |
| 3 | 1,620 | 260 | | 12 | | 272 | 3.15 x $10^{11}$ | 1.89 x $10^{12}$ |
| 162 | 4,723,920 | 787,310 | | 12 | | 787,322 | 1.08 x $10^8$ | 6.48 x $10^8$ |
| 840 | 127,008,000 | 21,167,990 | | 12 | | 21,168,002 | 4.02 x $10^6$ | 2.42 x $10^7$ |
| 841 | 127,310,580 | 21,218,420 | | 12 | | 21,218,432 | 4.01 x $10^6$ | 2.41x $10^7$ |
| 842 | 127,613,520 | 21,268,910 | | 12 | | 21,268,922 | 4.00 x $10^6$ | 2.40 x $10^7$ |
| 843 | 127,916,820 | 21,319,460 | | 12 | | 21,319,472 | 3.99 x $10^6$ | 2.39 x $10^7$ |

## HEXAGON HIERARCHIES

In a tessellation of hexagons, one hexagon is surrounded by six others, as in figure 4. If one connects the centers of the outer ring of six, one produces a hexagon three times as large as the original hexagon. One may continue in this manner to get hexagons of size 3, 12, 27, 48, . . . 3n$^2$, where n = 1,2,3, . . .

If one connects opposite vertices of the ring of six as shown in figure 5, one produces a hexagon four times as large as the original hexagon. One may continue in this manner to get compositions of size 1, 4, 9, 16, 25, . . . n$^2$, again where n = 1,2,3, . . .

If one makes a ring of six clusters of seven hexagons around another cluster of seven hexagons, as shown in figure 6, one produces a hexagon 21 times as large as a basic hexagon. This maneuver is called a sevenfold composition. The large hexagon is seven times larger than a threefold hexagon. One may continue in this manner to get compositions of size 1, 7, 19, 37, . . . 3n$^2$ - 3n + 1, where n = 1, 2, 3.

If one relaxes the constraint of requiring the vertex to be in the center hexagon of the cluster and allows the vertex to be in the center of any hexagon of the cluster, then one gets another family of compositions and decompositions, such as in figure 7.

## The Chevron and Intermediate Orientations

The structures in figures 1, 2, and 3 are called the “triangle” orientation. What if one applies the threefold decomposition to the structures in figures 1, 2, and 3? Then one obtains structures as in figure 7. Structures of this sort have the chevron orientation. In this case, each face has 1.5 hexagons, plus three-fifths of a pentagon. Across 20 faces, there are 12 pentagons and 30 hexagons. In general, there are 10n$^2$ –10 hexagons and 12 pentagons for a total of 10n$^2$ + 2 polygons. What resolutions yield FHM- and FIA-sized hexagons? The answer is about 280 for an FHM-sized hexagon and about an order 1,458 for an FIA-sized hexagon. Note that an order 1,458 hexagon is one twenty-seventh the size of an order 162 hexagon of the opposite orientation, and that an order 280 hexagon of this (the “chevron”) orientation is 27 times as large as an order 840

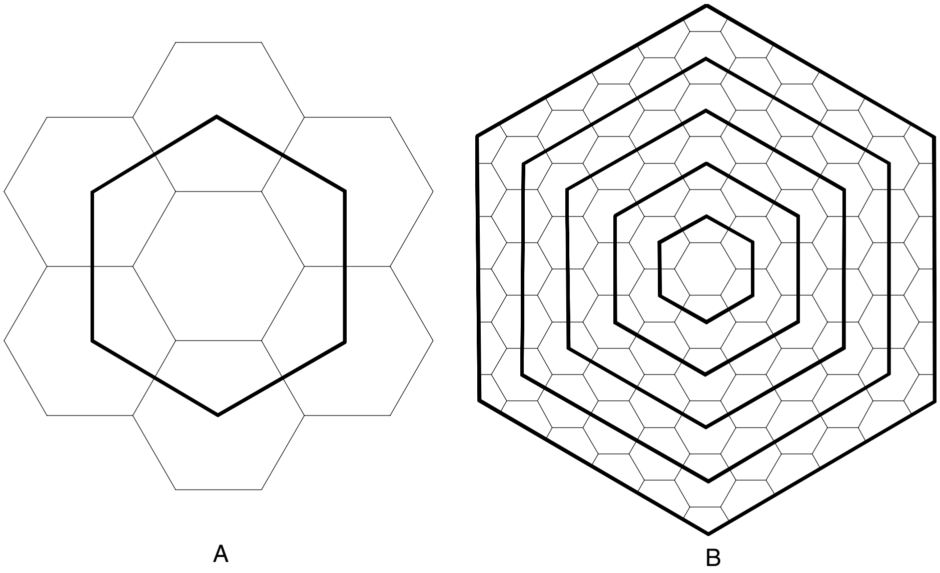


Figure 4—Illustration of (A) the threefold composition method and (B) its generalization.

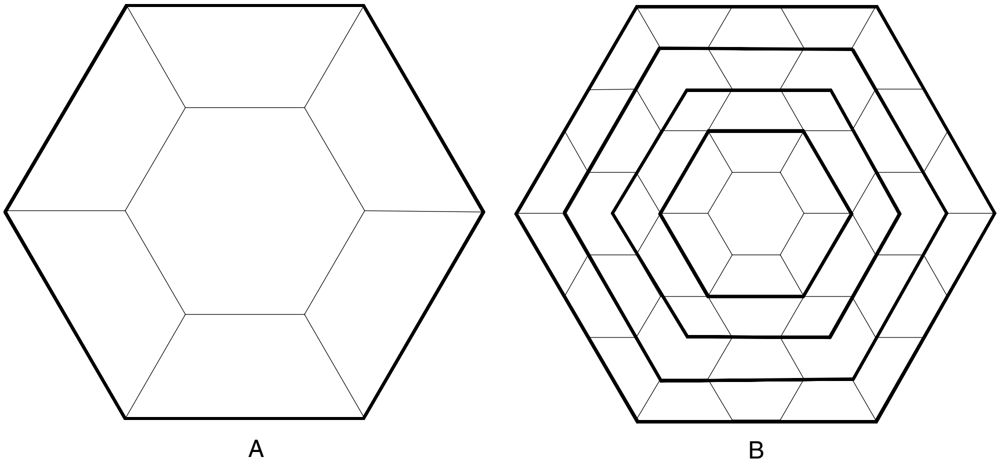


Figure 5—Illustration of (A) the fourfold decomposition method and (B) its generalization

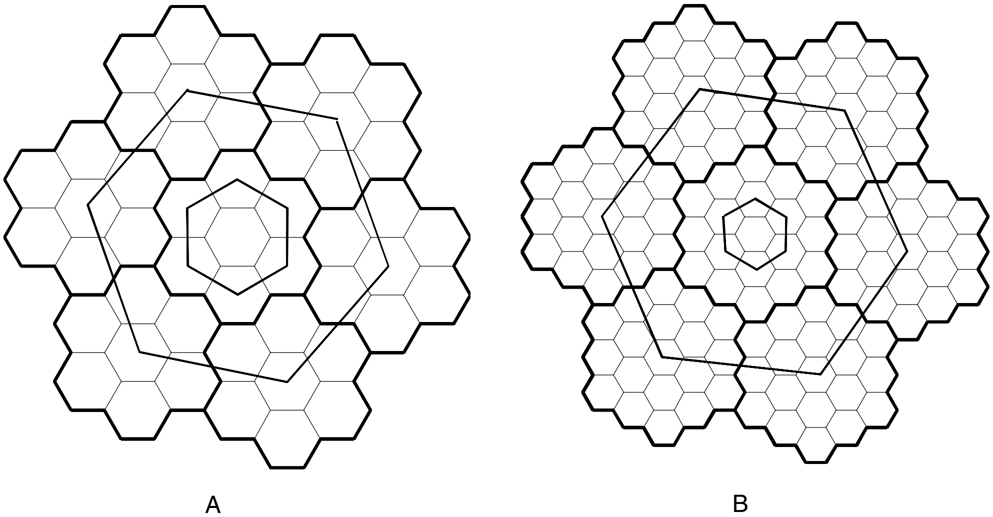


Figure 6—Illustration of (A) the sevenfold and (B) the nineteenfold decomposition.

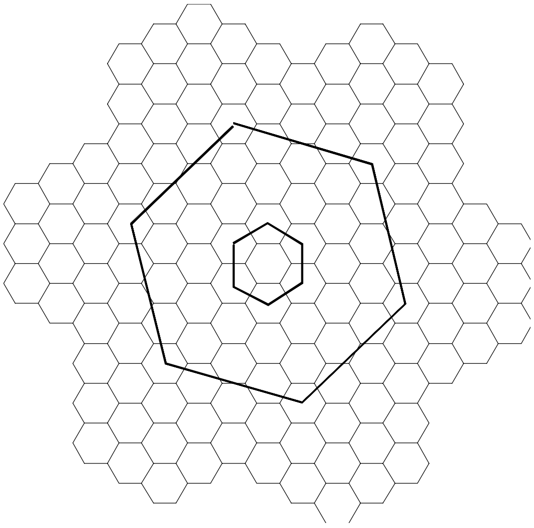


Figure 7—Thirteenfold decomposition.

hexagon of the opposite (the "triangle") orientation. More exact numbers appear in table 2.

Are the triangle and the chevron orientations the only ones possible? No, but mathematically they are the easiest. A close inspection shows that the generalized threefold decomposition is the chevron orientation. In figure 4 (A) cut the figure from the center to the vertices of the bold hexagon. What you see is the chevron orientation. In the equation $T = h^2 + hk + k^2$, substitute h = n and k = n. The result is $T = 3n^2$. The generalized fourfold decomposition is the triangle orientation. In figure 5 (A) cut the figure from the center to the vertices of the bold hexagon. There is the triangle orientation. In the equation $T = h^2 + hk + k^2$, substitute h = n and k = 0. What remains is $T = n^2$, and the generalized sevenfold is an intermediate case. Substitute h = n and k = n - 1 in the equation $T = h^2 + hk + k^2$. What remains is $T = 3n^2 – 3n + 1$, the generalized sevenfold equation.

**Table 2—Number and size of polygons under the chevron orientation at various resolutions**

| Order | Triangles | Hexagons | + | Pentagons | = | Polygons | Area of triangle | Area of hexagon |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | *$m^2$* | *$m^2$* |
| 1 | 60 | 20 | | 12 | | 32 | $8.50 \times 10^{12}$ | $5.10 \times 10^{13}$ |
| 2 | 240 | 30 | | 12 | | 42 | $2.13 \times 10^{12}$ | $1.28 \times 10^{13}$ |
| 3 | 540 | 80 | | 12 | | 92 | $9.45 \times 10^{11}$ | $5.67 \times 10^{12}$ |
| 280 | 4,704,000 | 783,990 | | 12 | | 784,002 | $1.08 \times 10^{8}$ | $6.51 \times 10^{8}$ |
| 281 | 4,737,660 | 789,600 | | 12 | | 789,612 | $1,08 \times 10^{8}$ | $6.46 \times 10^{8}$ |
| 1457 | 127,370,940 | 21,228,480 | | 12 | | 21,228,492 | $4.01 \times 10^{6}$ | $2.40 \times 10^{7}$ |
| 1458 | 127,545,840 | 21,257,630 | | 12 | | 21,257,642 | $4.00 \times 10^{6}$ | $2.40 \times 10^{7}$ |
| 1459 | 127,720,860 | 21,286,800 | | 12 | | 21,286,812 | $4.00 \times 10^{6}$ | $2.40 \times 10^{7}$ |

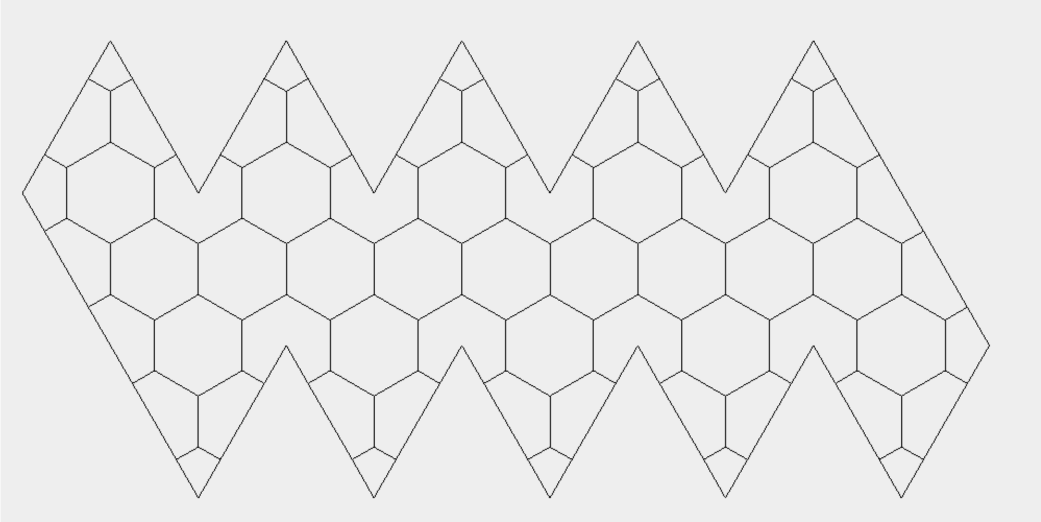


Figure 8—The chevron orientation of hexagon tessellation.

As was mentioned in the introduction, virologists refer to the "triangle" orientation as the structure of a virus with triangulation number T = 3. Triangulation numbers work in the following way:

Start with a plane of tessellated hexagons, as in figure 9. From the origin, go north by h hexagons, marking a spot at H1. Also, go northeast by h hexagons, marking a spot at H2. From H1, go north by k hexagons, marking a spot at K1. From H2, go northwest by k hexagons, marking a spot at K2. In the figure, h = 3 and k = 2. The result is an equilateral triangle with vertices at the origin, K1, and K2.

This equilateral triangle has a triangulation number $T = h^2 + hk + k^2$, or in this case, 19. An icosahedron tessellated with hexagons in this manner would have 10T – 10 hexagons, 12 pentagons, and, of course, 10T + 2 polygons. To tesselate an icosahedron with hexagons, one may continue in this fashion making sure that adjoining triangles match partial hexagons—except at the poles and the ends of the cuts, where pentagons are formed.

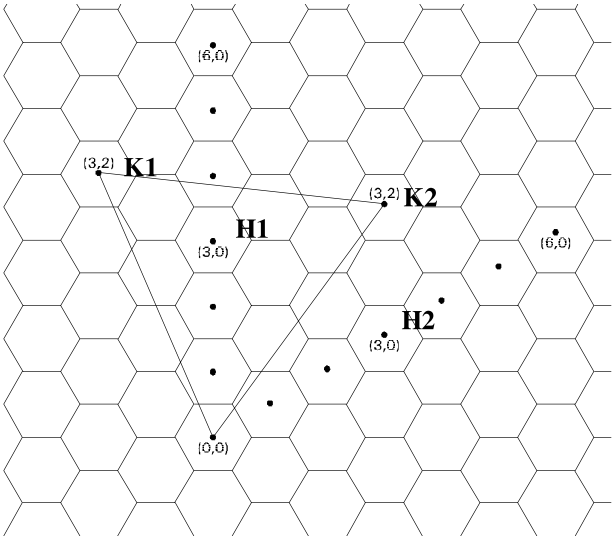


Figure 9—Method of constructing triangulation numbers.

## Icosahedroning the Sphere

When the icosahedron is tessellated with hexagons in the desired way, the next problem is projecting points and lines on the icosahedron to points and lines on the sphere. This problem has baffled cartographers for decades. Among the first to offer a solution was Fisher (1943). His solution is used to this day. He used a combination of aspects of the gnomonic projection. Fisher's map appears in Dahlberg (1997) among other places.

One apparent disadvantage of Fisher's map is that the combination of aspects appears to abruptly shift from straight parallels to curved parallels. The lines of 30° N. and 30° S. look as though they appear on the map twice—once as curved lines and once as straight lines. Thus, directions could be ambiguous and certain points on the earth might appear on the earth in two different places. Snyder used a variation of the Lambert Azimuthal Equal Area projection, and Buckminster Fuller devised a method as well (Pitre 2000).

If you want to map the icosahedron to the sphere, it is desirable that the points on the main triangles match. That way, no point on the earth appears on the map more than once and directions are unambiguous. To make computations easier, you may want to start with the North Pole at the top apex and the South Pole at the bottom apex. Then, split the sphere into five parts with each cut 72° from the next.

One tempting solution is to map parallels on the earth to straight lines in the triangle orientation. Doing so, one can see that 5 triangles (out of 20) join at each of the poles. One can show that one-quarter of a sphere's area is above 30° N., and one-quarter below 30° S. So simply map straight lines in the triangle to parallels on the earth. Meridians sweep out equal areas of the triangle. One flaw in that method is that when half-hexagons are joined from large triangles, the resultant figures look more like pentagons than hexagons.

The next attempt might be to map straight lines on the triangle to great circles on the earth. The cuts are still mapped to meridians (which are great circles). Straight lines are mapped to great circles, which follow such equations as atan(c•sinθ). Munem and Foulis (1984) give the area of a sphere as:

$$\int_{0}^{\pi}\int_{0}^{2\cdot\pi} R^{2}\cdot\sin(f)\,dt\,df \qquad (1)$$

where

R = the radius,
t = longitude, and
f = colatitude.

One may adapt this equation to:

$$\int_{-\pi}^{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} R^{2}\cdot\cos(f)\,df\,dt \qquad (2)$$

where

f = latitude.

If one wants the area bounded by any 72° wedge, the North Pole, and a particular great circle, one gets:

$$\int_{\frac{\pi}{2}-\frac{\pi}{5}}^{\frac{\pi}{2}+\frac{\pi}{5}}\int_{\operatorname{atan}(c\cdot\sin(t))}^{\frac{\pi}{2}} R^{2}\cdot\cos(f)\,df\,dt \qquad (3)$$

Integrate this to get:

$$\frac{2}{5}\cdot\pi\cdot R^{2}-2\cdot\operatorname{asin}\left(\frac{1}{4}\cdot\sqrt{2}\cdot\sqrt{5-\sqrt{5}}\cdot\frac{c}{\sqrt{1+c^{2}}}\right)\cdot R^{2} \qquad (4)$$

If one wants one-twentieth of the total area of the sphere, one sets this result equal to $\pi R^2/5$, and then solves for c, which turns out to be equal to 0.618034. If one substitutes this value into atan(k•sinθ), one sees that the great circle traces a route from 26.565 N. at the endpoints (the 54th and 126th meridians) to an apex of 31.717 N, at the 90th meridian.

If one wants a mathematically simple method, one can just project any point (x,y,z) on the triangular face of the icosahedron to a point (X,Y,Z) on a sphere of radius R in the following way:

$$r = x^2 + y^2 + z^2 \qquad (5)$$

$$X = (R/r)\bullet x \qquad (5a)$$
$$Y = (R/r)\bullet y \qquad (5b)$$
$$Z = (R/r)\bullet z \qquad (5c)$$

This method leads to hexagons of roughly equal size. Exactly equal sized hexagons are desirable, but not crucial to planning a forest inventory (Snyder 1992).

One could adapt the gnomonic method to a more equal area projection by observing that in the triangle orientation, the first row has one triangle, the second three, the third five, . . . and $k^2$ triangles up to and including the kth row. If there are $n^2$ total triangles on the face of a major triangle, one may set expression (4) equal to $(\pi R^2/5)\bullet(k^2/n^2)$ and solve for c. Also, meridians would not be equally spaced on the triangle. Observe that:

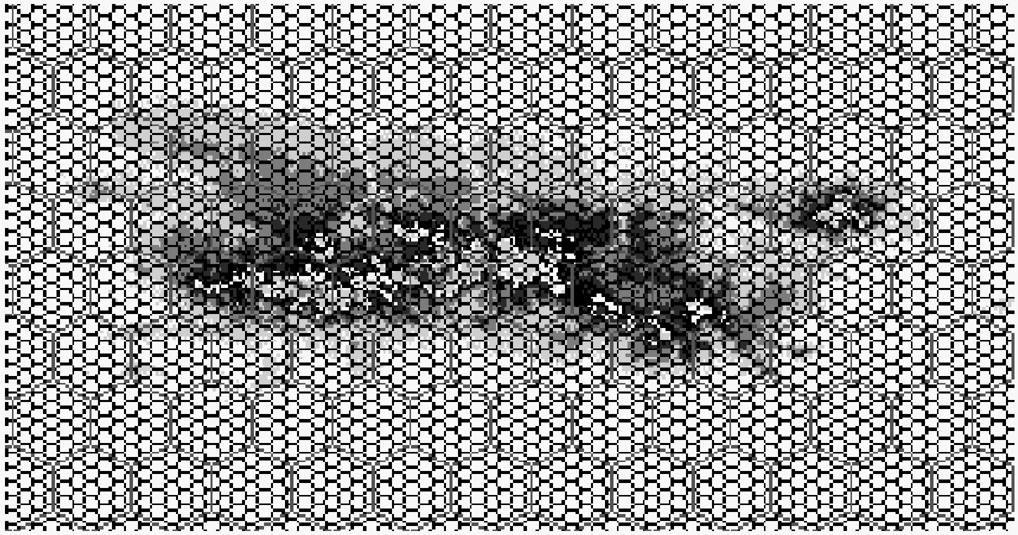


Figure 10—Hexagon grid for Puerto Rico and the U.S. Virgin Islands

$$\int_{\frac{\pi}{2}}^{\frac{\pi}{2}+k} \int_{\operatorname{atan}(c \cdot \sin(t))}^{\frac{\pi}{2}} R^2 \cdot \cos(f)\, df\, dt \qquad (6)$$

is equal to:

$$R^2 \cdot k - \operatorname{asin}\left(\sin(k) \cdot \frac{c}{\sqrt{1+c^2}}\right) \cdot R^2 \qquad (7)$$

which in general is not linear in k. At latitudes near Puerto Rico and the U.S. Virgin Islands, it makes little difference whether one maps straight lines on the icosahedron to parallels on the earth, great circles on the earth, or loxodromes (lines of constant direction) on the earth. However, at latitudes near Alaska, it does make a difference.

In conclusion, figure 10 is a proposed grid system for Puerto Rico and the Virgin Islands. This grid may need to be altered in order to accommodate Snyder's (1992) assumptions and starting points as well as for the particular needs of the FIA and FHM programs in the Caribbean. At this writing, various resolutions of the grid are being explored. In the conterminous United States, FIA-sized hexagons have been grouped into sets of 27 for the purpose of constructing traditional FHM-sized hexagons. One could also overlay another network of sixteenfold hexagons.

## LITERATURE CITED

**Carr. D. [and others].** 1999. ISEA discrete global grids. Statistical Computing and Statistical Graphics Newsletter. 8 (2/3): 31–39.

**Dahlberg, R.** 1997. Interrupting the world map. In: Robinson, Arthur H.; Snyder, John P., eds. Matching the map projection to the need. Bethesda, MD: American Congress on Surveying and Mapping. [Number of pages unknown].

**ESRI.** 1991. Map projections and coordinate management. Redlands, CA: Environmental Systems Research Institute, Inc. 86 p.

**Fisher, I.** 1943. A world map on a regular icosahedron by gnomonic projection. The Geographical Review. 33.4: 605–619.

**Johnson, J.; Speir, J.** 1997. Quasi-equivalent viruses: a paradigm for protein assemblies. Journal of Molecular Biology. 269: 665–675.

**Maggio, E.** 1994. Bouncing balls of carbon: the discovery and promise of fullerenes. In: Bartusiak, Marcia. A positron named Priscilla. Washington DC: National Academy Press. 315 p.

**Munem, M.; Foulis, D.** 1984. Calculus with analytic geometry. New York: Worth Publishers, Inc. 1,004 p.

**Pearson, F**. 1990. Map projections: theory and applications. Boca Raton, FL: CRC Press, Inc. 372 p.

**Pitre, T.** 2000. PC-SIG newsletter. Issue 6.

**Snyder, J.P.** 1992. An equal-area map projection for polyhedral globes. Cartographica. 29(1): 10–21.

**Tegmark, M.** 1996. An icosahedron-based method for pixelizing the celestial sphere. Astrophysics Journal Letters. 470: L81–L84.

# ACCURACY OF REMOTELY SENSED CLASSIFICATIONS FOR STRATIFICATION OF FOREST AND NONFOREST LANDS[1]

**Raymond L. Czaplewski and Paul L. Patterson**[2]

**Abstract**—We specify accuracy standards for remotely sensed classifications used by FIA to stratify landscapes into two categories: forest and nonforest. Accuracy must be highest when forest area approaches 100 percent of the landscape. If forest area is rare in a landscape, then accuracy in the nonforest stratum must be very high, even at the expense of accuracy in the forest stratum. Accuracy in both strata must be at least 90 percent to achieve appreciable gains in efficiency. We recommend that new remotely sensed data be used to re-stratify landscapes whenever the area in forestland decreases by five percent or more since the previous stratification. Efficiency can increase up to 15 percent with formation of an "indeterminate" stratum, which contains elements that are most likely to be misclassified.

## INTRODUCTION

The USDA Forest Service's Forest Inventory and Analysis (FIA) uses a systematic sample of field plots to characterize forest conditions over 300,000,000 hectares of forest and woodland ecosystems in the USA. FIA measures 364,000 1-ha field plots, 120,500 of which are forested. Remote sensing can improve accuracy of FIA statistical estimates. For example, FIA interprets aerial photography for a systematic sample of 9,400,000 plots to improve estimates of forest area and population totals. Stratification with wall-to-wall Landsat satellite data could replace photo-interpretation within the next few years.

We specify accuracy standards for remotely sensed classifications that will be used for stratification into two categories: forest and nonforest. Application of our recommendations requires assumptions, such as the expected prevalence of forestlands in the population. We make generalizations regarding the loss in efficiency caused by change in land use since acquisition of remotely sensed data. These generalizations help determine how old remotely sensed images can become before new imagery is needed for stratification. We also consider formation of a stratum for elements that are most likely to be misclassified. We make recommendations that can help determine a priori the size of this "indeterminate" stratum.

## SAMPLE SURVEY ESTIMATORS

Assume a population is subdivided into two sub-populations, such as forest and nonforest. Our goal is estimation of the proportion P($A_j$) of each sub-population $A_i$ in the population, where 0<P($A_j$)<1. A simple transformation converts this proportion into a percentage or an area (e.g., number of hectares). Assume every element of the sampled population is composed of one and only one sub-population category, which justifies the binomial distribution. We introduce the "error matrix" for remotely sensed classifications and make the connection to statistical stratification, start with the estimator, give the estimators for simple random sampling and for stratification, and define the "design effect" as a measure of the gain in statistical efficiency with stratification.

### Error Matrix for Remote Sensing

The "error matrix" or "confusion matrix" describes accuracy in the remote sensing literature (e.g., Congalton 1991). Let P($B_j$) be the proportion of the population in stratum $B_j$, let P($A_j \cap B_j$) denote the proportion of the population that is jointly in sub-population $A_j$ and remotely sensed stratum $B_j$, and let P($A_j | B_j$) denote the proportion of sub-population $A_j$ given that the remotely sensed stratum is $B_j$, where P($A_j | B_j$)=P($A_j \cap B_j$)/P($B_j$). Figure 1 gives the mathematical notation that we use for the error matrix. We assume remotely sensed classifications are used to define each stratum, and remote sensing measures the size, or area, of each stratum, *i.e.*, P($B_j$).

The ideal stratification occurs when each sub-population occurs in one and only one stratum (Cochran 1977). However, remote sensing does not have 100 percent accuracy, and each remotely sensed stratum usually contains both sub-populations. For example, let stratum $B_1$ be classified as forest with wall-to-wall Landsat data; however, not all sites that are truly forested will be included in this stratum.

The sample of field plots is used to estimate the distribution P($A_j | B_j$) of each sub-population $A_j$ within each remotely sensed stratum $B_j$. Remote sensing improves statistical estimates of each sub-population P($A_j$) by introducing ancillary data, namely precise measurement of the size P($B_j$) for each stratum.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Project Leader and Mathematical Statistician, USDA Forest Service, Rocky Mountain Research Station, 2150 Centre Ave., Bldg. A, Suite 350, Fort Collins, CO 80526–1891, respectively.

1a. Joint probabilities

| | | |
|---|---|---|
| $\mathrm{P}\left(A_1 \cap B_1\right)$ | $\mathrm{P}\left(A_1 \cap B_2\right)$ | $\mathrm{P}\left(A_1\right) = \sum_{j=1}^{m} \mathrm{P}\left(A_2 \cap B_j\right)$ |
| $\mathrm{P}\left(A_2 \cap B_1\right)$ | $\mathrm{P}\left(A_2 \cap B_2\right)$ | $\mathrm{P}\left(A_2\right) = \sum_{j=1}^{m} \mathrm{P}\left(A_2 \cap B_j\right)$ |
| $\mathrm{P}\left(B_1\right) = \sum_{i=1}^{m} \mathrm{P}\left(A_i \cap B_1\right)$ | $\mathrm{P}\left(B_2\right) = \sum_{i=1}^{m} \mathrm{P}\left(A_i \cap B_2\right)$ | $\sum_{i=1}^{m} \sum_{j=1}^{m} \mathrm{P}\left(A_i \cap B_j\right) = 1$ |

1b. Conditional probabilities (accuracy) within strata

| | |
|---|---|
| $\mathrm{P}\left(A_1 \mid B_1\right) = \dfrac{\mathrm{P}\left(A_1 \cap B_1\right)}{B_1}$ | $\mathrm{P}\left(A_1 \mid B_2\right) = \dfrac{\mathrm{P}\left(A_1 \cap B_2\right)}{B_2}$ |
| $\mathrm{P}\left(A_2 \mid B_1\right) = \dfrac{\mathrm{P}\left(A_2 \cap B_1\right)}{B_1}$ | $\mathrm{P}\left(A_2 \mid B_2\right) = \dfrac{\mathrm{P}\left(A_2 \cap B_2\right)}{B_2}$ |
| $\sum_{i=1}^{m} \mathrm{P}\left(A_i \mid B_1\right) = 1$ | $\sum_{i=1}^{m} \mathrm{P}\left(A_i \mid B_2\right) = 1$ |

Figure 1—The "error matrix", or "confusion matrix", describes classification accuracy with remotely sensed data. The goal is estimation of the prevalence or size of each sub-population, i.e., P($A_j$). Post-stratification uses the distribution of sub-population proportions in each stratum, i.e., P($A_j$|$B_j$), and the size of each stratum, i.e., P($B_j$), to improve statistical estimates of P($A_j$) **.**

## Simple Random Sampling

Assuming simple random sampling, Cochran (1977) gives the estimated proportion of sub-population $A_j$ and its variance as:

$$\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{SRS}}\left(A_i\right) &= \frac{1}{n} \sum_{a=1}^{n} x_a \\
V_{\mathrm{SRS}}\left[\hat{\mathrm{P}}\left(A_i\right)\right] &\equiv \left\{\begin{matrix}\text{the variance of the estimate } \hat{\mathrm{P}}\left(A_i\right) \\ \text{from the simple random sample (SRS)}\end{matrix}\right\} \\
&= \frac{1}{n} \mathrm{P}\left(A_i\right)\left[1 - \mathrm{P}\left(A_i\right)\right] \frac{N - n}{N - 1} \\
&\cong \frac{1}{n} \mathrm{P}\left(A_i\right)\left[1 - \mathrm{P}\left(A_i\right)\right]
\end{aligned} \tag{1}$$

where

$$\begin{aligned}
x_a &= \begin{cases} 1 \text{ if sample unit } a \text{ is in sub-population } A_i \\ 0 \text{ otherwise} \end{cases} \\
N &\equiv \left\{\text{the total number of units in the population}\right\} \\
n &\equiv \left\{\text{the sample size, } n \text{ is assumed very small compared to } N\right\}
\end{aligned}$$

## Stratification

Consider a simple random sample of field plots, each of which is classified into one and only one sub-population. Remotely sensed classifications place each field plot into one and only one stratum, and remote sensing measures the area of each stratum. Cochran (1977) gives the estimated proportion of sub-population $A_i$ in the total population, and its variance:

$$
\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{STR}}\left(A_i\right) &= \sum_{j=1}^{m} \mathrm{P}\left(B_j\right)\hat{\mathrm{P}}\left(A_i \mid B_j\right) \\
V_{\mathrm{STR}}\left[\hat{\mathrm{P}}\left(A_i\right)\right] &\equiv \begin{Bmatrix}\text{the variance of the estimate } \hat{\mathrm{P}}\left(A_i\right) \\ \text{from the stratified random sample (STR)}\end{Bmatrix} \\
&= \sum_{j=1}^{m} \frac{\mathrm{P}\left(B_j\right)^2}{n_j} \mathrm{P}\left(A_i \mid B_j\right)\left[1 - \mathrm{P}\left(A_i \mid B_j\right)\right]\frac{N_j - n_j}{N_j - 1} \\
&\cong \frac{1}{n}\sum_{j=1}^{m} \mathrm{P}\left(B_j\right)\mathrm{P}\left(A_i \mid B_j\right)\left[1 - \mathrm{P}\left(A_i \mid B_j\right)\right]
\end{aligned}
\tag{2}
$$

where

$$
\begin{aligned}
\mathrm{P}\left(B_j\right) &= N_j/N \\
N_j &\equiv \left\{\text{the number of units in stratum } j\right\} \\
\mathrm{P}\left(A_i \mid B_j\right) &\equiv \begin{Bmatrix}\text{the proportion of sub-population } A_i \\ \text{in stratum } B_j\end{Bmatrix} \\
\hat{\mathrm{P}}\left(A_i \mid B_j\right) &= \frac{1}{n_j}\sum_{a \in n \cap a \in B_j} x_a \cong \frac{1}{\mathrm{P}\left(B_j\right)n}\sum_{a \in n \cap a \in B_j} x_a \\
n_j &\equiv \begin{Bmatrix}\text{sample size in stratum } j, \\ n_j \text{ much smaller than } N_j\end{Bmatrix} \cong \mathrm{P}\left(B_j\right)n
\end{aligned}
$$

Equation (2) is sufficient to compute approximations for the variance of a post-stratified estimate. However, the expected variance is often needed for survey planning when the stratum sizes $\mathrm{P}(B_j)$ and conditional probabilities $\mathrm{P}(A_i|B_j)$ are not yet observed. In the following sections, we make realistic assumptions and simplifications that make it easier to anticipate gains from stratification and specify accuracy standards for remote sensing.

## Design Effect

The improvement in statistical efficiency with stratification of a simple random or systematic sample is quantified by the ratio of variances, which is designated the "design effect" by Särndal and others (1992). We denote the design effect as *k*, and it is approximated with equations (1) and (2) as:

$$
k = \frac{V_{\mathrm{STR}}\left[\mathrm{P}\left(A_i\right)\right]}{V_{\mathrm{SRS}}\left[\mathrm{P}\left(A_i\right)\right]} \cong \frac{\sum_{j=1}^{m} \mathrm{P}\left(B_j\right)\mathrm{P}\left(A_i \mid B_j\right)\left[1 - \mathrm{P}\left(A_i \mid B_j\right)\right]}{\mathrm{P}\left(A_i\right)\left[1 - \mathrm{P}\left(A_i\right)\right]}
\tag{3}
$$

If stratification improves the estimate, then *k* must be less than 1. Since all variances are positive, *k*>0. In the following sections, we use the design effect to simplify the mathematics and draw broad generalizations.

For two strata (*m* = 2), the design effect *k* in equation (3) simplifies to:

$$
\begin{aligned}
k &= \frac{V_{str}\left[\mathrm{P}\left(A_1\right)\right]}{V_{srs}\left[\mathrm{P}\left(A_1\right)\right]} \\
&= \frac{\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\left[1 - \mathrm{P}\left(A_1 \mid B_1\right)\right] + \mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\left[1 - \mathrm{P}\left(A_1 \mid B_2\right)\right]}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} \\
&= \frac{\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\left[\mathrm{P}\left(A_2 \mid B_1\right)\right] + \mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\left[\mathrm{P}\left(A_2 \mid B_2\right)\right]}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)}
\end{aligned}
\tag{4}
$$

Now, we express equation (4) in a form that is more useful in deriving the generalizations that follow. First, we use figure 1 to define two useful equalities:

$$
\begin{aligned}
\mathrm{P}\left(A_1\right) &= \begin{cases} \left[\mathrm{P}\left(A_1 \cap B_1\right)\right] + \left[\mathrm{P}\left(A_1 \cap B_2\right)\right] \\ \left[\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\right] + \left[\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\right] \end{cases} \\
\left[\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\right] &= \mathrm{P}\left(A_1\right) - \left[\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\right] \\
\left[\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_2 \mid B_1\right)\right] &= \mathrm{P}\left(A_2\right) - \left[\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_2 \mid B_2\right)\right]
\end{aligned}
\tag{5}
$$

Equation (4) can be expressed using the equalities in equation (5) as:

$$
\begin{aligned}
k\left[\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)\right] &= \left[\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_2 \mid B_1\right)\right]\mathrm{P}\left(A_1 \mid B_1\right) + \left[\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\right]\mathrm{P}\left(A_2 \mid B_2\right) \\
&= \left\{\begin{matrix} \left[\mathrm{P}\left(A_2\right) - \mathrm{P}\left(B_2\right)\mathrm{P}\left(A_2 \mid B_2\right)\right]\mathrm{P}\left(A_1 \mid B_1\right) + \\ \left[\mathrm{P}\left(A_1\right) - \mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\right]\mathrm{P}\left(A_2 \mid B_2\right) \end{matrix}\right\} \\
&= \left\{\begin{matrix} \mathrm{P}\left(A_2\right)\mathrm{P}\left(A_1 \mid B_1\right) - \mathrm{P}\left(B_2\right)\mathrm{P}\left(A_2 \mid B_2\right)\mathrm{P}\left(A_1 \mid B_1\right) + \\ \mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2 \mid B_2\right) - \mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right) \end{matrix}\right\} \\
&= \left\{\begin{matrix} \mathrm{P}\left(A_2\right)\mathrm{P}\left(A_1 \mid B_1\right) + \mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2 \mid B_2\right) \\ - \mathrm{P}\left(B_2\right)\mathrm{P}\left(A_2 \mid B_2\right)\mathrm{P}\left(A_1 \mid B_1\right) - \mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right) \end{matrix}\right\} \\
&= \left\{\begin{matrix} \mathrm{P}\left(A_2\right)\mathrm{P}\left(A_1 \mid B_1\right) + \mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2 \mid B_2\right) \\ - \left[\mathrm{P}\left(B_2\right) + \mathrm{P}\left(B_1\right)\right]\mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right) \end{matrix}\right\}
\end{aligned}
\tag{6}
$$

Since P($B_1$)+ P($B_2$)=1, equation (6) may be rewritten as:

$$
\begin{aligned}
k &= \frac{\mathrm{P}\left(A_2\right)\mathrm{P}\left(A_1 \mid B_1\right) + \mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2 \mid B_2\right) - \mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} \\
&= \frac{\mathrm{P}\left(A_2\right)\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} + \frac{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} - \frac{\mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} \\
&= \frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)} + \frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)} - \frac{\mathrm{P}\left(A_1 \mid B_1\right)\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)}
\end{aligned}
\tag{7}
$$

Subtract *k* in equation (7) from 1:

$$
\begin{aligned}
1 - k &= 1 - \left\{\left[\frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right] + \left[\frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\right] - \left[\frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\right]\left[\frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right]\right\} \\
&= 1 - \left[\frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right] - \left\{\left[\frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\right] - \left[\frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\right]\left[\frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right]\right\} \\
&= \left[1 - \frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right] - \frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\left[1 - \frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right] \\
&= \left[1 - \frac{\mathrm{P}\left(A_1 \mid B_1\right)}{\mathrm{P}\left(A_1\right)}\right]\left[1 - \frac{\mathrm{P}\left(A_2 \mid B_2\right)}{\mathrm{P}\left(A_2\right)}\right]
\end{aligned}
\tag{8}
$$

Rearranging equation (8):

$$\begin{aligned}
1-k &= \left[\frac{\mathrm{P}(A_1)}{\mathrm{P}(A_1)} - \frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2)}{\mathrm{P}(A_2)} - \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \\
&= \left[\frac{\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \\
&= \frac{\left[\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)\right]\left[\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)\right]}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} \\
&= \left[-\frac{\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_2)}\right]\left[-\frac{\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)}\right] \\
&= \left[\frac{\mathrm{P}(A_1 \mid B_1) - \mathrm{P}(A_1)}{1 - \mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2 \mid B_2) - \mathrm{P}(A_2)}{1 - \mathrm{P}(A_2)}\right]
\end{aligned} \tag{9}$$

Note that the final expression in equation (9) is the product of two terms, each of which is independent of each other and the sizes of the strata P($B_i$). This feature greatly simplifies subsequent algebra.

## SYMMETRICAL 2x2 ERROR MATRIX

We now derive expressions for the size of the two strata given the sizes for both sub-populations P($A_i$). We assume the relative accuracy is identical in both strata (see below). Under this assumption, we show that both margins of the 2x2 error matrix in figure 1a are identical, i.e., P($A_i$)= P($B_i$), and the off-diagonal joint probabilities are identical, i.e., P($A_1 \cap B_2$)=P($A_2 \cap B_1$). The symmetry under these conditions facilitates derivations in other sections.

### Relative Accuracy

Using equation (9), define “relative accuracy” as follows:

$$\frac{\mathrm{P}(A_i \mid B_i) - \mathrm{P}(A_i)}{1 - \mathrm{P}(A_i)} \equiv \text{relative accuracy in stratum } i \tag{10}$$

If the relative accuracies are identical in both strata, then the following proceed from equation (9):

$$\begin{aligned}
\frac{\mathrm{P}(A_i \mid B_i) - \mathrm{P}(A_i)}{1 - \mathrm{P}(A_i)} &= \sqrt{1-k} \\
\mathrm{P}(A_i \mid B_i) &= \sqrt{1-k}\left[1 - \mathrm{P}(A_i)\right] + \mathrm{P}(A_i)
\end{aligned} \tag{11}$$

### Symmetric Margins

From figure 1b and equation (11), the off-diagonal conditional probabilities are:

$$\begin{aligned}
\mathrm{P}(A_2 \mid B_1) &= 1 - \mathrm{P}(A_1 \mid B_1) \\
&= 1 - \left\{\sqrt{1-k}\left[1 - \mathrm{P}(A_1)\right] + \mathrm{P}(A_1)\right\} \\
&= \left[1 - \mathrm{P}(A_1)\right] - \sqrt{1-k}\left[1 - \mathrm{P}(A_1)\right] \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_2) \\
\mathrm{P}(A_1 \mid B_2) &= 1 - \mathrm{P}(A_2 \mid B_2) \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)
\end{aligned} \tag{12}$$

By definition, the size of sub-population $A_1$ equals the sum of sub-population $A_1$ in each of the two strata $B_1$ and $B_2$. Using figure 1a and equations (11) and (12):

$$
\begin{aligned}
\mathrm{P}(A_1) &= \mathrm{P}(A_1 \cap B_1) + \mathrm{P}(A_1 \cap B_2) \\
&= \mathrm{P}(B_1)\left\{\mathrm{P}(A_1 \mid B_1)\right\} + \mathrm{P}(B_2)\left\{\mathrm{P}(A_1 \mid B_2)\right\} \\
&= \mathrm{P}(B_1)\left\{\sqrt{1-k}\,\mathrm{P}(A_2) + \mathrm{P}(A_1)\right\} + \left[1 - \mathrm{P}(B_1)\right]\left\{\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\right\} \\
&= \left\{\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \\ \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} + \left\{\begin{matrix}\left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] - \\ \left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right]\mathrm{P}(B_1)\end{matrix}\right\} \\
&= \left\{\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \\ \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} + \left\{\begin{matrix}\left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] + \\ -\,\mathrm{P}(A_1)\,\mathrm{P}(B_1) + \sqrt{1-k}\,\mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} \\
&= \left[\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \sqrt{1-k}\,\mathrm{P}(A_1)\,\mathrm{P}(B_1) \\ +\,\mathrm{P}(A_1)\,\mathrm{P}(B_1) - \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right] + \left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] \\
&\left.\begin{aligned}
&= \sqrt{1-k}\left[\mathrm{P}(A_2) + \mathrm{P}(A_1)\right]\mathrm{P}(B_1) + \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \sqrt{1-k}\,\mathrm{P}(B_1) + \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)
\end{aligned}\right\} \text{ since } \left[\begin{matrix}\mathrm{P}(A_2) + \\ \mathrm{P}(A_1)\end{matrix}\right] = 1
\end{aligned}
\tag{13}
$$

Solving equation (13) for P($B_1$), we show that both margins of the 2x2 error matrix in figure 1a are identical, i.e., P($A_i$)= P($B_i$):

$$
\begin{aligned}
\sqrt{1-k}\,\mathrm{P}(B_1) &= \mathrm{P}(A_1) - \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \left[1 - 1 + \sqrt{1-k}\right]\mathrm{P}(A_1) \\
\mathrm{P}(B_1) &= \mathrm{P}(A_1) \\
1 - \mathrm{P}(B_1) &= 1 - \mathrm{P}(A_1) \\
\mathrm{P}(B_2) &= \mathrm{P}(A_2)
\end{aligned}
\tag{14}
$$

### Symmetric Off-Diagonals

Finally, we show that both off-diagonal joint probabilities in the 2x2 error matrix (fig. 1a) are identical when the relative accuracies in both strata are identical:

$$
\begin{aligned}
&\left.\begin{aligned}
\mathrm{P}(A_1 \cap B_2) &= (A_2)\,\mathrm{P}(A_1 \mid B_2) \\
&= (A_2)\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\,\mathrm{P}(A_2)
\end{aligned}\right\} \text{for } \mathrm{P}(B_2) = \mathrm{P}(A_2) \\
&\left.\begin{aligned}
\mathrm{P}(A_2 \cap B_1) &= (A_1)\,\mathrm{P}(A_2 \mid B_1) \\
&= (A_1)\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_2) \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\,\mathrm{P}(A_2)
\end{aligned}\right\} \text{for } \mathrm{P}(B_1) = \mathrm{P}(A_1) \\
&\mathrm{P}(A_1 \cap B_2) = \mathrm{P}(A_2 \cap B_1)
\end{aligned}
\tag{15}
$$

### Symmetric Matrix of Joint Probabilities

Assuming the relative accuracies are identical in both strata, as in equations (11) to (15), the matrix of joint probabilities from figure 1a is symmetric, as given in figure 2.

## ACCURACY STANDARDS

Classification accuracy P($A_i$|$B_i$) in stratum $B_i$ must be greater than the proportion of sub-population P($A_i$) in the population, i.e., P($A_i$|$B_i$)>P($A_i$); otherwise, the design effect $k$ will be greater than one in equation (9). For example, a 10,000-km$^2$ geographic area truly contains 7,000-km$^2$ of forest cover. Stratification will improve precision if, and only if, the remotely sensed forest category has at least 70 percent accuracy [*i.e.*, 0.7<P($A_1$|$B_1$)<1.0], and the remotely sensed nonforest category has at least 30 percent accuracy [*i.e.*, 0.3<P($A_2$|$B_2$)<1.0]. However, accuracy must be far greater before the gain in precision is substantial, as we now discuss.

Examples of desired accuracy of remotely sensed classifications are given for five different levels of design effect $k$ in table 1. For example, a "substantial" gain in table 1 is defined as a design effect of $k$=0.5, meaning:

| | | |
|---|---|---|
| $P(A_1 \cap B_1) = P(A_1)\begin{bmatrix} P(A_1) + \\ P(A_2)\sqrt{1-k} \end{bmatrix}$ | $P(A_1 \cap B_2) = \left[1 - \sqrt{1-k}\right] P(A_1) P(A_2)$<br>$= P(A_2 \cap B_1)$ | $P(A_1)$ |
| $P(A_2 \cap B_1) = \left[1 - \sqrt{1-k}\right] P(A_1) P(A_2)$<br>$= P(A_1 \cap B_2)$ | $P(A_2 \cap B_2) = P(A_2)\begin{bmatrix} P(A_2) + \\ P(A_1)\sqrt{1-k} \end{bmatrix}$ | $P(A_2)$ |
| $P(B_1) = P(A_1)$ | $P(B_2) = P(A_2)$ | 1 |

Figure 2—Error matrix for two sub-populations and corresponding strata when the relative accuracy in each stratum is identical (equations (11) to (15)). These assumptions permit generalizations, while maintaining realistic scenarios.

- The estimate with stratification has half the variance of the estimate with simple random sampling;
- Estimates from simple random sampling would require a two-fold increase in the number of field plots to achieve the same variance with stratification; and
- The confidence interval with stratification is approximately 71 percent ($\sqrt{0.5} \times 100\%$) smaller than that with simple random sampling.

Figure 3 illustrates the relative precision of stratified estimates for each level of design effect in table 1. The discrete levels in table 1 simplify mathematical generalizations that follow.

Assume that the relative accuracies are identical in both strata, as in equation (11). Figure 4 displays the classification accuracy $P(A_i|B_i)$ in stratum $B_i$ required to meet various levels of gain in statistical efficiency (*k* in table 1). From figure 4, classification accuracy in a stratum must be nearly perfect if its corresponding sub-population is very prevalent, *i.e.*, $P(A_i)$»1.0, while the accuracy need not be nearly as great for a rare sub-population, *i.e.*, $P(A_i)\approx 0$. Czaplewski and Patterson (in preparation) show that figure 4 is applicable classification systems having three or more sub-populations under certain assumptions.

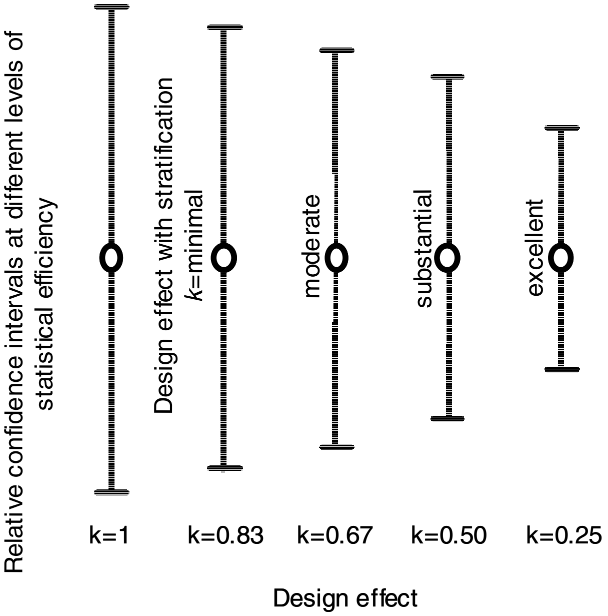


Figure 3—Relative confidence intervals for different levels of design effect *k* in table 1.

**Table 1—Levels of gain from stratification used in comparisons**

| Design effect $k$=$V_{STR}/V_{SRS}$ | Gain in efficiency through post-stratification | Increase in effective number of plots[a] gained through stratification | Relative variance of stratified sampling compared to simple random sampling 100×$V_{STR}/V_{SRS}$ | Relative standard error[b] of stratified sampling compared to simple random sampling 100× $\sqrt{V_{STR}/V_{SRS}}$ |
|---|---|---|---|---|
| | | | *Percent* | |
| $k$=(1/1.0)=1.00 | "None" | None | 100 | 100 |
| $k$=(1/1.2)=0.83 | "Minimal" | 1.2-fold | 83 | 91 |
| $k$=(1/1.5)=0.67 | "Moderate" | 1.5-fold | 67 | 82 |
| $k$=(1/2.0)=0.50 | "Substantial" | 2-fold | 50 | 71 |
| $k$=(1/4.0)=0.25 | "Excellent" | 4-fold | 25 | 50 |

[a] The increase in sample size *n* that would be required to achieve the same variance without stratification.
[b] Approximately proportional to the confidence interval.

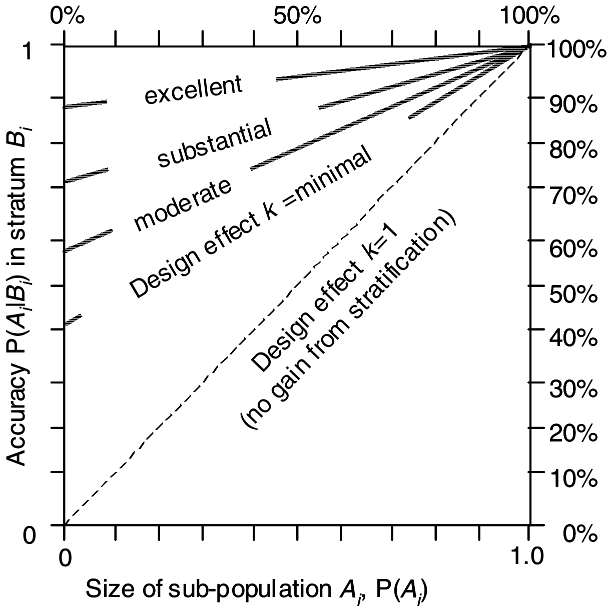


Figure 4—Classification accuracy $P(A_i|B_i)$ in stratum $B_i$ required for different levels of statistical gain ($k$ in table 1) as a function of sub-population size $P(A_i)$. These results assume that the relative accuracy is identical for all strata as in equation (11).

## LOSS OF EFFICIENCY OVER TIME

Landscapes change over time through land management, succession, disturbance and shifting land use. Current field plots may be stratified using remotely sensed data that were acquired many years ago. Some portion of the total "classification error" is caused by changes in the landscape, not by the original accuracy of the remotely sensed classifications. How old can the remotely sensed data become before its value for stratification becomes seriously degraded?

Assume the size of sub-population $A_1$ at time $t$ decreases by some fraction D of its original size at time 0, where 0<D<1. Since the remotely sensed data were acquired at time $t$=0, the stratum sizes $P(B_i)$ are the same at times 0 and $t$. Assume changes in the landscape between times 0 and $t$ are independent of the remotely sensed classification at time 0. The decrease in the size of sub-population $A_1$ causes a corresponding increase in sub-population $A_2$. Finally, assume both strata at time 0 have the same relative accuracy. The error matrix in figure 5, which corresponds to figure 2, captures these assumptions. Under these conditions, the design effect $k_t$ at time $t$ equals:

$$k_t = 1 - \left\{ \frac{(1 - \Delta)(1 - k_0)\left[1 - P(A_1)_0\right]}{1 - P(A_1)_0 + \Delta P(A_1)_0} \right\} \tag{16}$$

Equation (16) can be transformed into a more general expression that simultaneously covers all levels of the design effect $k_0$:

$$\frac{k_t - k_0}{1 - k_0} = \frac{\Delta}{1 - (1 - \Delta)\Delta P(A_1)_0} \tag{17}$$

Figure 6 is a graphical display of equation (17). When there has been little change in the landscape between time 0 and $t$ ($\Delta \approx 0$), there is little change in design effect ($k_t \approx k_0$) using remotely sensed data acquired at time 0, and there is little loss in statistical efficiency. However, as the net decrease in sub-population size ($\Delta$) becomes larger, the design effect approaches one. This means that the variance with stratification is nearly equal to that under simple random sampling, i.e., the gain in efficiency through stratification is almost entirely lost.

When a sub-population is very common, i.e., $P(A_i) \approx 1$, even a small decrease in sub-population size between time 0 and $t$ causes major losses in efficiency. However, if the

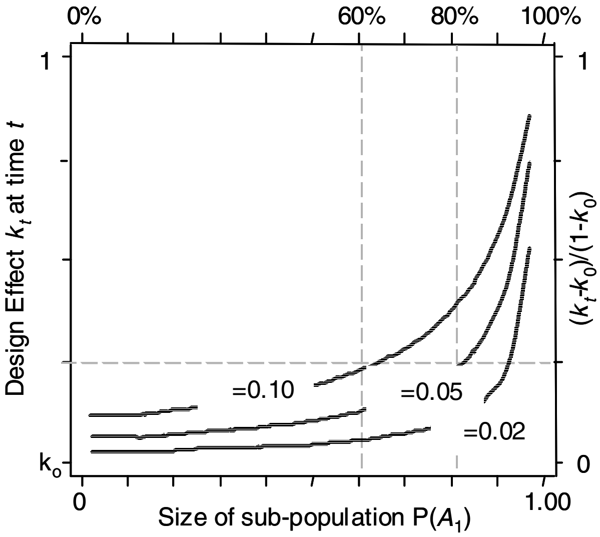


Figure 6—Design effect as function of loss rate for one of the sub-populations. As the rate becomes faster, the design effect approaches 1, meaning that the variance with stratification is no better than the variance with simple random sampling.

| | | |
|---|---|---|
| $P(A_1 \cap B_1)_t$<br>$= (1 - \Delta) P(A_1 \cap B_1)_0$ | $P(A_1 \cap B_2)_t$<br>$= (1 - \Delta) P(A_1 \cap B_2)_0$ | $P(A_1)_t = (1 - \Delta) P(A_1)_0$ |
| $P(A_2 \cap B_1)_t = \begin{bmatrix} P(A_2 \cap B_1)_0 + \\ \Delta P(A_1 \cap B_1)_0 \end{bmatrix}$ | $P(A_2 \cap B_2)_t = \begin{bmatrix} P(A_2 \cap B_2)_0 + \\ \Delta P(A_1 \cap B_2)_0 \end{bmatrix}$ | $P(A_2)_t = \begin{bmatrix} P(A_2)_0 + \\ \Delta P(A_1)_0 \end{bmatrix}$ |
| $P(B_1)$ | $P(B_2)$ | 1 |

Figure 5—Error matrix that includes change (D at time $t$) in sub-population $A_1$ after acquisition of the remotely sensed data (time $t$=0) that are used to specify strata $B_1$ and $B_2$.

sub-population represents less than 80 percent of the total population at time 0, and the net decrease at time *t* is less than 5 percent ($\Delta$=0.05), then the decrease in efficiency is less than 25 percent (fig. 6). Most extensive landscapes have less than 80 percent forest.

As example, let sub-population $A_1$ represent forest land-use, and the stratification is based on ten-year old remotely sensed data. Assume 1 percent per year of the original forest land-use is converted to another land use, such as agriculture or urban. The rate of change over 10 years is $\Delta\approx0.10$. Assume there are no conversions back to forest land-use. If forest land-uses occupy less than 60 percent of the landscape, then the stratification based on ten-year old remotely sensed data retains 75 percent of its efficiency (fig. 6). Loss of statistical efficiency is most rapid in those landscapes dominated by forest land-uses.

Czaplewski and Patterson (in preparation) extend this model to cases in which changes occur in both sub-populations. They analyze steady state conditions, in which the changes in sub-population $A_1$ exactly equal the changes in sub-population $A_2$. They find that statistical efficiency also decreases over time, and the rate of loss in efficiency can be higher in a dynamic stead-state landscape than a landscape that is not at equilibrium.

## INDETERMINATE STRATUM

Some population elements (e.g., pixels) are classified with less confidence than other elements with remotely sensed data. For example, the maximum likelihood classifier, which is widely used for image processing, computes the probabilities of a pixel being a member of each remotely sensed category. The pixel is assigned to the category with the highest probability, even if the largest probability is relatively low for some pixels. This often occurs with mixed pixels, or pixels near the boundary of a multivariate cluster. As another example, a binary-tree classifier assigns each element into a single category, but the algorithm estimates the probability of correct classification using its training data. Even with unsupervised classifiers, all multivariate clusters do not have the same proportion of predominate labeling sites. We investigate the opportunity to increase statistical efficiency by creating a new stratum that contains pixels which are classified with less confidence than other pixels. We label this stratum as the “indeterminate stratum.”

The matrix in figure 7 gives one example that is numerically tractable. Let $d_{ij}$ represent the quantity of elements that are removed from sub-population *i* in stratum *j*. We start by moving a small quantity of elements into the indeterminate stratum and increase the quantity until the gain in statistical efficiency is maximized. The first elements removed are those that are most difficult to successfully classify.

The size of each $d_{ij}$ is modeled by functions $f_{i,j}(c_i)$, where $c_i$ starts at 0 and incrementally increases towards 1 until the optimum is realized (fig. 8). These functions have the following conditions:

1. For the matrix in figure 7, $d_{ij}=f_{i,j}(c_i)$ for $0<c_i<1$ in stratum $B_i$. If the sub-population is correctly classified in stratum $B_i$, $f_{i,i}(c_i)$ is a linear function of $(c_i)$. If the sub-population is not correctly classified in stratum $B_i$, $f_{j,i}(c_i)$ is a non-linear function of $(c_i)$ so that we can impose Conditions 3 and 4 that follow. Figure 8 illustrates these two functions.

$$\begin{aligned} d_{i,i} &= f_{i,i}(c_i) \\ &= c_i \, \mathrm{P}(A_i \cap B_i) \\ d_{j,i} &= f_{j,i}(c_i) \\ &= \frac{\mathrm{P}(A_i \cap B_i)}{\alpha}\left(1 - e^{-\alpha c_i}\right) \end{aligned} \tag{18}$$

2. The number of elements correctly classified in a stratum must always be larger than the number of elements that are incorrectly classified, i.e., $\mathrm{P}(A_i|B_i)>0.5$. For large values of the design effect *k* (meaning that the original stratification yields little gain in efficiency over simple random sampling), and a when stratum is very rare, this condition is not always met. Therefore, the following constraint is placed on $\mathrm{P}(A_i)$ in equation (18).

$$\mathrm{P}(A_1) > \frac{1 - 2\sqrt{1-k}}{2\left(1 - \sqrt{1-k}\right)}$$

3. Some classification errors are not removed until virtually all elements in the stratum are shifted into the indeterminate stratum (fig. 8). The last elements to be removed from the stratum ($c\approx1$) have almost no classification error. The value of a in equation (18) is numerically determined so that the following condition is true for $f_{j,i}(c_i)$, i.e., when the sub-population is not correctly classified in stratum $B_i$.

$$f_{j,i}(c_i = 1) = \mathrm{P}(A_j \cap B_i)$$

4. The highest proportion of classification errors are removed from each stratum during the first incremental

| | | | |
|---|---|---|---|
| $\mathrm{P}(A_1 \cap B_1) - d_{1,1}$ | $d_{1,1} + d_{1,2}$ | $\mathrm{P}(A_1 \cap B_2) - d_{1,2}$ | $\mathrm{P}(A_1)$ |
| $\mathrm{P}(A_2 \cap B_1) - d_{2,1}$ | $d_{2,1} + d_{2,2}$ | $\mathrm{P}(A_1 \cap B_2) - d_{2,2}$ | $\mathrm{P}(A_2)$ |
| $\mathrm{P}(A_1) - (d_{1,1} + d_{2,1})$ | $d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}$ | $\mathrm{P}(A_2) - (d_{1,2} + d_{2,2})$ | 1 |

Figure 7—Error matrix in which an “indeterminate” stratum, which includes sites that are most likely to be misclassified. This stratum can increase statistical efficiency by increasing the classification accuracy in the original two strata.

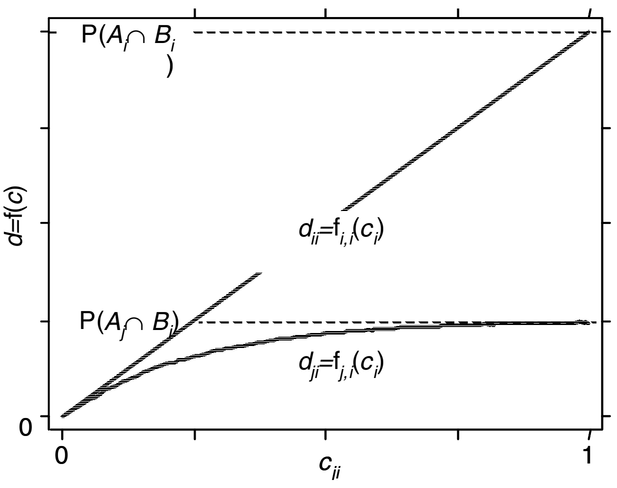


Figure 8—Functions $f_{i,j}(c_i)$ for the size of each $d_{ij}$, which is used to shift likely classification errors into the "indeterminate" stratum. See figure 7 and equation (18).

shifts of elements into the indeterminate stratum, where $c_i \approx 0$. We assume a 50:50 mixture of the two subpopulations among these first elements. This represents those elements that are most difficult to correctly classify. This constraint is imposed by making the first derivatives of $f_{i,i}(c_i)$ and $f_{j,i}(c_i)$ identical when $c_i=0$ (fig. 8).

$$\left[\frac{\partial}{\partial c_i} f_{i,i}\left(c_i = 0\right)\right] = \left[\frac{\partial}{\partial c_i} f_{j,i}\left(c_i = 0\right)\right]$$

5. As elements are moved into the indeterminate stratum, accuracies increase in both of the original strata. We force their relative accuracies to remain equal so that $c_2$ can be expressed as a function of $c_1$. This reduces the number of variables in our evaluation. However, we make an exception to this constraint when $c_1=1$, meaning all of stratum $B_1$ is moved into the indeterminate stratum.

$$\left[\frac{\dfrac{P\left(A_1 \cap B_1\right) - d_{1,1}}{P\left(A_1\right) - d_{1,1} - d_{2,1}} - P\left(A_1\right)}{1 - P\left(A_1\right)}\right] = \left[\frac{\dfrac{P\left(A_2 \cap B_2\right) - d_{2,2}}{P\left(A_2\right) - d_{1,2} - d_{2,2}} - P\left(A_2\right)}{1 - P\left(A_2\right)}\right]$$

$$c_2 = f\left(c_1\right) \quad \text{for } c_1 < 1$$

6. The shift of elements into the indeterminate stratum stops when the design effect in equation (4) reaches its minimum within the interval $0<c_i<1$, i.e., the optimum improvement in statistical efficiency.

$$\left[\frac{\partial}{\partial c_1} k\right] = 0$$

7. We relax Condition 5 (above) when $c_1=1$, meaning the optimum in Condition 6 is not realized as $c_1$ reaches 1. This situation approximately occurs whenever $P(A_i)<1.15k-0.67$ in our model from figure 7 and equation (18). In this situation, we merge stratum $B_1$ with the indeterminate stratum. This returns us to two strata, where the merged stratum contains stratum $B_1$ plus elements removed from stratum $B_2$. We use $c_2$ to increase the quantity of elements shifted from stratum $B_2$ into this new stratum until we achieve the optimum in Condition 6.

We were unable to find an algebraic solution to this formulation; therefore, we developed a numerical solution. The following describe our results.

We found that statistical efficiency does increase with addition of an indeterminate stratum, at least using the model in figure 7 and equation (18). Let $k_{opt}$ represent the design optimal effect with addition of the indeterminate stratum. Figure 9 shows the proportional improvement in the design effect ($k_{opt}/k$) relative to the initial prevalence of stratum $B_1$ for different initial design effects (table 1). The optimal gain in efficiency exceeds 15 percent ($k_{opt}/k<1-0.15$) when classification accuracy is high, *i.e.*, the initial design effect is excellent; however, the gain is less than 5 percent when the initial design effect is marginal.

Given the model in figure 7 and equation (18), optimal size of the indeterminate stratum varies with prevalence of the two strata, as shown in figure 10. The optimal size is under 10 percent of the population when the design effect is excellent (classification accuracy is high), but it can approach 30 percent when the design effect is marginal (fig. 10). The optimal proportion of the indeterminate stratum that originates from each of the original strata in figure 7 is given in equation (19) and illustrated in figure 11:

$$\begin{aligned}\frac{d_{1,1} + d_{2,1}}{d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}} &\approx 0.5 - \left(\frac{e^{4.3k}}{25.6}\right)\left[P\left(A_1\right) - 0.5\right] \\ \frac{d_{1,2} + d_{2,2}}{d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}} &\approx 0.5 + \left(\frac{e^{4.3k}}{25.6}\right)\left[P\left(A_1\right) - 0.5\right]\end{aligned} \qquad (19)$$

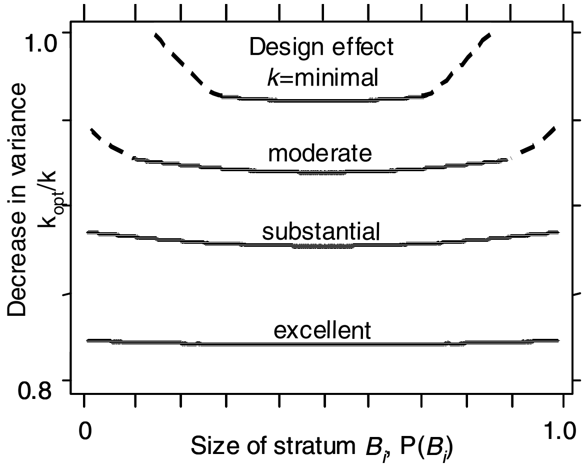


Figure 9—Proportional improvement in the design effect ($k_{opt}/k$) with addition of an indeterminate stratum for different initial design effects (table 1). The dashed lines indicate when stratum $B_1$ is merged with the indeterminate stratum to optimize efficiency (Condition 7).

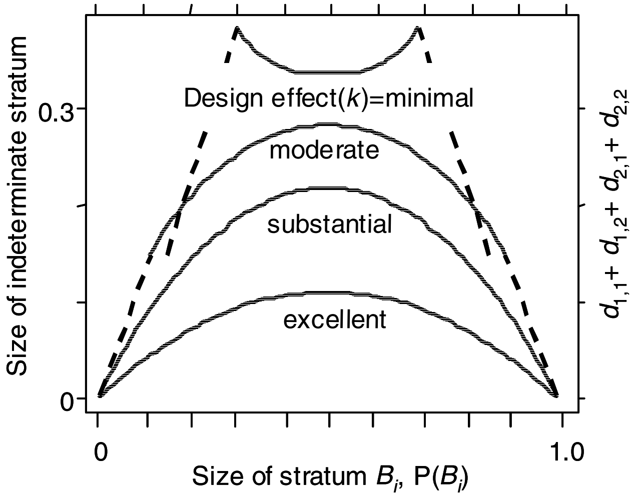


Figure 10—The optimal size of the indeterminate stratum, expressed as a proportion of the total population. Dashed lines indicate when stratum $B_1$ is merged with the indeterminate stratum to optimize efficiency (Condition 7).

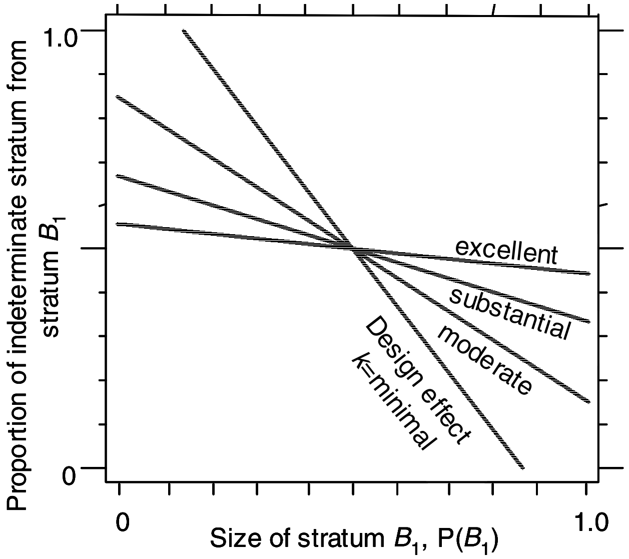


Figure 11—Proportion of indeterminate stratum originally part of stratum $B_1$, which corresponds to equation (19).

Let $P(A_1|B_1)_{OPT}$ represent the optimal classification accuracy in stratum $B_1$ after formation of the indeterminate stratum. Figure 12 shows $P(A_1|B_1)_{OPT}$ as a function of the original accuracy in stratum $B_1$, i.e., $P(A_1|B_1)$. We found that this relationship is approximately $P(A_1|B_1)_{OPT} \approx [0.75\ P(A_1|B_1)+0.25]$ in our model (figure 7 and equation 18), regardless of the initial design effect.

If accuracy is marginal and the stratum size is small (i.e., $P(B_j)<1.15k-0.67$), then the indeterminate stratum should be merged with the rare stratum to increase efficiency. This situation corresponds to the dashed lines in figure 9. However, gains do not exceed 5 percent with the model in figure 7 and equation 18. When the gain in efficiency is optimal, the indeterminate stratum contains 100 percent ($c_i$=1) of the rare stratum, plus approximately 25 percent ($c_j \approx$ 0.25) of the common stratum. The latter portion contains the most likely classification errors in stratum $B_j$.

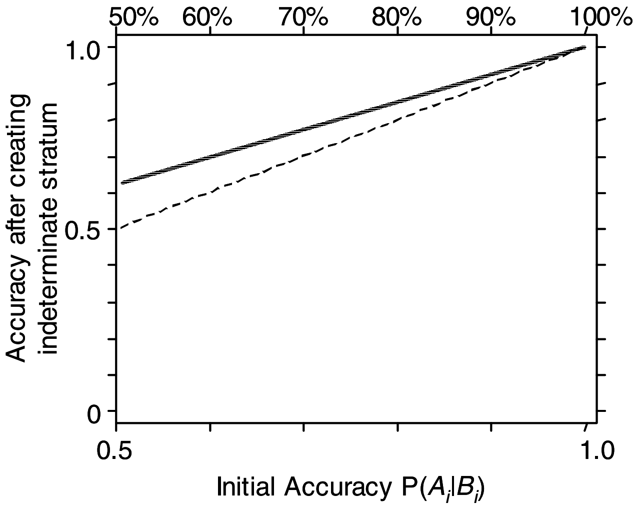


Figure 12—Optimal classification accuracy in stratum $B_1$ after formation of the indeterminate stratum, which contains elements that are most likely to have classification errors.

We recommend that the size of the indeterminate stratum be specified before exploring the sample data after they are collected; this avoids "over-fitting" to a given sample. Over-fitting can bias the estimated sampling error, thus producing a variance estimate that is smaller than its true value. In other words, our estimate would not be as precise as we assume, and analyses of these estimates can produce false conclusions. Figures 9 and 10, and equation 19, provide a priori specifications that can help practitioners follow our recommendation.

## REFERENCES

**Cochran, W.G.** 1977. Sampling techniques. New York. John Wiley. 428 p.

**Congalton, R.G.** 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment. 37: 35–46.

**Czaplewski, R.L.; Patterson, P.L.** [In preparation]. Classification accuracy for stratification with remotely sensed data. Submitted to Forest Science.

**Särndal, C.E.; Swensson, B; Wretman, J.** 1992. Model assisted survey sampling. New York: Springer-Verlag. 694 p.

# REMOTE SENSING PRECISION REQUIREMENTS FOR FIA ESTIMATION[1]

**Mark H. Hansen**[2]

**Abstrac**t—In this study the National Land Cover Data (NLCD) available from the Multi-Resolution Land Characteristics Consortium (MRLC) is used for stratification in the estimation of forest area, timberland area, and growing-stock volume from the first year (1999) of annual FIA data collected in Indiana, Iowa, Minnesota, and Missouri. These estimates show that with improvements in the classification, ground plot location and geo-rectification methods, we should be able to meet the accuracy standard for timberland area estimates (3 percent per million acres of timberland) in most areas. To meet the accuracy standard for growing-stock volume estimates (5 percent per billion cubic feet of growing stock) under the base sampling intensity (one plot per 5,937 ac on a 5-year cycle) that FIA has implemented in the North Central Region, we must be able to create meaningful volume classes from TM imagery or find other means to improve our estimates. Improvements in classification methods are most important in areas where forest land makes up a small portion of the total land area. Simulations based on observed FIA data, an existing classification of Landsat TM data, and various assumptions are used to examine the expected accuracy of FIA estimates when a complete cycle of annual inventories (5 years) has been completed.

## INTRODUCTION

FIA has made a commitment to an annual forest inventory with a base inventory of one ground plot every 5,937 ac with one nth of the plots measured every year over a cycle of n years. In the North Central region, with the cooperative efforts of state agencies, plans are to measure plots on a cycle of n=5 years, with a 20 percent sample of the plots measured each year. Once the annual forest inventory is fully implemented, basic estimates of current conditions (e.g., timberland area and growing-stock volume) can be made from the moving average of estimates made from all plots measured over the full cycle. The first year's ground plot measurement data collected under this system in the North Central region are now available for the states of Indiana, Iowa, Minnesota, and Missouri. These plots were measured between October 1998 and September 1999.

These first year plot data, together with a thematic GIS layer based on the classification of Landsat TM data for stratification purposes, are used here to produce estimates and sampling errors for forest area, timberland area, and growing-stock volume. Stratification methods similar to those used in the last two periodic inventories conducted by NCFIA (Hansen and Wendt 2000) are used with data produced by MRLC (Vogelmann and others, 1998) for stratification into four classes. Here I examine the sampling errors that can be expected once the entire cycle of data is available (4 years from now) to see how close to meeting national accuracy standards we can expect our estimates will be.

Sampling errors can be reduced by various means including improving stratification, measuring additional sample plots, and using other estimation schemes. Measuring additional sample plots is extremely costly. Other estimation methods are possible and are a topic of current discussion that goes beyond the scope of this paper. Here I examine how much better the classification of Landsat TM data must be in order to reduce sampling errors to meet the national accuracy standards. Stratified random sampling estimation is used throughout the study.

## DATA

NCFIA sampled 5,240 systematic plot locations in the four-state study area using the national FIA plot design, a cluster of four 1/24th acre fixed area subplots. Of these plot locations, 1,467 (28 percent) contained some forest land. These plots were located across the landscape following the grid system described by Brand and others (2000). Data available for each plot included the geographic position of the plot center (measured by GPS and/or digitized from geo-referenced Landsat imagery), the proportion of forest and timberland area on the plot (an observation from 0 to 100 percent), and the growing-stock volume (cubic feet per acre) on the plot. The methods, procedures, and definitions used to observe the attributes of interest are available on-line at http://fia.fs.fed.us/manuals/.

FIA estimates are commonly reported at the state and unit (group of counties) level. Figure 1 shows these units for the study area. In this study I have classified units as sparse (less than 10 percent forest), mixed (10 to 45 percent forest), or heavy (greater than 45 percent forest) based on the most recent periodic FIA inventory (fig. 2). Table 1 presents the total number of plot locations and the number that contained forest land for each unit.

The thematic GIS layer based on the classification of Landsat TM data used in this project is the National Land Cover Data (NLCD) prepared by the Multi-Resolution Land Characteristics Consortium (MRLC). MRLC used dual date (leaves-on and leaves-off) imagery that was resampled into a 30- by 30-m pixel format, an Anderson level 2 classification scheme, various ancillary data sources, and a single

[1] Paper presented at the Second Annual Forest and Inventory (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108.

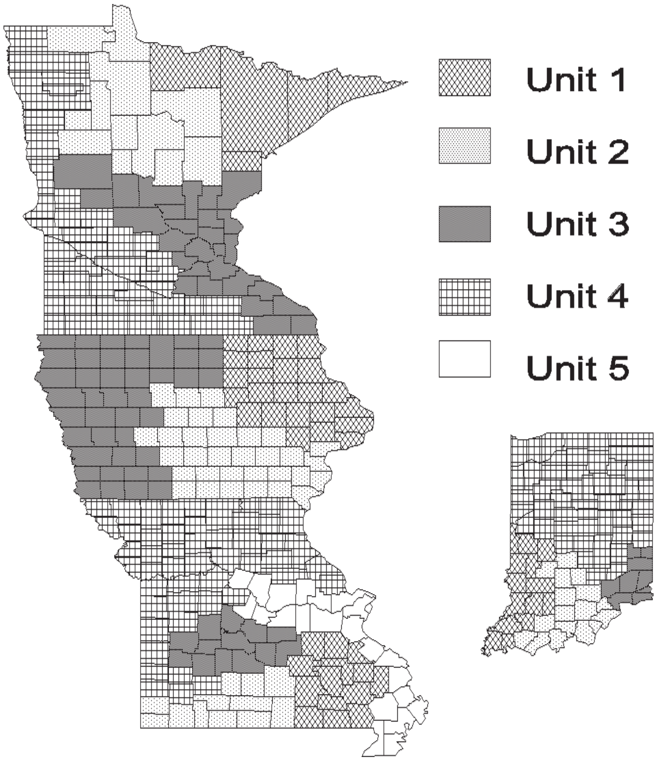


Figure 1—FIA inventory units in the study area.

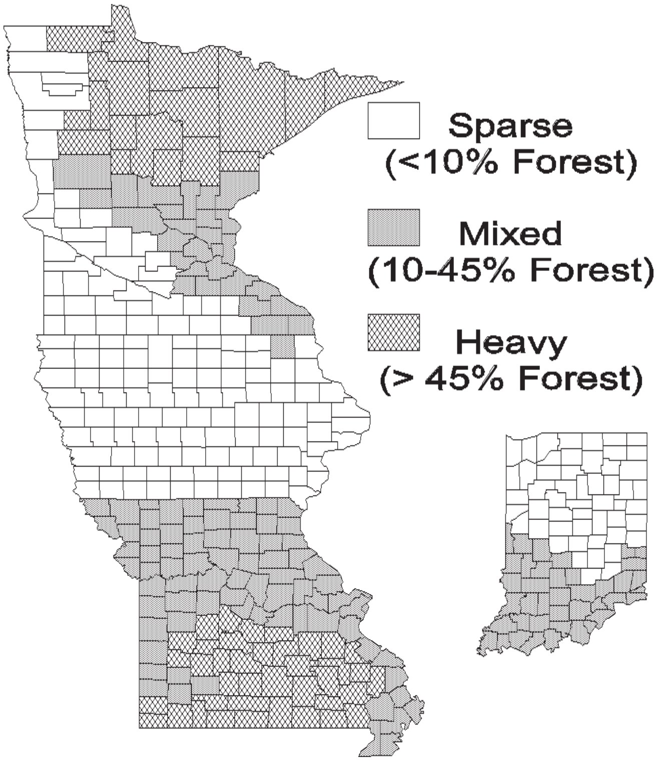


Figure 2—Sparse, mixed, and heavy forest FIA units.

pixel minimum mapping unit to classify geo-referenced imagery into the 21 classes shown in table 2. Imagery dates ranged from 1988 to 1994 over the study area. Detailed documentation of the methods used in classification and on-line access to the data are available at http://www.epa.gov/mrlc/.

## ESTIMATION

### Stratified Random Sampling

All estimates and sampling errors presented here are based on stratified random sampling estimators with stratification after the selection of the sample (poststratification) presented by Cochran (1977) with finite population correction ignored. The estimate of the population mean is the weighted average of the observed strata means from the sample

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1}$$

and the estimated sampling error is a weighted function of the within strata sampling errors

$$s_{\bar{y}_h} = \sqrt{\frac{1}{n}\sum_{h=1}^{L} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)\, s_h^2} \tag{2}$$

where *L* is the number of strata, *n* is the total number of observations (plots), $\bar{y}_h$ and $s_h^2$ are the observed (estimated) mean and variance in statum h, and $W_h$ is the stratum weight (proportion of the total population in stratum h).

**Table 1—Number of total plot locations and plot locations that contained forest land by FIA unit**

| | | | Number of plot locations | |
|---|---|---|---|---|
| | State | Unit | Total | Forest |
| Sparse (< 10 percent forest) | IN | 4 | 458 | 56 |
| | IA | 1 | 311 | 34 |
| | IA | 2 | 375 | 56 |
| | IA | 3 | 516 | 20 |
| | MN | 4 | 663 | 30 |
| | | Subtotal | 2,323 | 196 |
| Mixed (10 to 45 percent forest) | IN | 1 | 124 | 35 |
| | IN | 2 | 134 | 61 |
| | IN | 3 | 53 | 22 |
| | MN | 3 | 413 | 93 |
| | MO | 4 | 664 | 155 |
| | MO | 5 | 265 | 89 |
| | | Subtotal | 1,653 | 455 |
| Heavy (> 45 percent forest) | MN | 1 | 313 | 235 |
| | MN | 2 | 412 | 220 |
| | MO | 1 | 193 | 144 |
| | MO | 2 | 179 | 113 |
| | MO | 3 | 167 | 104 |
| | | Subtotal | 1,264 | 816 |
| | | Grand total | 5,240 | 1,467 |

Table 2—National land cover data classes

| Class | Description |
|---|---|
| 11 | Open water |
| 12 | Perennial ice/snow |
| 21 | Low intensity residential |
| 22 | High intensity residential |
| 23 | Commercial/industrial/transportation |
| 31 | Bare rock/sand/clay |
| 32 | Quarries/strip mines/gravel pits |
| 33 | Transitional |
| 41 | Deciduous forest |
| 42 | Evergreen forest |
| 43 | Mixed forest |
| 51 | Shrubland |
| 61 | Orchards/vineyards/other |
| 71 | Grasslands/herbaceous |
| 81 | Pasture/hay |
| 82 | Row crops |
| 83 | Small grains |
| 84 | Fallow |
| 85 | Urban/recreational grasses |
| 91 | Woody wetlands |
| 92 | Emergent herbaceous wetlands |

Here, the thematic GIS data (the pixels and their classification) define the population and sampling frame. These GIS data divides the population (total area of an inventory unit) into equal size pixels (sampling units) where each pixel has a distinct class (stratum) assigned to it. The ground plots provide an observation of the attribute of interest (y) for the specific pixel that contains the plot center. This observation is always a per unit area observation such as volume per acre. In the case of area estimation, for example, the area of timber land, this per unit area observation is a value from 0 to 1 that is the proportion of the ground plot that was observed to be timber land. Most estimates reported by FIA are totals rather than means and are the product of the estimated population mean $(\bar{y}_{st})$and the known total area of the population $(A)$ that is obtained from Bureau of Census data.

## FIA Accuracy Standards

FIA has set national accuracy standard for its inventories. These standards are defined for a specified area or volume. The standard for the estimate of total timberland area is 3 percent per million acres and for growing-stock volume 5 percent per billion cubic feet. The equation

$$e = \frac{(\text{observed sampling error})\sqrt{\text{estimated total volume or area}}}{\sqrt{\text{specified volume or area}}} \quad (3)$$

converts the observed sampling error as a percent for an estimate to a specified volume or area standard basis (typically 1 million ac or 1 billion ft$^3$). For example, an inventory that yields an estimated area of timberland of 4 million acres with a 2.0 percent sampling error would not meet the standard because $(2.0\text{ percent})\sqrt{4{,}000{,}000\text{ acres}}/\sqrt{1{,}000{,}000\text{ acres}} =$ 4.0 percent per million acres, which is greater than the accuracy standard (3 percent per million acres) for timberland area estimates. In this paper equation 3 is used to convert sampling errors to a per million acres or per billion cubic feet basis. Also, to convert observed sampling errors from the first year of annual inventory data to a full cycle basis when estimates will be made based on the moving average of 5 years of observations, sampling errors are divided by the square root of 5. Dividing sampling errors by the square root of 5 is equivalent to increasing the sample size by a factor of 5 (based on the assumption of a representative sample of the population), which simulates estimation based on the average of five independent estimates.

## Stratification

The NLCD data were used to create four strata (1-nonforest interior, 2-nonforest edge, 3-forest edge, 4-forest interior). This was accomplished in four steps:

1. NLCD classes 33 (transitional), 41 (deciduous forest), 42 (evergreen forest), 43 (mixed forest), 51 (shrubland), and 91(woody wetlands) were grouped into a single class (forest), and all other classes were grouped into a second class (nonforest).

2. A clump and sieve operation (ERDAS, 1997) was applied to this two-class image to create a two-class image with a minimum mapping unit of 1 ac (4 pixels).

3. Forest pixels within 2 pixels of any nonforest pixel were classified forest edge; all other forest pixels (those not within 2 pixels of nonforest) were classified forest interior.

4. Nonforest pixels within 2 pixels of any forest pixel were classified nonforest edge; all other nonforest pixels (those not within 2 pixels of forest) were classified nonforest interior.

In step 1, the NLCD shrubland class was included in the initial forest grouping because we found that some of the lands classified as shrubland by NLCD contained enough trees to meet the FIA definition of forest land. An example portion of the final reclassified image is shown in figure 3.

U.S. Bureau of Census data files were used in this study to provide the total area within each FIA unit and divide the imagery into FIA inventory units. TIGER county boundary files and ERDAS IMAGINE software were used to perform the data manipulation required to define the strata, match ground plot data to the appropriate pixels, and summarize the number of pixels by class and inventory unit. Oracle SQL programs were written to produce the estimates and sampling errors. In the estimation, the Bureau of Census information, together with the NLCD geo-referenced data, define the population (total area sampled).

## ESTIMATION WITH FOUR STRATA

The sampling errors of the area estimates are very dependent on the quality of the stratification. The estimate of forest area will have a low sampling error if the stratification is good, that is, if the forest interior stratum contains most of

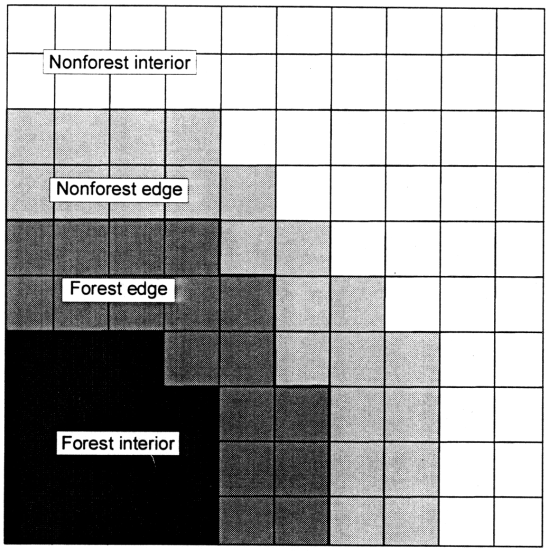


Figure 3—Example portion of the reclassified NLCD image.

the plots that are 100 percent forest, the nonforest interior stratum contains most of the that plots are 100 percent nonforest, and the two edge strata contain the plots that are on the forest/nonforest interface.

In meeting the national FIA accuracy standard for timberland area estimation, stratification becomes more important in areas where timberland is a small portion of the total land area. In populations that are more than 85 percent timberland, simple random sampling estimation will produce sampling errors less than 3 percent per million acres and statification is not required to meet the area accuracy standard at the current ground plot intensity. In populations that are 20 percent timberland or less, simple random sampling will produce sampling errors in excess of 7 percent per million acres. Figure 4 shows the expected sampling errors based on simple random sampling across the complete range of percent timberland when the total timberland area in the population is 1 million ac.

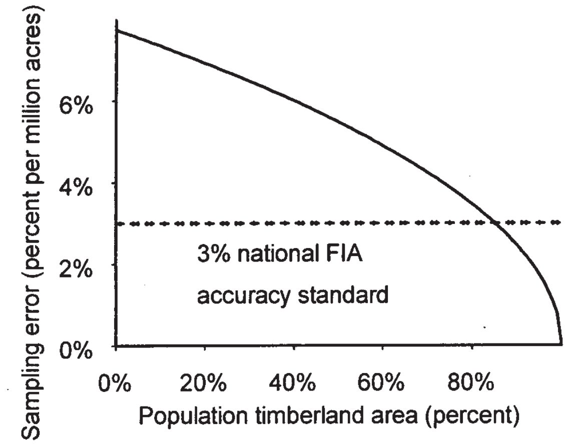


Figure 4—Expected sampling error (timberland area) from a simple random sample in a population of 1,000,000 acres of timberland and a sampling intensity of one plot per 5,937 ac.

The mean percent forest land within each of the four strata for all of the FIA units in the study area are shown in figure 5. This figure shows that the NLCD data with reclassification did a fairly good job of stratification. This figure is arranged with the heavily forested units at the top of the vertical scale and the sparsely forest units at the bottom. In all but one inventory unit (MO-2) the nonforest interior stratum contained less than 5 percent forest area, and in only one unit (MN-4) was the forest interior observed to contain less than 80 percent

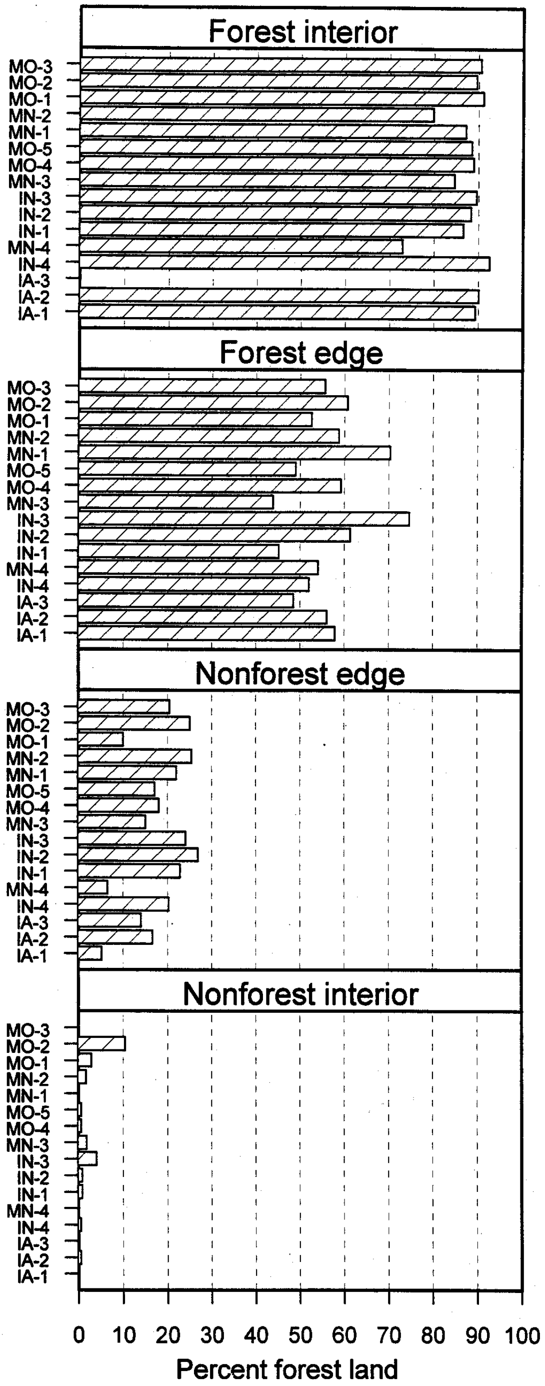


Figure 5—Mean observed percent forest land by FIA unit and the original four strata created from the NLCD data and the four steps described.

forest land. Because there was very little forest interior in western Iowa (IA-3), in that unit I combined the forest interior and forest edge strata into the forest edge strata in all analyses in this paper.

Table 3 shows the estimate of total forest area, timberland area, and growing-stock volume (based on the 1999 plot data and the stratification described previously) for the three groups of FIA units shown in figure 2 (sparse, mixed, and heavy forest) and sampling errors based on various assumptions, each a progressive improvement in the stratification. The first sampling error column in this table (sampling error for a single year estimate) is the actual estimated sampling error based on the data from only 1 year. This sampling error represents a sampling intensity of one plot per 29,685 ac and is the observed sampling error of the estimates in the first column of numbers (single year estimate). All other sampling errors in this table have been converted to a per million acres or per billion cubic feet basis for the moving average estimate given 5 years of data based on equation 3 divided by the square root of 5. The different columns simulate progressive improvements in classification that could possibly improve the estimation but still use the same four strata.

The column labeled “nothing added” is simply the expected sampling error with 5 years of data based on stratification using only the existing NLCD data and the procedures to define four strata as previously described. This column was computed by applying equation 3 to the numbers in the two columns to its left (to convert to per million or per billion) and dividing by the square root of five (to simulate the addition of 4 more years of data at the same intensity). It should be noted that the expected sampling errors in all cases failed to meet the FIA accuracy standards and that they are highest in the sparsely forested unit and lowest in the heavily forested unit. Without improvements to the classification, we should not expect to meet accuracy standards, especially in areas that are sparsely forested.

In an attempt to reduce sampling error by improving the stratification, I have simulated the effect that having perfect knowledge of all reserved lands would have on the estimates. Under this scenario, reserved lands would be treated as a subpopulation. This was simulated by moving reserved plots into a different population and moving a proportional area from the NLCD data. Reserved lands are areas such as parks and wilderness areas where timber harvest is prohibited.

By definition, plots on reserved lands have observed values of zero for timberland area and growing-stock volume. However, they are often forested and thus increase the variability in the forest strata in the estimation of timberland area and growing-stock volume. In application, the treatment of reserved lands as a subpopulation would be possible through the acquisition of good maps and/or GIS layers that show reserved lands from various land management agencies. This has been done in the past when aerial photos were manually interpreted but has not yet been

**Table 3—Estimates and sampling errors based on stratified random sampling estimation and the existing four strata with various improvements to the classification**

| | | | | Sampling error (percent per million acres or billion cubic feet) after 5 years with various improvements in classification | | | | |
|---|---|---|---|---|---|---|---|---|
| Units | Item | Single year estimate (million acres or billion cubic feet) | Sampling error for a single year estimate (% of estimate) | Nothing added | Add GIS layer for reserved land | Add GIS layer for reserved land and census water | Add GIS layer for reserved land, census water, and most farm lands | Add GIS layers and improve location data |
| Sparse | Forest area | 4.27 | 4.74 | 4.38 | 4.31 | 4.31 | 3.67 | 3.04 |
| | Tmbld area | 3.79 | 5.5 | 4.79 | 4.68 | 4.68 | 4.13 | 3.71 |
| | GS volume | 4.01 | 8.34 | 7.46 | 7.33 | 7.33 | 7.02 | 6.73 |
| Mixed | Forest area | 10.99 | 2.86 | 4.23 | 4.2 | 4.2 | 3.66 | 2.96 |
| | Tmbld area | 10.51 | 3.02 | 4.38 | 4.31 | 4.31 | 3.81 | 3.12 |
| | GS volume | 12.61 | 4.56 | 7.24 | 7.18 | 7.18 | 6.84 | 6.44 |
| Heavy | Forest area | 22.76 | 1.49 | 3.19 | 3.19 | 3.11 | 2.97 | 2.56 |
| | Tmbld area | 21.05 | 1.8 | 3.69 | 3.38 | 3.29 | 3.18 | 2.83 |
| | GS volume | 19.41 | 3.26 | 6.42 | 6.26 | 6.22 | 6.16 | 6.03 |

implemented into the NLCD classification. I simulated adding such a GIS layer for reserved lands by moving all the plots on reserved lands and their associated area to a new stratum. In all but one case, this action reduced sampling errors (one estimate of forest area showed no change in sampling error), although the sampling errors are still above the national accuracy standard. The biggest reductions in sampling errors came in the heavily forested units where large areas of reserved forest lands are found (e.g., Boundary Waters Canoe Area and Voyageurs National Park). Here, sampling errors for timberland area and growing-stock volumes decreased substantially (3.69 to 3.38 and 6.42 to 6.26 percent, respectively). However, sampling errors for forest area did not change (3.19 percent).

Adding information about census water is simulated in the next column using the same techique. The decrease in sampling errors that this produces is again only seen in the heavily forested units such as northern Minnesota where water is a fairly large portion of the total area. In application this could be done through GIS layers that are available from the Bureau of Census.

In the next column I simulated what would happen if we had access to a good GIS layer that could identify 80 percent of the agricultural lands in each unit. Since several government agricultural programs either have this type of GIS information available or are in the process of creating GIS information, it may be possible to identify major agricultural areas known to be nonforest and segment them in the estimation. In effect, the addition of these GIS layers is simply a method to treat areas we know are different as a subpopulation. The addition of these various GIS layers did reduce sampling errors, but it did not produce any sampling errors below the national accuracy standards. Other GIS layers for things such as urban areas, transportation, and other nonforest areas by definition may be available and could possibly improve the stratification somewhat.

Closer examination of the data showed that many of the ground plots that were totally misclassified (forest interior plots that contained no forest land or nonforest interior plots that contained 100 percent forest land) were within 2 pixels of an edge stratum. Sixteen of the total 24 (67 percent) nonforest interior plots that contained 100 percent forest land were within 2 pixels of nonforest edge and 29 of the total 54 (54 percent) forest interior plots that contained no forest land were within 2 pixels of forest edge. This suggests that poor geo-registration of the image and/or plot poor location information may be responsible for much of the error. To simulate the effect that improvements in location information (either ground plots or pixels) could reasonably have on the final estimates, I randomly moved 40 percent of these misclassified interior plots to the adjacent edge strata, along with a proportional amount of the total area, and produced the last column of table 3. In the heavily forested units, sampling errors for timberland area (2.83 percent) were less than the national standard, and in the mixed and sparse units they were reduced considerably (3.12 and 3.71 percent, respectively) and were not far above the standard. Sampling errors for growing-stock volume were reduced, but are still considerably above the 5 percent per billion cubic foot standard (6.03 to 6.73 percent). Figure 6 contrasts the mean percent forest land in these revised strata with the mean from the original strata by inventory unit. The net effect of the changes to the original data was an increase in average percent forest land in almost every case in the forest interior, forest edge, and nonforest edge strata. The net effect in the nonforest interior stratum was mixed. However, this stratum was greatly reduced in size relative to the other strata with the addition of the GIS layer for agricultural lands.

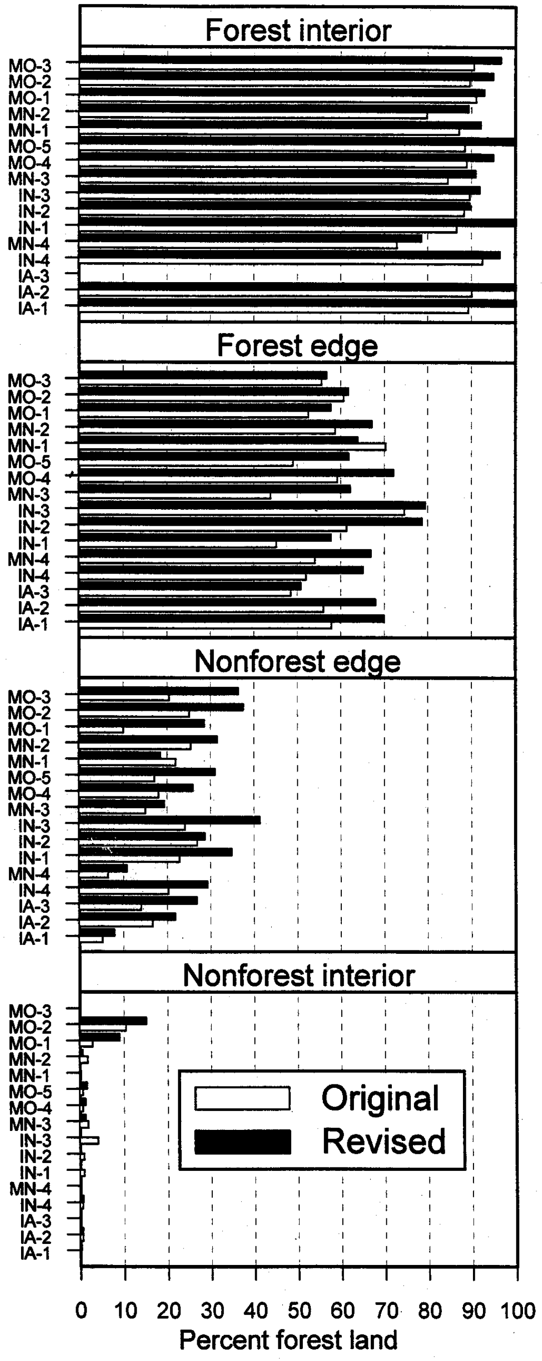


Figure 6—Mean percent forest land in original stratification contrasted with the mean percent forest land after simulation to improve the stratification by the addition of GIS layers and improved location information.

## ESTIMATION WITH ADDITIONAL VOLUME STRATA

It is apparent that strata are needed to improve volume estimates, or that some other method must be found to reduce sampling errors to attain the national accuracy standard for volume estimates. To reduce sampling errors from 6 to 5 percent by increasing sampling intensity would require 44 percent more plots. That would be very expensive. Here I simulate adding volume strata to see how the sampling errors for growing-stock volume change.

Initially, I subdivided both the forest interior and forest edge strata into two strata each. I did this by generating a value based on a function of observed basal area, stand age, and a random number, ordered the plots based on this value, and placed the top half in one stratum (high volume) and the other half in another stratum (low volume). The goal was to divide each of the two original forest strata into two equal size strata that approximate a fairly good classification. This procedure created four strata (forest interior-low, forest interior-high, forest edge-low, and forest edge-high) from the forest inerior and forest edge strata. Table 4 summarizes observed volume per acre in these classes. The high and low volume strata have significantly different means. Figure 7 shows the distribution of volume per acre for these strata. These simulated strata are different without being too good. There is significant overlap as would be expected in any real classification, and both classes have large numbers of plots with zero volume per acre. The effect that adding these two volume strata has on the estimation is shown in table 5. Sampling errors for the area estimates change only a little compared to large reductions in the sampling errors on volume. This stratification did not meet the national accuracy standard for growing-stock volume in any of the units.

**Table 4—Summary of observed volume per acre with the addition of two volume strata**

| Units | Stratum | Mean | Standard deviation |
|---|---|---|---|
| | | *Cubic feet per acre* | |
| Sparse | Forest interior - low | 636 | 613 |
| | Forest interior - high | 1,670 | 674 |
| | Forest edge - low | 365 | 693 |
| | Forest edge - high | 853 | 734 |
| Mixed | Forest interior - low | 690 | 578 |
| | Forest interior - high | 1,662 | 810 |
| | Forest edge - low | 444 | 591 |
| | Forest edge - high | 1,101 | 858 |
| Heavy | Forest interior - low | 444 | 404 |
| | Forest interior - high | 1,138 | 703 |
| | Forest edge - low | 362 | 442 |
| | Forest edge - high | 1,002 | 743 |

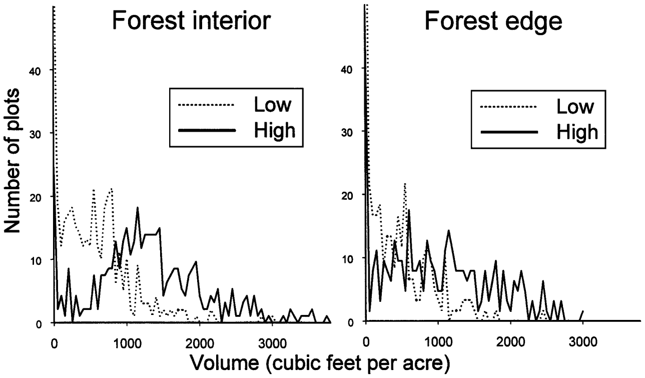


Figure 7—Distribution of plot volume per acre within the simulated high and low volume strata.

A second stratification into three volume classes (high, medium, and low) was created using a similar procedure. These strata are summarized in table 6 with the distributions shown in figure 8. Table 5 show the 5 percent per billion cubic feet standard was met in the heavily forest units (4.91 percent) but not quite met in the mixed and sparse units (5.31 and 5.66, respectively). The addition of the three volume strata also improved the timberland area estimates enough to meet the accuracy standard in the mixed forest units by stratifying many of the misclassified forest plots into the low volume strata. In the sparse forest units, even better stratification would be needed to meet both the area and volume standards. Since area and volume estimation are linked, it is difficult to meet volume standards without also meeting the area standards.

These simulations suggest that fairly good volume classification is needed to meet national accuracy standards for growing-stock volume estimation. Figure 9 shows the

**Table 5—Sampling errors with the addition of simulated volume classes**

| | | Sampling error after five years with various improvements in classification | | |
|---|---|---|---|---|
| Units | Item | Without volume classes | With 2 volume classes | With 3 volume classes |
| | | *%/million ac or billion ft$^3$* | | |
| Sparse | Forest area | 3.04 | 3.07 | 2.97 |
| | Tmbld area | 3.71 | 3.78 | 3.69 |
| | GS volume | 6.73 | 6.32 | 5.66 |
| Mixed | Forest area | 2.96 | 2.99 | 3.01 |
| | Tmbld area | 3.12 | 3.15 | 3.15 |
| | GS volume | 6.44 | 5.64 | 5.31 |
| Heavy | Forest area | 2.56 | 2.57 | 2.58 |
| | Tmbld area | 2.83 | 2.83 | 2.83 |
| | GS volume | 6.03 | 5.29 | 4.91 |

**Table 6—Summary of observed volume per acre with the addition of three volume strata**

| Units | Stratum | Mean | Standard deviation |
|---|---|---|---|
| | | *Cubic feet per acre* | |
| Sparse | Forest interior - low | 425 | 506 |
| | Forest interior - medium | 1,017 | 559 |
| | Forest interior - high | 1,840 | 556 |
| | Forest edge - low | 273 | 744 |
| | Forest edge - medium | 577 | 336 |
| | Forest edge - high | 1,106 | 821 |
| Mixed | Forest interior - low | 584 | 616 |
| | Forest interior - medium | 798 | 467 |
| | Forest interior - high | 1,790 | 740 |
| | Forest edge - low | 311 | 559 |
| | Forest edge - medium | 668 | 453 |
| | Forest edge - high | 1,410 | 878 |
| Heavy | Forest interior - low | 339 | 366 |
| | Forest interior - medium | 646 | 372 |
| | Forest interior - high | 1,291 | 690 |
| | Forest edge - low | 425 | 506 |
| | Forest edge - medium | 1,017 | 559 |
| | Forest edge - high | 1,840 | 556 |

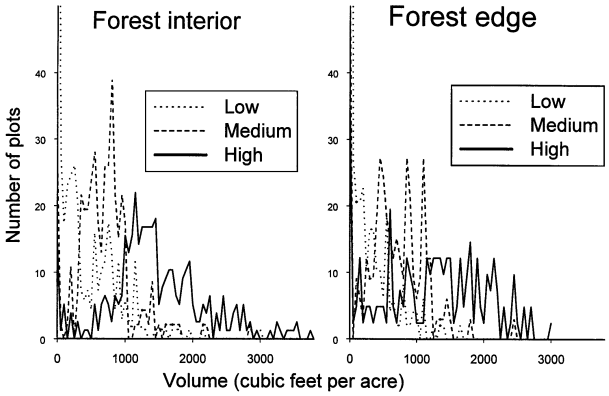


Figure 8—Distribution of plot volume per acre within the simulated high, medium, and low volume strata.

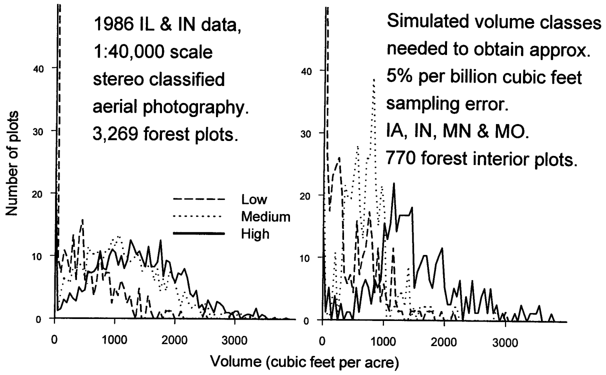


Figure 9— Distribution of the volume per acre data summarized in table 7. The data on the left are from manual stereo interpretation of aerial photos (1980's Illinois and Indiana inventory data). The data on the right are the simulated Landsat TM classification (1999 forest interior strata from Iowa, Indiana, Minnesota, and Missouri) that produced volume estimates of approximately the national standard (5 percent per billion cubic feet). Total number of plots in all strata have been adjusted to an equal basis to aid in visual comparison.

distribution of FIA plots measured in the mid-1980's FIA inventories in Illinois and Indiana. In those inventories, 3,269 FIA plots were classified into nine size-density classes using stereo pairs of 1:40,000 aerial photos. Based on this classification, I created the three best strata I could find by collapsing the three classes with the highest average volume per acre into one strata (high) and similarly collapsing the other classes to create medium and low volume strata. This figure also shows the distribution of the forest interior plots in the simulated three volume classes. To assist in visual comparison, I have adjusted the scale to a common total number of plots in all strata. Table 7 summarizes volume per acre for these two data sets. The two distributions have similar shapes. However, it appears that the simulated strata do a better job of distinguishing between the high and medium volume plots and that the low volume stratum produced from the aerial photos was somewhat better than the simulated low volume stratum.

There are a number of differences between the two data sets contrasted in table 7 and figure 9. The data sets are from different regions (Illinois and Indiana vs Indiana, Iowa, Missouri, and Minnesota), used different ground plot designs, and were measured at different times. They are only shown here to illustrate that the quality of volume classification needed from Landsat TM to meet or nearly meet existing national FIA accuracy standards is about the same as we were obtaining from manual interpretation of aerial photos.

**Table 7—Comparison of observed volume per acre within three volume classes, manual stereo interpretation of aerial photos (1980's Illinois and Indiana inventory data) vs simulated Landsat TM classification (1999 forest interior strata from Iowa, Indiana, Minnesota and Missouri)**

| | Method | | | |
|---|---|---|---|---|
| | Manual stereo interpretation of aerial photos | | Simulated Landsat TM classification | |
| Class | Mean | S.D. | Mean | S.D. |
| Low | 224 | 426 | 402 | 456 |
| Medium | 1,003 | 725 | 706 | 420 |
| High | 1,272 | 797 | 1,466 | 738 |

## DISCUSSION

The analysis and simulations I have done here suggest that through the use of Landsat TM imagery, forest/nonforest classification, ancillary data sources (GIS layers), and improved registration and GPS locations, we will meet or nearly meet the FIA accuracy standards for area estimation at the one plot per 5,937 ac sampling intensity we have implemented. To meet sampling accuracy standards for growing-stock volume estimation, additional stratification is needed. Major investments in improving stratification must be made to obtain the degree of classification needed to meet the volume accuracy standards. To meet these accuracy standards given this sampling intensity, we must be able to identify volume classes from remote sensing with about the same degree of accuracy as we did using manual interpretation of aerial photos in past applications.

## ACKNOWLEDGMENTS

I thank Daniel G. Wendt, Supervisory Computer Specialist at North Central Research Station, for his assistance in processing the Landsat TM data used in this study.

## REFERENCES

**Brand, G.J.; Nelson, M.D.; Wendt, D.G.; Nimerfro, K.K.** 2000. The hexagon/panel system for selecting FIA plots under an annual inventory. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C., eds. Proceedings of the 1st annual forest inven-tory and analysis symposium; 1999 November 2–3; San Antonio, TX. Gen. Tech. Rep. NC–213. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 8–13.

**Cochran, W.G.** 1977. Sampling techniques. New York: John Wiley. 413 p.

**ERDAS,** 1997. ERDAS field guide, 4th ed. Atlanta, GA: ERDAS, Inc. 656 p.

**Hansen, M.H.; Wendt, D.G.** 2000. Using classified landsat thematic mapper data for stratification in a statewide forest inventory. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C., eds. Proceedings of the 1st annual forest inventory and analysis symposium; 1999 November 2–3; San Antonio, TX. Gen. Tech. Rep. NC–213. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 20–27.

**Vogelmann, J.E.; Sohl, T.; Campbell; P.V.; Shaw, D.M.** 1998. Regional land cover characterization using Landsat thematic mapper data and ancillary data sources. Environmental Monitoring and Assessment. 51: 415–428.

# A NONPARAMETRIC GEOSTATISTICAL METHOD FOR ESTIMATING SPECIES IMPORTANCE[1]

**Andrew J. Lister, Rachel Riemann, and Michael Hoppus**[2]

**Abstract**—Parametric statistical methods are not always appropriate for conducting spatial analyses of forest inventory data. Parametric geostatistical methods such as variography and kriging are essentially averaging procedures, and thus can be affected by extreme values. Furthermore, non normal distributions violate the assumptions of analyses in which test statistics are generated and compared to a theoretical distribution, such as analysis of variance or stepwise multiple linear regression. Here, we offer guidelines and an example of the use of the indicator approach for dealing with nonparametric data distributions, using data from a study conducted in northern Vermont and New Hampshire.

## INTRODUCTION

Recently authors have used the USDA Forest Service's database of Forest Inventory and Analysis (FIA) plots to produce maps of species distributions (Iverson and others 1999, Moeur and Riemann Hershey 1999, Riemann Hershey and others 1997), pockets of high-value commercial trees (King 2000), and forest distribution (Zhu 1994). The methods used to produce these maps have varied from geostatistical simulation (Riemann Hershey and others 1997, King 2000) to advanced multivariate regression-based techniques (Zhu 1994, Moeur and Riemann Hershey 1999, Iverson and others 1999). Few of the techniques, however, have addressed the theoretical and practical problems associated with analyzing highly skewed distributions using parametric statistics. For example, positively skewed data distributions can affect semivariance calculations and kriging weights if the extreme values are located within patches of homogeneous patches of low values. Similarly, traditional statistical methods, such as multiple linear regression, determine the significance of a given model by calculating an F statistic and comparing it to a theoretical distribution. If the data from which the model was built are not normally distributed, erroneous inferences can be made.

Geostatistical techniques such as ordinary kriging and its variants do not inherently require normally distributed data; rather, they assume a multi-point Gaussian random function, described thoroughly in Isaaks and Srivastava (1989), Goovaerts (1997) and Myers (1994). The random function model, which can actually be thought of as a conceptual model, was formulated in part to account for the inherent uncertainty surrounding a set of spatially referenced observations. A random function, in effect, is a set of random variables for each location within a given spatial domain. A random variable is a variable whose values at any location are determined by some probabilistic mechanism. In other words, a reported estimate is drawn from a distribution of estimates that have some probability of occurring at the estimate's location.

Each data point can be conceived of as a random variable whose true value is known, and each estimate to be made is a linear combination of random variables (the known data).

The distribution of values making up a random variable can be described by a cumulative distribution function (CDF), or, as with class variables, by a probability density function (PDF). At any unknown location, a CDF constructed without any additional information regarding the form of the random variable would resemble that in figure 1A. In this situation, the best estimate of a variable at an unknown location would be the sample mean. However, in many earth science datasets, data are spatially dependent, and this knowledge can be used to update the CDF to one that might resemble that in figure 1B. In this instance, one might choose different percentiles of the CDF as an estimate to report, depending upon the goals of the study.

For example, assume there is a set of spatially referenced observations of the importance of species X within a study area. The importance of this species might be dependent upon variables such as soil chemistry, climatic factors, topographic relationships, or the presence or absence of other species. In general, these factors can be assumed to vary relatively smoothly across space. It can be inferred, thus, that areas with high levels of species X are surrounded by other areas of high levels of species X. In other words,

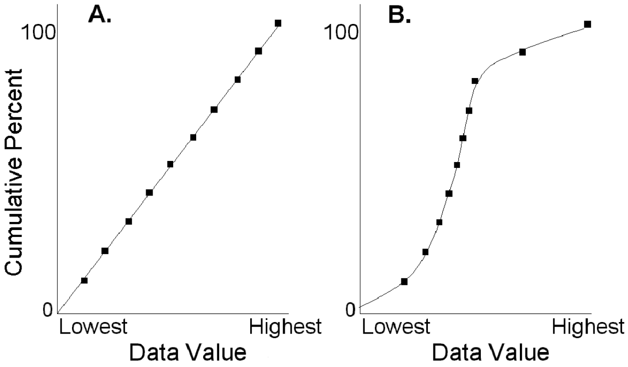


Figure 1—Example of two cumulative distribution functions defining the random variable at an unknown location: A—when no additional information is known about the values; B—when additional information, such as the form of the model of spatial dependence, is known. Points can be defined along B using indicator geostatistics.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Forester, Research Forester, and Research Forester, USDA Forest Service, Northeastern Research Station, 11 Campus Blvd., Newtown Square, PA 19008, respectively.

the random function exhibits spatial dependence, or autocorrelation. Autocorrelation, an index of similarity analogous to variance, can be calculated for points separated by distances placed in discrete distance classes. Once the relationship between autocorrelation and separation distance is modeled using variography, the variogram can be used in the estimation procedure (e.g., kriging) to update the CDF to create a conditional CDF (CCDF). Therefore, a given location's CDF is altered, or conditioned on the surrounding data, using the model of spatial dependence constructed for the random function. Under the multi-Gaussian assumption (i.e., that the random variables composing the random function are normally distributed), the mean of the random variable is the simple kriging estimate at that location, and the variance is the simple kriging variance (Isaaks and Srivastava 1989, Goovaerts 1997, Myers 1994).

Again, it is important to note that the non-normality of a distribution of samples does not necessarily imply a non-multiGaussian random function. An observed set of samples can be thought of as one realization of the random function; i.e., the samples could theoretically have had an infinite number of distributions. In geostatistics, the problem with non-normal distributions is that the modeling procedure (the variography) and the estimation procedure (the kriging) are essentially averaging techniques, and can be affected by small numbers of extreme values, a common phenomenon in earth science datasets. To resolve this, sample data with highly skewed distributions are normal score transformed (Deutsch and Journel 1998, Goovaerts 1997). In this procedure, the original CDF is mapped onto the standard normal CDF, giving the transformed distribution perfect symmetry, with a mean of zero and a standard deviation of one. The relationship between the two distributions is defined on a case-by-case basis so that after the estimation procedure is carried out, back transformation can be performed (Deutsch and Journel 1998).

The normal score transform is useful when performing conventional statistical estimation as well. For example, with stepwise multiple regression, variable inclusion and parameter estimates are determined by calculating test statistics that are compared to a reference distribution that was created under the assumption of normality. Large deviations from normality can lead to undesirable outcomes such as heteroscedasticity in regression residuals (Zar 1984), and thus should be carefully evaluated.

## INDICATOR GEOSTATISTICS

Another estimation method used when working with highly skewed data distributions is indicator geostatistics. The goal of this paper is to present the general theory and methodology behind indicator kriging (IK), and to present an example of indicator kriging with varying local means (IKLVM). Both are nonparametric geostatistical approaches that avoid many of the abovementioned pitfalls. In IK and IKLVM, the CCDF is constructed by defining discrete points across the entire range of data values (fig. 1B), and then interpolating between these points to arrive at the completed CCDF for each point to be estimated. IK is a univariate approach, while IKLVM allows for the incorporation of ancillary covariates into the estimation procedure.

## INDICATOR KRIGING

In IK, the first step in defining the CCDF values at a given location is to determine a series of threshold values or cutoffs from within the range of data values. In practice, deciles of the sample data distribution are chosen because the goal is to define the CCDF for the entire data range. Once these thresholds have been determined, each value is coded as a "1" or a "0", with "1" being assigned to values below that threshold, and "0" being assigned to values above the threshold. Thus in the example where deciles of the distribution are used as cutoffs, 10 datasets consisting of 1's and 0's will be created, one for each cutoff.

The second step in defining the CCDF values is to model the spatial autocorrelation for each of the coded datasets (e.g., deciles) using variography. In order to construct smoothly varying CCDF's, Goovaerts (1997) recommends using the same model or combination of basic models for all of the variograms. For example, the indicator variograms (correlograms) for each of seven percentiles of a dataset shown in figure 2 should be modeled using the same basic structure.

Once these variograms have been created, IK is performed. In IK, each estimate is actually a weighted average of the sample data (1's and 0's) surrounding it, with the weights being derived from the variogram. An IK estimate can be interpreted as the probability of an outcome of 1, or, more specifically, the probability of the actual value at that location being below the threshold used to code the data. This process is repeated for each point to be estimated, and for each threshold dataset. In effect, in keeping with the above decile example, 10 continuous probability maps are created with the value of each pixel being the probability of falling below the threshold used to code the data. Thus, 10 pairs of x,y coordinates (cutoff value, probability of being less than that cutoff value) can be obtained for each location, and 10 points can be placed along the CCDF as in figure 1B. Interpolating between and extrapolating beyond these discrete points to fill in the CCDF should be undertaken with care; guidelines are given in Goovaerts (1997) and Deutsch and Journel (1998).

Figure 2—Examples of indicator variograms obtained from spruce-fir importance data from a study in northern New Hampshire and Vermont. See Lister and others (2000) for details. All of the variograms from different cutoff levels show some degree of spatial dependence.

The final step of the process is to choose a percentile of the CCDF to report as a value. Goovaerts (1997) gives a detailed discussion of different criteria that might be employed to make this decision. In essence, the decision will be based on some sort of an optimality criterion defined by the goals of the study. The following section will not only present an example of an indicator technique in practice, but also elucidate some of the steps one might take to develop an optimality criterion for the choice of the estimate to report.

## INDICATOR KRIGING WITH LOCALLY VARYING MEAN—AN EXAMPLE

A study was conducted using data from 760 FIA plots in northern New Hampshire and Vermont (for details see Lister and others 2000). The relative importance (relative basal areas) of a combination of red spruce (*Picea rubens* Sarg.) and balsam fir (*Abies balsamea* L.) was determined on each plot, and IKLVM was applied. IKLVM is in principle identical to univariate IK, however the CCDF values are determined by a combination of logistic regression (Montgomery and Peck 1982) and simple kriging of the residuals of the logistic regression. The technique is implemented in a manner similar to that of simple kriging with varying local means, described in Metzger (1997), Majure and others (1996), and Hunner and others (1998). The general expression for the logistic regression estimate is

$$E(y \mid xi..xn) = \frac{\exp(b0 + b1 * x1 + ... + bn * xn)}{1 + \exp(b0 + b1 * x1 + ... + bn * xn)}$$

where E(y) can be interpreted as the probability of an outcome of “1” occurring (assuming that the data are coded as 0 and 1), b0…bn are the coefficients and x1…xn are the ancillary data layers.

The first step of the process was to determine the appropriate thresholds as described above. The frequency histogram of the data indicates a strong positive skewness (fig. 3), with 30 percent of the data values having 0 percent spruce-fir importance. Consequently, the cutoffs chosen for indicator coding were the 30th – 90th deciles of the original data’s distribution, or values of 0, 3, 10, 17, 27, 40 and 60 percent spruce-fir importance. Once these seven coded data sets were constructed as described above, logistic regression was applied to determine, for each point to be estimated, the probability of being below or equal to one of the seven cutoff values (an outcome of “1”).

Figure 3—Frequency histogram of the spruce-fir importance data used in the example. The distribution exhibits a strong positive skew, with 50 percent of the data falling in the first class.

In order to build the logistic regression model, exhaustively sampled data layers (i.e., secondary data that were collocated with FIA plots and at all points to be estimated) were chosen based on a combination of user judgment, exploratory data analysis, and stepwise logistic regression. The variables included in the final regression model were Landsat band 4, digital elevation model (DEM) -derived slope and the square root of elevation, and the square root of latitude. One logistic regression model was built for each of the seven cutoffs. All regressors contributed significantly to the model for each cutoff at the 0.01 level, with the majority being significant at the 0.0001 level.

The value of each pixel in the maps in figure 4 was predicted using the logistic regression model for that cutoff. Each value represents the probability that a pixel falls below the cutoff used to code the data. For example, the upper left map represents the probability of a pixel’s value being lower than or equal to the original data’s 30th percentile, which is 0 percent spruce-fir importance. The highest probabilities (lighter pixels) of having 0 percent spruce-fir importance are seen in areas such as valleys or in clearly nonforest areas, where the logistic regression procedure yielded low values. As the cutoff values increase, the amount of area which probably falls below that cutoff’s level of spruce-fir increases until finally, at the 90th percentile, spruce-fir forest has a relatively high probability of occurring at importance levels of at most 60 percent everywhere except forested ridge tops far from roaded areas (shown as darker areas on the map). In these areas, the forests have a high chance of having greater than 60 percent spruce-fir importance.

The next step of the IKLVM procedure was to calculate the regression residuals from each model by subtracting the probabilities from the regression output from the coded data. These residuals are then assessed for spatial dependence using variography. In our example, the inverted correlograms (hereafter referred to as variograms) of the residuals indicate that spatial dependence does exist in the regression residuals (fig. 5). None of the variograms exhibited substantial anisotropy, i.e. the model of spatial continuity did not change with direction.

The next to final step was to use simple kriging to estimate for each map the error at every point in the study area, based on the variograms of the logistic regression residuals (fig. 5). These error (residual) maps were then combined with the regression-based maps with simple addition to arrive at updated maps of IKLVM probability estimates. It is these updated estimates that were used to complete the CCDF.

To reiterate, we created IKLVM maps for each of our cutoff values. If we were to stack these maps one on top of the other and randomly sample any pixel of the stack, we would obtain a set of x, y coordinates that could be used to locate

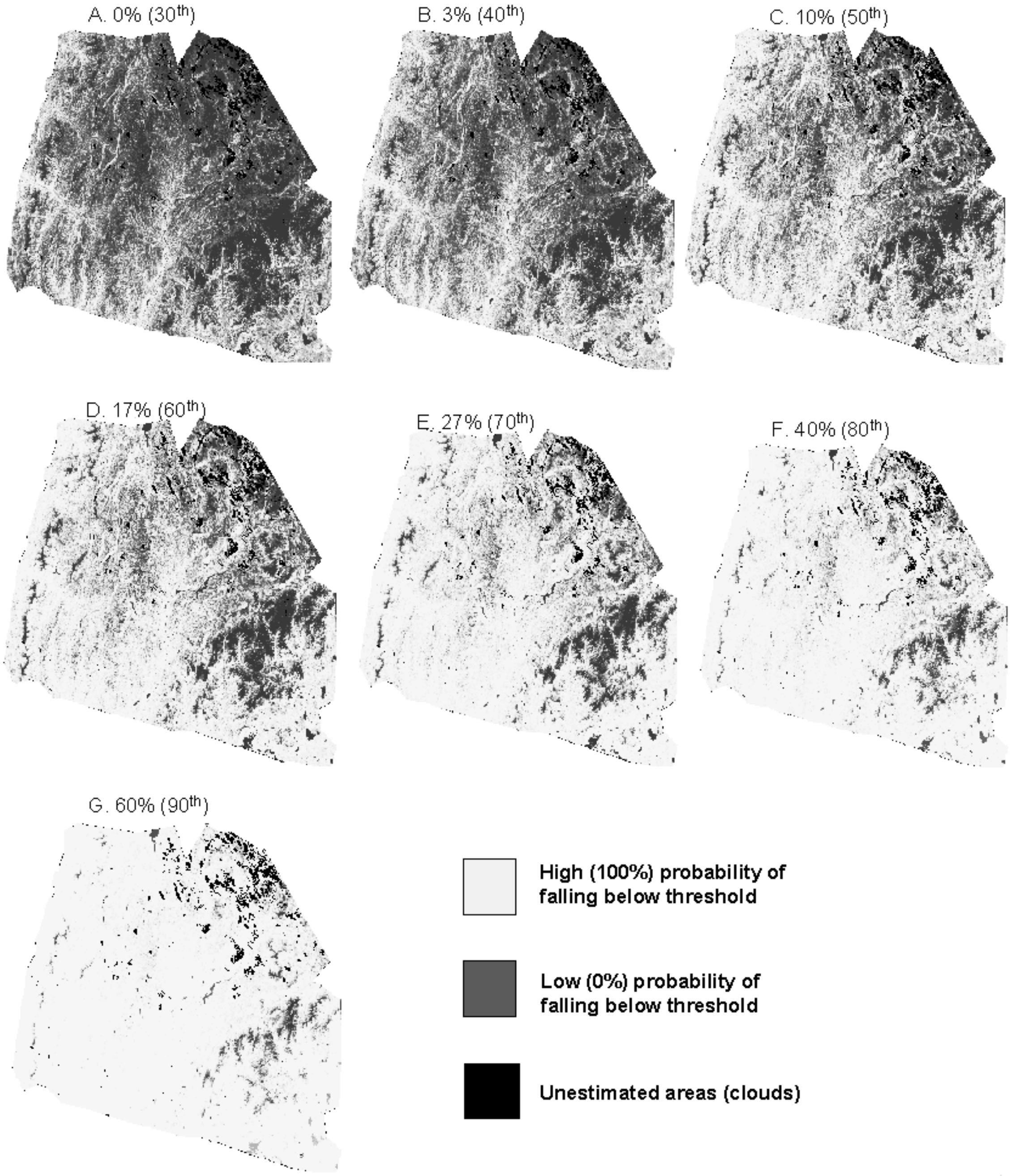

A. 0% (30th)
B. 3% (40th)
C. 10% (50th)
D. 17% (60th)
E. 27% (70th)
F. 40% (80th)
G. 60% (90th)
High (100%) probability of falling below threshold
Low (0%) probability of falling below threshold
Unestimated areas (clouds)

Figure 4—Logistic regression maps constructed from data coded as 1 or 0, based on whether they fall below the indicated threshold percentage of spruce-fir importance. Light pixels have higher probabilities of falling below that cutoff than dark pixels. The effects of topography are readily apparent in the southeast part of the study area.

discrete points on a CCDF similar to that in figure 1B. It is important to remember that each pixel in the map has its own CCDF. In our example, we chose to fill in the CCDF by implementing linear interpolation between points in the center of the distribution and hyperbolic interpolation in the tails. This choice was made based on examination of the resulting CCDF's and assessment of their plausibility, as well as on the suggestions of Deutsch and Journel (1998) and Goovaerts (1997).

Once we created our percentile maps and arrived at a CCDF for each pixel, the final step in the IKLVM approach was to choose a percentile of the distribution to report as the final estimate. Goovaerts (1997) and Deutsch and Journel (1998) discuss criteria that can be used to make this decision. In general, they suggest that the user establish an "optimality criterion", or set of conditions that a "good" estimate must satisfy, and then use this criterion to make the choice. Figure 6 shows scatterplots for both the model fit (A-G) and a set of

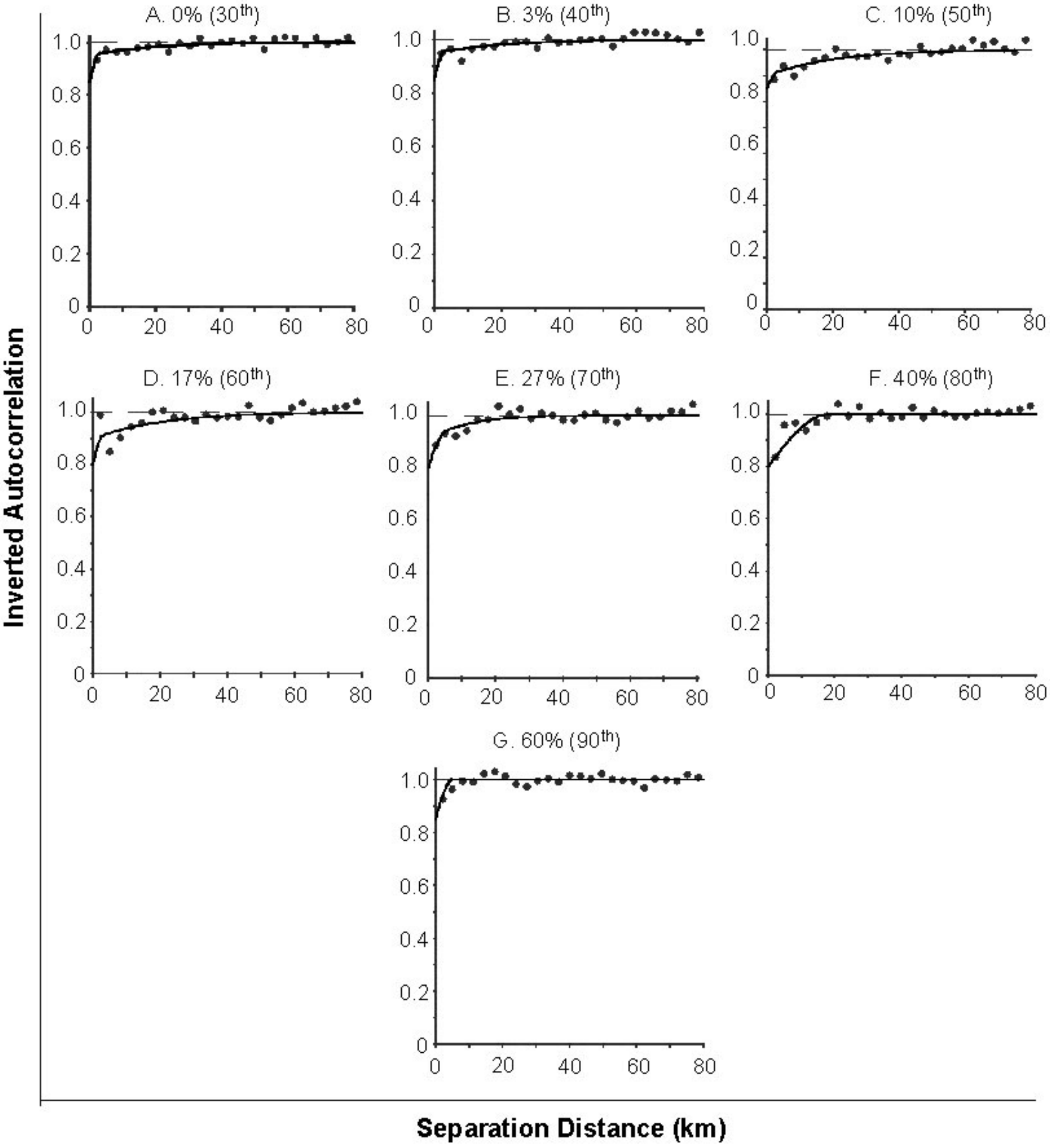


Figure 5—Variograms of residuals from the logistic regression procedure for the indicated cutoff. All residual variograms exhibit spatial dependence.

validation data's fit (H-N) for a range of deciles of the CCDF (the 20th through the 80th) that probably encompasses the final estimate. The dashed line passing through the cloud of points is the 45-degree line, along which all points would fall if the model perfectly predicted the sample points. The dark line is a least squares best fitting line describing the relationship between the actual value (x axis) and the estimate (y axis). The closer the agreement between the least squares best fitting line and the 45-degree line, the more accurate the model is, on average.

It is apparent that for low percentiles, the model dramatically underestimates (fig. 6). For larger percentiles, however, the situation is reversed with values being over predicted. Intermediate values tend to be predicted the best near the middle of the distribution. For some applications, the user might be very concerned about correctly estimating values close to zero, for example, when trying to accurately locate areas with small amounts of some rare but valuable tree species. Before investing in field reconnaissance, a user might want to be as certain as possible that the species occurs at a location, so he or she might choose a percentile where the values are underestimated, for example, the 20th or the 30th percentiles. Similarly, a user might want to find all areas where there are large amounts of a species of interest, in which case the user might choose a percentile

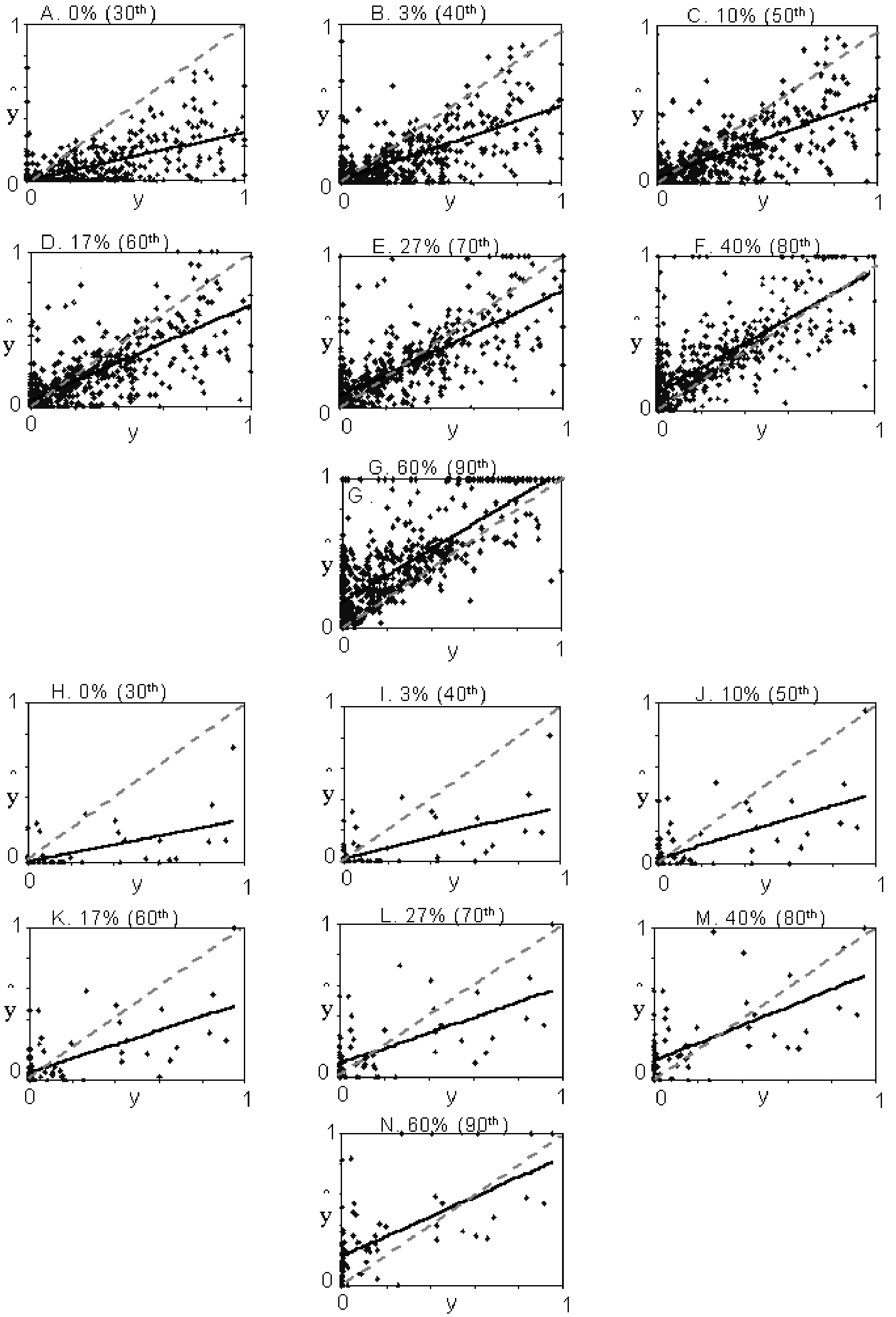


Figure 6—Scatterplots of the model fit of estimates from the 20th to 80th deciles (A-G, respectively), and for the 70 validation data (H-M, respectively). The actual data value is on the x axis, and the predicted value is on the y axis. The gray dashed line is the 45-degree line (perfect agreement), and the black solid line is the least squares best fitting line through the points.

such as the 80th or 90th. The percentile of the CCDF to report as an estimate is therefore chosen based on the relative impact of overestimation or underestimation.

Another criterion might be avoidance of heteroscedasticity of the residuals. The percentile values surrounding the median (figure 6D, E, F and K, L, and M) appear to exhibit roughly equal variance for the entire range of estimates, with the points approximately following the 45-degree line. In our example, we might choose the 70th percentile based on this criterion.

Another criterion that can be applied is the ability of a given percentile to produce estimates with a distribution that resembles that of the original data, either for certain areas of the distribution, or for the distribution as a whole. Figure 7 shows a histogram of estimates from each percentile compared to that of the original data. Using this criterion, the estimates from the 40th percentile (the "x" symbol) of the CCDF's most closely agree with the original data in the first class, which encompasses the lowest importance values for spruce fir (< 0.05, or 5 percent importance) and has the largest class occupancy.

Table 1 shows the results of a quantitative method of comparing the estimate histograms with that of the original data. One might seek to minimize the squared difference between percentages of estimates in each class for the different techniques. In addition, one might want to weight these differences by the magnitude of class occupancy because differences in very populous classes might be more important than differences in less populous classes. For example, for the first class in table 1, the value of 301.991 was arrived at by squaring the difference between the

Figure 7—Histograms of the original data (dashed line) and the estimates from each of seven percentiles (solid lines; see legend for details). The 40th percentile histogram has the closest agreement with the original data's histogram.

**Table 1—Quantitative assessment of the differences between the histograms in figure 7. For each percentile, the weighted average squared difference between the percentile histogram and the original data's histogram was calculated for each class. The 40th percentile has the minimum weighted average difference**

| | Percent of | Squared difference | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Class | original data | 20th | 30th | 40th | 50th | 60th | 70th | 80th |
| 1 | 50.000 | 301.991 | 176.375 | 2.450 | 33.863 | 170.107 | 398.375 | 692.871 |
| 2 | 4.816 | 0.002 | 0.120 | 1.202 | 1.997 | 2.475 | 2.226 | 1.201 |
| 3 | 5.524 | 0.003 | 0.039 | 0.551 | 0.908 | 1.091 | 0.893 | 0.364 |
| 4 | 6.516 | 0.453 | 0.239 | 0.002 | 0.101 | 0.253 | 0.343 | 0.168 |
| 5 | 4.816 | 0.210 | 0.101 | 0.020 | 0.191 | 0.488 | 0.713 | 0.641 |
| 6 | 2.833 | 0.010 | 0.002 | 0.021 | 0.093 | 0.214 | 0.377 | 0.466 |
| 7 | 4.533 | 0.337 | 0.266 | 0.100 | 0.033 | 0.001 | 0.014 | 0.036 |
| 8 | 2.691 | 0.074 | 0.052 | 0.007 | 0.003 | 0.039 | 0.137 | 0.282 |
| 9 | 2.975 | 0.121 | 0.090 | 0.042 | 0.013 | 0.001 | 0.046 | 0.192 |
| 10 | 3.258 | 0.228 | 0.200 | 0.107 | 0.060 | 0.011 | 0.013 | 0.194 |
| 11 | 1.416 | 0.016 | 0.011 | 0.001 | 0.000 | 0.007 | 0.048 | 0.209 |
| 12 | 1.841 | 0.047 | 0.042 | 0.023 | 0.008 | 0.001 | 0.008 | 0.103 |
| 13 | 0.992 | 0.002 | 0.003 | 0.002 | 0.001 | 0.001 | <0.001 | <0.001 |
| 14 | 1.275 | 0.010 | 0.008 | 0.008 | 0.007 | 0.005 | 0.004 | 0.002 |
| 15 | 1.133 | 0.007 | 0.007 | 0.006 | 0.005 | 0.003 | 0.003 | 0.002 |
| 16 | 2.125 | 0.096 | 0.070 | 0.064 | 0.059 | 0.052 | 0.051 | 0.042 |
| 17 | 0.992 | 0.010 | 0.010 | 0.003 | 0.003 | 0.003 | 0.002 | 0.001 |
| 18 | 0.992 | 0.010 | 0.010 | 0.004 | 0.004 | 0.003 | 0.002 | 0.002 |
| 19 | 0.283 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 20 | 0.992 | 0.010 | 0.010 | 0.005 | 0.001 | 0.031 | 0.159 | 0.589 |
| Mean squared difference: | | 15.981 | 9.350 | 0.243 | 1.966 | 9.199 | 22.412 | 38.743 |

amount of the 20th percentile estimates in class 1 (74.6 percent) and the amount of the original data in class 1 (50 percent). This squared difference was multiplied by 0.5, which is the proportion of the values of the original data in class 1. Because this first class contains 50 percent of the data, any difference observed in this class is more important then one observed in, e.g., class 10, which contains less than 4 percent of the data. We can thus weight each difference by the percentage of original data in that class, and then determine the smallest weighted average squared difference between the actual data and the estimates in order to choose that as the percentile to report. In our example, the choice would be the 40th percentile.

## FINAL POINTS

The indicator approach shows itself to be much less restrictive than traditional approaches such as parametric regression, or geostatistics under the multiGaussian assumption. It makes no assumptions about the underlying shape of the CCDF describing the random variable at any location, and it also allows for the incorporation of secondary data, as in the IKLVM procedure. The benefits of incorporating additional “soft information” into the estimation procedure become readily apparent when examining the resulting final maps from univariate IK and multivariate IKLVM (fig. 8). The amount of detail available in the regression-based map is dramatically higher than that found in the univariate-derived map. This is due to the ability of the technique to account for sharp changes in the landscape over short distances. Univariate IK, on the other hand, assumes that a smooth transition occurs between levels of the primary variable in the intervening spaces between the plots; therefore it fails to take into account the fine-scale features.

In conclusion, the random function model allows us to implement indicator geostatistical methods that can alleviate concerns about non-normal data distributions. The use of the indicator approach also allows us to define optimality criteria for reporting a final estimate or creating a map of an environmental variable. We have found that these approaches, especially IKLVM, are useful tools for modeling forestry data.

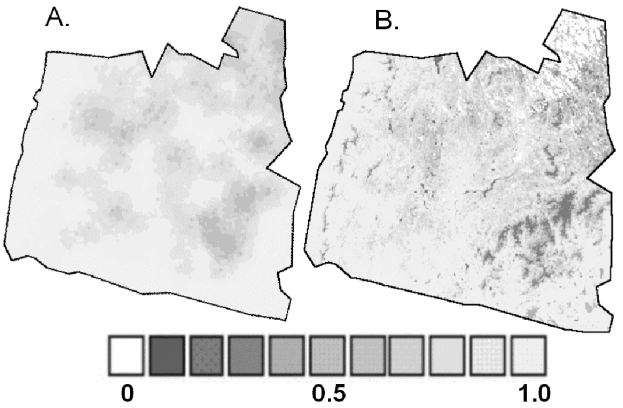


Figure 8—Comparison of the final univariate IK map (A) and the IKLVM multivariate map (B). The IKLVM map reveals much more of the fine scale spatial heterogeneity that exists across the landscape than does the IK map.

## REFERENCES

**Deutsch, C.V.; Journel, A.G.** 1998. GSLIB: Geostatistical software library and user’s guide. 2nd ed. New York: Oxford University Press. 369 p.

**Goovaerts, P.** 1997. Geostatistics for natural resources evaluation. New York: Oxford University Press. 483 p.

**Hunner, G.; Reich, R.M.; Mower, H.T.** 1998. Modeling forest stand structure using spatial statistics. In: Proceedings of the 2nd southern forestry GIS conference; 1998 October 27–29; Athens, GA: 103–120.

**Isaaks, E.; Srivastava, R.M.** 1989. An introduction to applied geostatistics. New York: Oxford University Press. 561 p.

**Iverson, L.R.; Prasad, A.M.; Hale, B.J.; Sutherland, E.K.** 1999. An atlas of current and potential future distributions of common trees of the eastern United States. Gen. Tech. Rep. NE-265. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station. 41 p.

**King, S.L.** 2000. Sequential Gaussian simulation vs. simulated annealing for locating pockets of high-value commercial trees in Pennsylvania. Annals of Operations Research. 95: 177–203.

**Lister, A.; Riemann, R.; Hoppus, M.** 2000. Use of regression and geostatistical techniques to predict tree species distributions at regional scales. 4th international conference on integrating GIS and environmental modeling (GIS/EM4): problems, prospects and research needs. 2000 September 2–8; Banff, AB, Canada.

**Majure, J.; Cressie, N.; Cook, D.; Symanzik, J.** 1996. GIS, spatial statistical graphics, and forest health. Third international conference on integrating GIS and environmental modeling. 1996 January 21–26; Sante Fe, NM: Santa Fe, National Center for Geographic Information and Analysis.

**Metzger, K.L.** 1997. Modeling forest stand structure to a ten meter resolution using Landsat TM data. Fort Collins, CO: Colorado State University, M.S. thesis. 123 p.

**Moeur, M.; Riemann Hershey, R.** 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. In: Lowell, K.; Jaton, A., eds. Spatial accuracy assessment: Land information uncertainty in natural resources. Chelsea, MI: Ann Arbor Press: 419–429.

**Montgomery, D.C.; Peck, E.A.** 1982. An introduction to linear regression analysis. New York: John Wiley. 504 p.

**Myers, D.E.** 1994. Statistical methods for interpolating spatial data. Journal of Applied Science and Computers. 1(2): 283–318.

**Riemann Hershey, R; Ramirez, M.A.; Drake, D.A.** 1997. Using geostatistical techniques to map the distribution of tree species from ground inventory data. In: Gregoire, T.; Brillinger, D.R.; Diggle, P.J., eds. Modeling longitudinal and spatially correlated data: methods, applications, and future directions. Lecture Notes in Statistics 122. New York: Springer Verlag: 187–198.

**Zhu, Z.** 1994. Forest density mapping in the lower 48 States: a regression procedure. U.S. Department of Agriculture, Forest Service. Res. Pap. SO–280. New Orleans, LA: Southeastern Forest Experiment Station. 11 p.

# Statistical Estimation and Modeling

*Moderator:*

**RON McROBERTS**
USDA Forest Service
Southern Research Station

# AREAL CONTROL USING GENERALIZED LEAST SQUARES AS AN ALTERNATIVE TO STRATIFICATION[1]

**Raymond L. Czaplewski**[2]

**Abstract**—Stratification for both variance reduction and areal control proliferates the number of strata, which causes small sample sizes in many strata. This might compromise statistical efficiency. Generalized least squares can, in principle, replace stratification for areal control.

## INTRODUCTION

Stratification with remotely sensed forest map improves precision of FIA estimates. However, FIA also uses stratification to assure that area estimates equal "official" acres in each county, as published by the U.S. Census Bureau. I term this latter function "areal control."

Cross-stratification by both remotely sensed strata and geopolitical boundaries proliferates the number of strata. Hence, sample sizes become small in many of these cross-classified strata. There is concern within FIA that these small sample sizes can degrade the statistical efficiency gained through stratification. This concern is heightened by the move to an annual FIA system, in which only 10 percent to 15 percent of the FIA field plots are remeasured each year.

I consider the use of stratification solely for variance reduction to avoid proliferation of strata. I present an alternative to stratification to areal control that constrains FIA estimates such that they agree with county acreages from the Census Bureau. I use a simple example of two strata (forest and nonforest) and two counties.

## PROBLEM FORMULATION

Let the 4x1 vector **z** represent the true area of forest and nonforest in each county. Equation (1) denotes the vector estimate of these areas, including the 4x4 covariance matrix $\mathbf{V}_z$ for estimation errors.

$$\hat{\mathbf{z}}=\begin{bmatrix}\hat{z}_1=\begin{array}{l}\text{Estimate of forest in county A}\\ \text{using remotely sensed strata}\end{array}\\ \hdashline \hat{z}_2=\begin{array}{l}\text{Estimate of nonforest in county}\\ \text{using remotely sensed strata}\end{array}\\ \hdashline \hat{z}_3=\begin{array}{l}\text{Estimate of forest in county B}\\ \text{using remotely sensed strata}\end{array}\\ \hdashline \hat{z}_4=\begin{array}{l}\text{Estimate of forest in county E}\\ \text{using remotely sensed strata}\end{array}\end{bmatrix} \qquad (1)$$

where

$$\mathbf{z} = \hat{\mathbf{z}} + \mathbf{u}_z$$
$$E\left[\mathbf{u}_z\mathbf{u}_z'\right] = \mathbf{V}_z$$

The estimates in equation (1) assume that remotely sensed data have already been used to separate Phase 2 plots into forest and nonforest strata, and the appropriate estimator is used to reduce variance through this stratification. However, the estimates in equation (1) are not stratified by Census Bureau county statistics for areal control. The following describes the alternative to stratification for areal control, in which the sample estimate is constrained so that summations of areal estimates for each county exactly equal the "official" acres in each county, as published by the U.S. Census Bureau.

Let the 2x1 vector **c** contain the "official" acres in each county. A sample estimate of **c** is available from a simple linear transformation of the vector estimate **z** from equation (1).

$$\hat{\mathbf{c}} = \begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\end{bmatrix}\begin{bmatrix}\hat{z}_1\\ \hat{z}_2\\ \hat{z}_3\\ \hat{z}_4\end{bmatrix} = \mathbf{H}\hat{\mathbf{z}} \qquad (2)$$

where

$$\mathbf{c} = \hat{\mathbf{c}} + \mathbf{u}_c$$
$$E\left[\mathbf{u}_c\mathbf{u}_c'\right] = \mathbf{H}\mathbf{V}_z\mathbf{H}'$$

In addition, the exact areas for each county are available from the Census Bureau.

$$\mathbf{c}=\begin{bmatrix}c_1=\{\text{Census Bureau area in county A}\}\\ \hdashline c_2=\{\text{Census Bureau area in county B}\}\end{bmatrix} \qquad (3)$$

**c** in equation (3) is a vector of constants, not an estimate, because the county acreages from the Census Bureau are known without error. Our objective is to constrain the vector estimate **z** in equation (1) such that the vector estimate **c** of county acreages in equation (2) agrees exactly with the Census Bureau statistics in equation (3).

## GENERALIZED LEAST SQUARES (GLS)

Let vector **B** represent the estimates of forest area that are constrained to agree with official statistics.

[1] Paper presented at the Second Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Project Leader, USDA Forest Service, Rocky Mountain Research Station, 2150 Centre Ave., Bldg. A, Suite 350, Fort Collins, CO 80526-1891.

$$\mathbf{B} = \left[\begin{array}{c} \beta_1 \\ \hdashline \beta_2 \\ \hdashline \beta_3 \\ \hdashline \beta_4 \end{array}\right] = \left[\begin{array}{l} \beta_1 = \texttt{acres of forest cover in county A} \\ \hdashline \beta_2 = \texttt{acres of non-forest cover in county A} \\ \hdashline \beta_3 = \texttt{acres of forest cover in county B} \\ \hdashline \beta_4 = \texttt{acres of non-forest cover in county A} \end{array}\right] \quad (4)$$

such that

$\beta_1 + \beta_2 =$ Exact Census Bureau acres in county A
$\beta_1 + \beta_2 =$ Exact Census Bureau acres in county B

Estimation of the **B** vector is the final goal.

The GLS solution estimates 4x1 vector **B** in Equation (4) from the 6x1 vector **y,** which is a concatenation of vector estimate **z** from equation (1), and a vector of independent ancillary estimates **k** (with covariance matrix $\mathbf{V}_k$), which will later contain the areal control statistics.

$$\hat{\mathbf{y}} = \left[\begin{array}{c} \hat{\mathbf{z}} \\ \hline \hat{\mathbf{k}} \end{array}\right] = \left[\begin{array}{c} \hat{z}_1 \\ \hdashline \hat{z}_2 \\ \hdashline \hat{z}_3 \\ \hdashline \hat{z}_4 \\ \hline \hat{k}_1 \\ \hdashline \hat{k}_2 \end{array}\right] \quad (5)$$

Define the linear model:

$$\mathbf{y} = \mathbf{XB} + \mathbf{u}_y \quad (6)$$

where

$$\mathbf{X} = \left[\begin{array}{c:c:c:c} 1 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 0 \\ \hdashline 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 1 \end{array}\right] = \left[\begin{array}{c} \mathbf{I} \\ \hline \mathbf{H} \end{array}\right]$$

$$\mathbf{V}_y = E\left[\mathbf{u}_y \mathbf{u}_y'\right] = \left[\begin{array}{c:c} \mathbf{V}_z & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{V}_k \end{array}\right]$$

The zero off-diagonal sub-matrices within the covariance matrix $\mathbf{V}_y$ in equation (6) correspond to the presumed independence between vectors **z** and **k** in equation (5).

The GLS estimator for **B** is defined as:

$$\hat{\mathbf{B}} = \left[\mathbf{X}'\left(\hat{\mathbf{V}}_y\right)^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\left(\hat{\mathbf{V}}_y\right)^{-1}\hat{\mathbf{y}} \quad (7)$$

Let the matrices in equation (7) be partitioned as follows:

$$\hat{\mathbf{B}} = \left[\left(\begin{array}{c} \mathbf{I} \\ \hline \mathbf{H} \end{array}\right)'\left(\begin{array}{c:c} \hat{\mathbf{V}}_z & \mathbf{0} \\ \hdashline \mathbf{0} & \hat{\mathbf{V}}_k \end{array}\right)^{-1}\left(\begin{array}{c} \mathbf{I} \\ \hline \mathbf{H} \end{array}\right)\right]^{-1}\left(\begin{array}{c} \mathbf{I} \\ \hline \mathbf{H} \end{array}\right)'\left(\begin{array}{c:c} \hat{\mathbf{V}}_z & \mathbf{0} \\ \hdashline \mathbf{0} & \hat{\mathbf{V}}_k \end{array}\right)^{-1}\left[\begin{array}{c} \hat{\mathbf{z}} \\ \hline \mathbf{k} \end{array}\right] \quad (8)$$

Using matrix algebra for partitioned matrices, equation may be rewritten as:

$$\begin{aligned} \hat{\mathbf{B}} &= \left[\left(\mathbf{I} \quad \mathbf{H}'\right)\left(\begin{array}{c:c} \hat{\mathbf{V}}_z^{-1} & \mathbf{0} \\ \hdashline \mathbf{0} & \hat{\mathbf{V}}_k^{-1} \end{array}\right)\left(\begin{array}{c} \mathbf{I} \\ \hline \mathbf{H} \end{array}\right)\right]^{-1}\left(\mathbf{I} \quad \mathbf{H}'\right)\left(\begin{array}{c:c} \hat{\mathbf{V}}_z^{-1} & \mathbf{0} \\ \hdashline \mathbf{0} & \hat{\mathbf{V}}_k^{-1} \end{array}\right)\left[\begin{array}{c} \hat{\mathbf{z}} \\ \hline \mathbf{k} \end{array}\right] \\ &= \left[\left(\mathbf{I} \quad \mathbf{H}'\right)\begin{pmatrix} \hat{\mathbf{V}}_z^{-1} \\ \hat{\mathbf{V}}_k^{-1}\mathbf{H} \end{pmatrix}\right]^{-1}\left(\mathbf{I} \quad \mathbf{H}'\right)\begin{pmatrix} \hat{\mathbf{V}}_z^{-1}\hat{\mathbf{z}} \\ \hat{\mathbf{V}}_k^{-1}\mathbf{k} \end{pmatrix} \\ &= \left[\hat{\mathbf{V}}_z^{-1} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1}\mathbf{H}\right]^{-1}\left[\hat{\mathbf{V}}_z^{-1}\hat{\mathbf{z}} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1}\mathbf{k}\right] \end{aligned} \quad (9)$$

Maybeck (1979, pp. 234-235) shows that the result in equation (9) is a maximum likelihood estimate under appropriate assumptions. Maybeck (p. 214) uses the following matrix inversion lemma to rewrite the first term in equation (9) into a numerically superior form:

$$\begin{aligned} \left[\hat{\mathbf{V}}_z^{-1} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1}\mathbf{H}\right]^{-1} &= \hat{\mathbf{V}}_z - \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}' + \hat{\mathbf{V}}_k\right]^{-1}\mathbf{H}\hat{\mathbf{V}}_z \\ &= \hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z \\ \text{where } \mathbf{G} &= \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}' + \hat{\mathbf{V}}_k\right]^{-1} \end{aligned} \quad (10)$$

The matrix inversion lemma in equation (10) reduces the dimensions of the matrices to improve numerical performance. The left-hand term in equation (10) involves inversion of a 4x4 matrix in my simple example, while the right-hand term involves inversion of a 2x2 matrix. When the number of areal controls is larger, say the Census Bureau area for each of 50 counties, and there are 2 remotely sensed strata, say forest and nonforest, then the left-hand term requires inversion of a 100x100 covariance matrix, while the right-hand term inverts a 50x50 matrix. Maybeck (1979, pp. 214-217) uses this lemma, then expands, regroups, and exploits algebraic identities to rewrite equation (9) as:

$$\begin{aligned} \hat{\mathbf{B}} &= \left[\hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z\right]\left[\hat{\mathbf{V}}_z^{-1}\hat{\mathbf{z}} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1}\mathbf{k}\right] \\ &= \hat{\mathbf{z}} + \mathbf{G}\left[\mathbf{k} - \mathbf{H}\hat{\mathbf{z}}\right] \end{aligned} \quad (11)$$

Equation (11) may be expressed in an equivalent “Joseph” form (Maybeck p. 237) as:

$$\hat{\mathbf{B}} = \left[\mathbf{I} - \mathbf{GH}\right]\hat{\mathbf{z}} + \mathbf{Gk} \quad (12)$$

Maybeck (pp. 215-216) derives the covariance matrix for the estimate in equation (11) as:

$$\mathbf{V}_{\hat{\mathbf{B}}} = \hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z \quad (13)$$

An alternative form for equation (13) is:

$$\mathbf{V}_{\hat{\mathbf{B}}} = \left[\mathbf{I} - \mathbf{GH}\right]\mathbf{V}_z\left[\mathbf{I} - \mathbf{GH}\right]' + \mathbf{G}\hat{\mathbf{V}}_k\mathbf{G} \quad (14)$$

While equation (14) is more complex than equation (13), the “Joseph” form of the covariance matrix in equation (14) has numerical advantages (Maybeck 1979, p. 237). However, equation (14) can still have numerical problems with illconditioned covariance matrices. Maybeck (pp. 368-405) develops equivalent estimation equations that improve numerical precision and stability by using square roots of the covariance matrices. Maybeck (1979) recommends diagnostics that test the distribution of multivariate residuals, or "innovations."

## APPLICATION OF GLS TO AREAL CONTROL IN FIA

Let the vector of independent ancillary estimates **k** in equation (5) contain the areal control statistics from the Census Bureau, i.e., **k**=**c** from equation (3). The asso-ciated covariance matrix in equations (10) through (14) equals the zero matrix, i.e., $\mathbf{V}_k=\mathbf{0}$, because **c** is a vector of constants. In this case, the usual form of the GLS estimator in equation (8) will not work because it requires the infeasible inverse of a singular covariance matrices. However, the equivalent estimators in equations (11) and (12) are feasible because the weighting matrix **G** in equation (10) simply equals:

$$\begin{aligned} \mathbf{G} &= \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}' + \mathbf{0}\right]^{-1} \\ &= \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}'\right]^{-1} \end{aligned} \qquad (15)$$

The matrix inversion lemma in equation (10) assumes that $\mathbf{V}_z$ and $\mathbf{V}_k$ are positive definite covariance matrices (Maybeck 1979, p. 213). This is obviously untrue when $\mathbf{V}_z=\mathbf{0}$ for equation (15), even though equation (15) is feasible. However, this violation of assumptions has never caused any problems for my numerical applications, and an alternative derivation should exist that does not need to assume that $\mathbf{V}_k$ is positive definite. In addition, **y** is a vector of proportions that sum to exactly one, and its covariance matrix $\mathbf{V}_z$ in equation (1) is not positive definite; neither is $\mathbf{HV}_z\mathbf{H'}$ positive definite, which means that its inverse in equation (15) does not exist. This latter problem can be solved by deleting one row in **c** and $\mathbf{HV}_z\mathbf{H'}$, and the corresponding row and column in covariance matrix $\mathbf{HV}_z\mathbf{H'}$, or using the pseudo-inverse.

FIA requires estimators that can be implemented within database structures. This typically requires that statistical estimators be implemented as expansion factors in the database. This criterion cannot be met exactly with the expressions in equations (11) or (12). Equation (11) might be best suited for application in a database. The database retains the expansion factors originally used to estimate vector **z,** which contains the statistics in FIA core tables for area. However, each element of **z**, and each corresponding cell in FIA core tables, must be "adjusted" by the associated element in the vector **G**[**c**-**Hz**], which contains positive and negative elements centered on zero. One way to implement these adjustment terms is to add a "record" to the FIA database that contains the estimated population totals for area in the FIA Inventory Unit (i.e., **z**).

## DISCUSSION

I show that there is at least one alternative to stratification for areal control. This alternative might solve problems with insufficient sample sizes in small cross-classified strata, especially as FIA shifts to an annualized system that measures 20 percent or less of its Phase Two field plots each year. In effect, the empirical relationship between remote sensing and field data is established using data across the entire multi-county region to avoid problems with small sample sizes in some strata. This is already captured in equation (1). Then, a linear multivariate estimator in equation (11) or (12) applies areal controls so that final estimates are constrained to agree with "official statistics" (e.g., Bureau of Census).

This same approach can be used to constrain FIA estimates to other "official statistics," such as the published area for each national forest, ranger district, park, or preserve. If it is not necessary to make FIA estimates agree with cross-classifications of administrative entities (e.g., area of each county on a national forest), then the areal controls can be applied sequentially through the estimators given above.

Because the number of cells in FIA core tables are so numerous, and the GLS estimators require inversion of covariance matrices that correspond to cells in FIA statistical tables, the proposed solution might only work for a few broad indicators (e.g., total forest area), which are typically margins in FIA core tables. The proposed solution might require ad hoc or special methods to implement in FIA information management system.

I plan to test the practical value of these estimators using Monte Carlo simulations.

## REFERENCES

**Maybeck, P.S.** 1979. Stochastic models, estimation, and control. New York: Academic Press. 423 p. Vol. 1.

# UPDATING INDIANA ANNUAL FOREST INVENTORY AND ANALYSIS PLOT DATA USING EASTERN BROADLEAF FOREST DIAMETER GROWTH MODELS[1 2]

**Veronica C. Lessard**[3]

**Abstract**—The Forest Inventory and Analysis (FIA) program of the North Central Research Station (NCRS), USDA Forest Service, has developed nonlinear, individual-tree, distance-independent annual diameter growth models. The models are calibrated for species groups and formulated as the product of an average diameter growth component and a modifier component. The regional models for the Eastern Broadleaf (Continental)—Province 222, defined by R.G. Bailey, are calibrated using periodic Forest Inventory and Analysis data within that ecoregion. Average annual diameter growth is the dependent variable. The independent variables include crown ratio, crown class, stand basal area larger than the subject tree, physiographic class, latitude, and longitude. North Central Forest Inventory and Analysis has begun implementing annual inventories in a number of states of the region. The diameter growth models have been applied to annual data from Indiana to test their effectiveness on an independent data set that was collected under a survey design that differs from the one on which the models were calibrated. The bias of estimates increased with increasing values of crown ratio and with decreasing values of crown class.

## INTRODUCTION

The North Central Research Station has developed nonlinear, individual-tree, distance-independent annual diameter growth models formulated as the product of an average diameter growth component and a modifier component and calibrated for species groups (Lessard and others submitted). The models may be used in two ways: (1) to update information on FIA plots not visited in the current year as a method of eliminating any lag in estimates of current conditions; and (2) to predict future forest resources.

Regional diameter growth models were developed for species groups within two ecoregions, the Laurentian Mixed Forest and the Eastern Broadleaf Forest (Lessard and McRoberts, in preparation). The objective of this study is to apply the Eastern Broadleaf Forest regional models to an independent data set, Indiana annual FIA data, and analyze the prediction performance.

## PROVINCE 222

The diameter growth models are calibrated on FIA data from the Eastern Broadleaf Forest (Continental)—Province 222, defined by Bailey (1995). Province 222 is a subdivision of the Hot Continental Division. Most precipitation in Province 222 occurs during the growing season and generally decreases in quantity as distance from the Atlantic Ocean increases. This province favors drought-resistant oak-hickory associations. Province 222 lies to the east of the prairie regions, south and west of the Laurentian Mixed Forest—Province 212 in the northern areas, and west of the Appalachian Mountains in the southern regions. It extends from the Minnesota/Canadian border in the north through Missouri and Tennessee in the south.

## CALIBRATION DATA

The diameter growth models were calibrated using FIA data across all ownership categories on land classified as timberland. Timberland was defined as non-reserved forestland that is producing or is capable of producing 20 ft$^2$/ac/yr of industrial wood. The FIA periodic 10-point cluster survey design and the data collection were described by Hansen and others (1992). ArcView GIS was used to overlay Bailey's eco-region map (Bailey and others 1994) on the FIA plot locations to select plots within Province 222. Growth models were calibrated using FIA data from the following states (the parentheses refer to the year of the inventory): Michigan (1980, 1993), Wisconsin (1983, 1996), Minnesota (1990, 1993), Illinois (1985, 1998), Indiana (1986, 1998), Iowa (1974, 1990), Ohio (1978, 1990), Missouri (1972, 1989), Kentucky (1974, 1987), and Tennessee (1989, 1996).

## INDIANA ANNUAL DATA

Data from both the old periodic 10-point cluster design and the new 4-point annual design plots were collected during the last periodic inventory in Indiana (1998). The new standard plot design is a cluster of four fixed-area subplots (24-foot radius) superimposed on four fixed-area micro-plots (6.8-ft radius). All trees 5.0 in dbh and larger are measured on the subplots and all trees 1.0-4.9 in dbh are measured on the micro-plots. Under the annual system, plots to be measured in each cycle are divided into five sub-cycles. Each sub-cycle is inventoried in a single year to complete the full inventory cycle in five years.

Two measurement intervals were included in the Indiana annual data set: (1) 1,358 trees (69 plots) in the 1998-1999 data (the last periodic, cycle 4 to cycle 5, sub-cycle 1); and (2) 1,503 trees (63 plots) in the 1998-2000 data (the last periodic, cycle 4 to cycle 5, sub-cycle 2).

[1] Paper presented at the Second Annual Forest Inventory and Analyis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] This study was partly funded by the USDA Forest Service, NCRS.

[3] Statistician, Natural Resources Inventory and Analysis Institute, 1992 Folwell Avenue, St. Paul, MN 55108 (formerly Research Forester, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108).

## MODEL FORM

The form of the diameter growth model is:

$$E(\Delta DBH) = AVERAGE * MODIFIER \text{ , where} \quad (1a)$$

$$AVERAGE = \beta_1 \exp(-\beta_2 DBH) DBH^{\beta_3} \text{ , and} \quad (1b)$$

$$\begin{aligned} MODIFIER = \; & \exp[\beta_4 (CR - 4) + \beta_5 (BAL - 50) + \beta_7 (CC - 3) + \\ & \beta_8 (PC - 5) + \beta_{10} (LNG - 89) + \beta_{11} ((LNG - 89)^2)/10 + \\ & \beta_{12} (LAT - 40) + \beta_{13} ((LAT - 40)^2)/10], \end{aligned} \quad (1c)$$

where DBH is diameter at breast height, CR is crown ratio (percent of tree height consisting of crown), CC is crown class in five categories ranging from dominant to suppressed, BAL is plot basal area per acre for trees larger than that of the subject tree, PC is physiographic class coded in the data set as 3, 4, 5, 6, or 7 (corresponding with xeric, xeromesic, mesic, hydromesic, and hydric, respectively), LNG is longitude, and LAT is latitude. The modeling methodology and assessment of fit are documented in Lessard and others (in review).

## BIAS ASSESSMENTS FOR MODELS APPLIED TO INDIANA ANNUAL DATA

Residual analysis was conducted to examine the adequacy of the diameter growth model predictions for an independent data set. The Eastern Broadleaf Forest regional models were applied to the FIA Annual Indiana data to obtain predicted diameter growth rates (in/yr) for individual trees. Predicted growth rates were compared to average annual observed change in DBH, calculated as the ratio of the difference in DBH at the two measurements and the number of years in the measurement interval. Residuals were calculated as differences between observed and predicted annual changes in diameter. To examine how well the models fit the data, percentile statistics (25th, 50th, 75th) were computed for the residuals by species group, and by classes of DBH, CR, BAL, CC, longitude, and latitude. Models were judged to be unbiased if zero was included in the range of values between the 25th and 75th percentiles for the residuals.

## RESULTS AND CONCLUSIONS

The models were generally unbiased when the residuals were examined by species group (table 1), and by classes of DBH (table 2), CR (table 3), BAL (table 4), CC (table 5), LNG (table 6), and LAT (table 7). Several exceptions did occur. The models overestimated growth rates for cottonwood, however there were only 17 cottonwood trees in the data set used to test the models. Median residuals generally increased with increasing CR classes and with decreasing CC sizes. However, the inter-quartile range of residual values included zero for all but the largest and smallest classes of CR (table 3) and for the smallest CC with only 3 observations (table 5).

To examine the trend found in the residuals with respect to CR more closely, median average annual growth rates were calculated by CR class for both the Indiana Annual data and the Eastern Broadleaf Forest Periodic data and compared Indiana Annual data growth rates were less than those of the calibration data for small CC values and greater for large CC

**Table 1—Analysis of residuals (calculated as the observed minus predicted values) sorted by species group**

| Species group | No. of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| | | *In/yr* | *In/yr* | *In/yr* |
| Softwoods | | | | |
| Eastern white pine | 36 | -0.117 | -0.031 | -0.003 |
| Red pine | 22 | -0.068 | -0.048 | 0.012 |
| Jack pine and Virginia pine | 93 | -0.022 | 0.037 | 0.099 |
| Shortleaf pine | 55 | -0.017 | 0.022 | 0.059 |
| Tamarack | 5 | -0.011 | 0.016 | 0.017 |
| Eastern redcedar | 100 | -0.026 | 0.021 | 0.111 |
| Other softwoods | 9 | -0.061 | -0.031 | 0.000 |
| Hardwoods | | | | |
| Select white oak | 158 | -0.065 | -0.027 | 0.018 |
| Other white oak | 32 | -0.063 | -0.031 | 0.046 |
| Northern red oak | 48 | -0.083 | -0.011 | 0.037 |
| Other red oak | 163 | -0.050 | 0.004 | 0.087 |
| Select hickory | 72 | -0.061 | -0.009 | 0.046 |
| Other hickory | 160 | -0.046 | -0.004 | 0.042 |
| Hard maple | 272 | -0.043 | -0.003 | 0.050 |
| Soft maple | 217 | -0.079 | -0.024 | 0.056 |
| Boxelder | 23 | -0.104 | -0.005 | 0.137 |
| American beech | 40 | -0.043 | -0.006 | 0.050 |
| White and green ash | 144 | -0.073 | -0.007 | 0.061 |
| Black ash | 7 | -0.102 | -0.052 | 0.014 |
| Aspen | 14 | -0.091 | 0.038 | 0.122 |
| Cottonwood | 17 | -0.182 | -0.097 | -0.035 |
| American basswood | 39 | -0.018 | 0.013 | 0.069 |
| Butternut and walnut | 77 | -0.057 | -0.014 | 0.053 |
| Black cherry | 153 | -0.084 | -0.014 | 0.075 |
| Elm | 168 | -0.060 | -0.005 | 0.060 |
| Hackberry | 28 | -0.056 | -0.015 | 0.089 |
| Sycamore | 27 | -0.110 | -0.040 | 0.014 |
| Yellow-poplar | 117 | -0.067 | 0.058 | 0.246 |
| Sweetgum | 35 | -0.026 | 0.032 | 0.113 |
| Tupelo | 40 | -0.046 | 0.002 | 0.071 |
| Sassafras | 125 | -0.059 | -0.029 | 0.006 |
| Flowering dogwood | 29 | -0.029 | -0.007 | 0.017 |
| Other commercial hardwoods | 68 | -0.089 | -0.025 | 0.064 |
| Noncommercial hardwoods | 69 | -0.033 | 0.010 | 0.066 |

**Table 2—Analysis of residuals (calculated as the observed minus predicted values) sorted by 5-inch dbh class**

| DBH class | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| *In* | | *In/yr* | *In/yr* | *In/yr* |
| 1-5 | 228 | 0.061 | 0.004 | -0.041 |
| 5-10 | 1,502 | 0.064 | -0.006 | -0.055 |
| 10-15 | 606 | 0.064 | -0.002 | -0.054 |
| 15-20 | 213 | 0.032 | -0.030 | -0.082 |
| 20-25 | 83 | 0.082 | -0.023 | -0.096 |
| 25+ | 30 | 0.087 | -0.006 | -0.089 |

**Table 3—Analysis of residuals (calculated as the observed minus predicted values) sorted by crown ratio**

| Crown ratio class | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| *Percent* | | *In/yr* | *In/yr* | *In/yr* |
| 0-9 | 33 | -0.047 | -0.073 | -0.103 |
| 10-19 | 288 | 0.011 | -0.032 | -0.067 |
| 20-29 | 697 | 0.031 | -0.022 | -0.066 |
| 30-39 | 771 | 0.065 | -0.002 | -0.055 |
| 40-49 | 470 | 0.071 | -0.001 | -0.052 |
| 50-59 | 219 | 0.113 | 0.025 | -0.034 |
| 60-69 | 104 | 0.159 | 0.065 | 0.004 |
| 70-79 | 55 | 0.231 | 0.064 | -0.034 |
| 80-99 | 25 | 0.150 | 0.130 | 0.049 |

**Table 4—Analysis of residuals (calculated as the observed minus predicted values) sorted by BAL**

| BAL class | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| *$\mathrm{Ft}^2$/ac* | | *In/yr* | *In/yr* | *In/yr* |
| 0-50 | 989 | -0.069 | -0.007 | 0.076 |
| 51-100 | 804 | -0.064 | -0.021 | 0.043 |
| 101-150 | 376 | -0.047 | -0.003 | 0.055 |
| 151-200 | 142 | -0.035 | 0.005 | 0.048 |
| 201-250 | 97 | -0.051 | -0.009 | 0.019 |
| 251-300 | 132 | -0.035 | 0.018 | 0.065 |
| 301-350 | 109 | -0.010 | 0.026 | 0.075 |
| 351-400 | 13 | 0.021 | 0.246 | 0.246 |

**Table 5—Analysis of residuals (calculated as the observed minus predicted values) sorted by CC**

| BAL class | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| | | *In/yr* | *In/yr* | *In/yr* |
| 1 | 3 | 0.286 | 0.185 | 0.147 |
| 2 | 79 | 0.154 | 0.052 | -0.011 |
| 3 | 1,457 | 0.074 | 0.004 | -0.057 |
| 4 | 626 | 0.048 | -0.017 | -0.060 |
| 5 | 497 | 0.027 | -0.020 | -0.052 |

**Table 6—Analysis of residuals (calculated as the observed minus predicted values) sorted by longitude**

| Longitude | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| *Degrees* | | *In/yr* | *In/yr* | *In/yr* |
| -87.55 | 230 | 0.076 | 0.004 | -0.051 |
| -87.05 | 364 | 0.066 | -0.005 | -0.063 |
| -86.55 | 1,066 | 0.041 | -0.013 | -0.057 |
| -86.05 | 301 | 0.085 | 0.012 | -0.043 |
| -85.55 | 405 | 0.080 | -0.005 | -0.058 |
| -85.05 | 296 | 0.072 | 0.005 | -0.067 |

**Table 7—Analysis of residuals (calculated as the observed minus predicted values) sorted by latitude**

| Latitude | Number of trees | Percentile | | |
|---|---|---|---|---|
| | | 25th | 50th | 75th |
| *Degrees* | | *In/yr* | *In/yr* | *In/yr* |
| 38.05 | 475 | 0.069 | 0.015 | -0.034 |
| 38.55 | 521 | 0.034 | -0.014 | -0.053 |
| 39.05 | 461 | 0.057 | -0.017 | -0.064 |
| 39.55 | 302 | 0.085 | 0.008 | -0.049 |
| 40.05 | 92 | 0.129 | 0.019 | -0.057 |
| 40.55 | 124 | 0.092 | -0.001 | -0.083 |
| 41.05 | 441 | 0.043 | -0.021 | -0.064 |
| 41.55 | 246 | 0.074 | -0.003 | -0.071 |

values. This follows the underestimation and overestimation patterns of the residuals (table 3).

In the context of their intended applications, the annual diameter growth models may be considered generally unbiased. However, because diameter growth rates with respect to CR tended to change from the time during which the calibration data were collected to the time the annual data was collected, exploration of methodology to capture these changes may improve the diameter growth predictions. Inclusion of climate variables in the model or application of model updating (e.g. Bayes) may improve the quality of diameter growth predictions.

## REFERENCES

**Bailey, R.G.** 1995. Description of the ecoregions of the United States. 2nd ed. Revised and expanded (1st ed. 1980). Misc. Pub. No. 1391 (rev.), Washington, DC: USDA Forest Service. 108 p. [with separate map at 1:7,500,000.]

**Hansen, M.H.; Frieswyk,T.; Glover, J.F.; Kelly, J.F.** 1992. The Eastwide forest inventory data base: users manual. USDA For. Serv. Gen. Tech. Rep. NC–GTR–151. 48 p.

**Lessard, V.C.; McRoberts, R.E.; Holdaway, M.R.** Diameter growth models using Minnesota Forest Inventory and Analysis data. Forest Science (in review).

**Lessard, V.C.; McRoberts, R.E.** [In preparation]. Diameter growth models for the Laurentian mixed forest province. USDA For. Serv. Gen. Tech. Rep.

**Lessard, V.C.; McRoberts, R.E.** [In preparation]. Diameter growth models for the easterm broadleaf forest province. USDA For. Serv. Gen. Tech. Rep.

# ESTIMATING THE UNCERTAINTY IN DIAMETER GROWTH MODEL PREDICTIONS AND ITS EFFECTS ON THE UNCERTAINTY OF ANNUAL INVENTORY ESTIMATES[1]

**Ronald E. McRoberts and Veronica C. Lessard[2]**

**Abstract**—Uncertainty in diameter growth predictions is attributed to three general sources: measurement error or sampling variability in predictor variables, parameter covariances, and residual or unexplained variation around model expectations. Using measurement error and sampling variability distributions obtained from the literature and Monte Carlo simulation methods, the uncertainty in 10-year diameter growth model predictions is estimated as are its effects on annual basal area estimates obtained using an annual inventory system. The results indicate that although annual diameter growth is difficult to predict precisely, the effects of the uncertainty in the growth predictions are greatly attenuated when diameter estimates are aggregated to estimate plot basal area and mean basal area over all plots.

## INTRODUCTION

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service has initiated an annual forest inventory system featuring measurement of a proportion of plots each year (McRoberts 1999). One approach to obtaining annual inventory estimates with this system is to use growth models to update to the current year data for plots measured in previous years and then base estimates on the data for all plots. If the updating procedure is sufficiently unbiased and precise, this approach provides nearly the same precision as using all plots but without the adverse effects of using out-of-date information. With this estimation approach in mind, a prototype set of individual tree, diameter at breast height (DBH) (1.37 m aboveground) growth models has been constructed and calibrated for use in updating FIA plot information.

The data used to calibrate the models were taken from measurements of forested Minnesota FIA plots for the 1977 (Spencer 1982) and 1990 periodic inventories (Miles et al 1995). Only trees alive and measured in both inventories were used. For each tree, average annual DBH growth was used as a surrogate for annual growth and was calculated as the ratio of the difference in DBH measurements for the two inventories and the number of growing seasons between measurements. Predictor variables were average DBH for the measurement interval, initial crown ratio (CR), initial crown class (CC), average plot basal area (BA), average plot basal area in trees larger than the subject tree (BAL), and physiographic class (PC). BA and BAL represent the sum of cross-sectional areas of live tree boles at breast height, and, unless otherwise noted, references to both BA and BAL are assumed to have been scaled to a per unit area basis.

The DBH growth models consist of the product of two components, an average component corresponding to regional average DBH growth with respect to DBH and a modifier component that adjusts DBH growth predictions in accordance with local plot and tree conditions. The average component is based on a 2-parameter gamma function with a constant multiplier and uses DBH as the predictor variable, while the modifier component consists of the product of exponential factors of which each incorporates a single predictor variable. Each factor in the modifier product expresses a multiplicative effect on growth predictions in terms of departures from regional or ecosystem averages for a single predictor variable. The general form of the DBH growth model is

$$E(\triangle DBH) = Ave(DBH)*Mod(CR,CC,BA,BAL,PC) \qquad [3a]$$

where E(.) denotes statistical expectation, ΔDBH is annual DBH growth,

$$Ave(DBH) = ß_1\ DBHß_2\ \exp(ß_3\ DBH) \qquad [3b]$$

and

$$\begin{aligned} &Mod(CR,CC,BA,BAL,PC) \\ &\quad = \exp[ß_4(CR-C_4)+ß_5(CC-C_5)+ß_6(BAL-C_6) \\ &\quad + ß_7(BA-C_7)+ß_8(PC-C_8)+ß_9(PC-C_8)^2], \end{aligned} \qquad [3c]$$

where the ßs are parameters to be estimated and the Cs are constants representing regional or ecosystem averages for the corresponding predictor variables. Using iteratively reweighted least squares techniques, the model was fit separately for individual species. If a parameter was not statistically significantly different than zero, its estimate was fixed at zero. Lessard and others (submitted) provide details of the fitting procedure and verification and validation of the models.

## THE ANNUALIZED INVENTORY DATABASE

An annualized database of plot and tree variables was constructed to evaluate the models. The database included measurements from forested FIA plots for both the 1977 (Spencer 1982) and the 1990 (Miles and others 1995) USDA Forest Service periodic inventories of Minnesota. Plots included in the 1977 inventory were actually measured between 1974 and 1978, while plots included in the 1990 inventory were actually measured between 1986 and 1991. Because additional investigations were necessary to

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Mathematical Statistician and Research Forester, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108, respectively.

estimate the uncertainty in model parameters estimated for each species, the data were further restricted to plots that included only the 15 most common tree species in Minnesota, i.e., if the species of any tree on a plot was not among the 15 most common species, the data for that plot was excluded from the database. The resulting database included data for 38,156 trees on 1,951 plots.

Plots measured for the 1977 and 1990 Minnesota inventories consisted of 10 subplots of which each is described as a variable radius plot due to the use of point sampling techniques. With these techniques, trees are selected with probability proportional to cross-sectional area rather than proportional to the frequency of occurrence in the population (Myers and Beers 1971). With point sampling, the number of trees in the population represented by a sample tree, termed the tree factor, varies by tree and is calculated as a scaling constant divided by the tree DBH. Tree factors are used to expand the measurements of sample trees to per unit area estimates.

Based on observations of individual trees with DBHs of at least 12.7 cm, an 11-year database was constructed that consisted of annual values for all model predictor variables and annual status with respect to survival, ongrowth, mortality, and harvest for each tree. Construction of the database required distributing total growth between inventories over varying numbers of years for individual trees in each of four categories: (1) survivor trees that were alive and measured at both inventories; (2) ongrowth trees that attained the 12.7-cm minimum DBH between inventories; (3) mortality trees that died between inventories due to causes other than harvest; and (4) harvest trees that were removed between inventories. For survivor trees, average annual DBH growth was calculated by dividing the total growth between inventories by the number of growing seasons between measurements. Measured DBH for the 1977 inventory was assigned to year 0, and DBHs for the 10 subsequent years were calculated by adding the average annual growth to the previous year's DBH. Because ongrowth trees were measured only in the 1990 inventory, DBH measurements for these trees were assigned to year 11, and DBHs for previous years back to year 0 were sequentially calculated by subtracting from current DBHs predictions of annual DBH growth obtained from the DBH growth models. Ongrowth status for these trees was designated in the year the tree attained the 12.7-cm minimum DBH. For mortality trees, a year of mortality between 1 and 10 was randomly selected from a uniform distribution and assigned to the tree independently of years of mortality assigned for other trees on the same plot. For harvest trees, a year of harvest between 1 and 10 was randomly selected from a uniform distribution and assigned to all trees harvested on the sample plot. For both mortality and harvest trees, DBHs measured in the 1977 inventory were assigned to year 0, and DBHs for subsequent years were calculated by adding previous year's DBHs and predictions of annual DBH growth obtained from the DBH growth models.

Calculation of unbiased estimates of change in BA (ΔBA) is difficult using data from variable radius plots (Van Deusen and others 1986). For these analyses, tree factors corresponding to year 0 were calculated for all trees and then held constant for the succeeding 10-year interval. Thus, annual database values of BA and BAL were calculated using the database of annual tree DBHs and the constant tree factors.

Although the procedures used to construct the annual database create somewhat greater uniformity in annual DBH growth, ongrowth, mortality, and harvest than would be observed, they represent a reasonable alternative. First, most other alternatives for distributing annual DBH growth or survivor trees would require either annual remeasurement or destructive sampling of all trees. Second, in the absence of precise knowledge of annual patterns of ongrowth, mortality, and harvest, uniform distributions represent overall long-term patterns that are reasonable for 10-year intervals. Finally, the impact on estimates of uncertainty, the primary entity of interest for this study, is expected to be minimal.

## THE SIMULATION PROCEDURES

Monte Carlo simulations were used to obtain estimates of uncertainty for model parameter estimates; 10-year ΔDBH and DBH predictions, plot BA estimates, and mean plot BA estimates; and annual inventory estimates of mean plot BA. Before the simulations could be implemented, uncertainty had to be quantified for three components: tree- and plot-level predictor variables, residual variability, and parameter estimates. In all situations, uncertainty in model predictor variables was assumed to be non-negligible.

### Uncertainty in Predictor Variables

Values of predictor variables are based on FIA field crew measurements and are subject to uncertainty. The tree-level predictor variables, DBH, CR, and CC, correspond to the measurement of a single physical entity, while the plot-level variables, BA, BAL, and PC, are sample estimates. Distributions for measurement errors for the tree-level predictor variables were obtained from the literature. McRoberts and others (1994) reported the results of a study in which 9-10 FIA field crews independently measured the same plots. They estimated a curve for describing the standard deviation of DBH measurements as a function of mean DBH. They also reported that distributions of ocular estimates of CR as percentages in the 0-1 range often deviated “0.3 around the median crew estimate. Nichols and others (1991) reported that when crews returned to plots later in the same growing season to obtain second ocular estimates of CC, 80 percent of estimates were unchanged while the remaining 20 percent were evenly distributed in the two adjacent classes. Uncertainty in BA and BAL estimates was simulated by using DBH measurements incorporating simulated DBH measurement error to calculate BA for each plot and BAL for each tree on each plot. Finally, because of the non-uniformity of plot soil, topographic, and vegetation conditions, PC is also subject to uncertainty due to sampling variability. However, because no empirical estimates of the sampling variability for PC are available, an arbitrary assumption was made that the coefficient of variation for PC is 10 percent.

### Residual Variability

Estimates of residual variability were obtained as by-products of calibrating the models. Residuals were assumed to follow a Gaussian distribution with zero mean but with heterogeneous variances. The standard deviations of the distributions of residuals were found to be adequately described as follows:

$$E[\ln(\hat{\sigma}_{res})] = \alpha_1+\alpha_2\ln(\hat{\Delta}DBH), \qquad [4]$$

where E(.) denotes statistical expectation of the quantity between the parentheses, ó$_{res}$, is the sample estimate of ó$_{res}$, and ΔDBH is predicted diameter growth from the models.

**Uncertainty in Model Parameter Estimates**

Model parameter covariances reflect uncertainty in the parameter estimates and must be included as a component of total uncertainty whenever model predictions are involved. When the models are relatively simple (eg., linear) and the uncertainty in predictor variables is negligible, parameter covariance estimates may be easily obtained using analytical methods. However, when the models are complex, nonlinear, and rely on predictors variables whose uncertainty cannot be assumed to be negligible, then Monte Carlo simulations are appropriate, if not also necessary, for reliably estimating these covariances. Failure to incorporate the uncertainty in the predictor variables results in underestimates of parameter covariances and, therefore, in underestimates of model prediction uncertainty. Using the distributions of uncertainty for the predictor variables and residual variation as previously described, distributions of model parameter estimates were obtained using a 4-step Monte Carlo procedure:

1. Simulated ΔDBH observations were obtained as the sums of two components: ΔDBH predictions obtained from the models using the parameter estimates obtained by calibrating the models to the observed data, and residuals randomly selected from a Gaussian distribution with zero mean and standard deviations obtained using equation [4] and the ΔDBH predictions.
2. Simulated values for predictor variables were obtained as sums of two components: observed values of the variables and either measurement error for DBH, CR, and CC, or sampling variability for PC obtained by randomly selecting values from the distributions previously described; using the simulated DBH observations, BA was calculated for each plot and BAL was calculated for each tree on each plot.
3. Model parameter estimates were obtained by fitting the models to the simulated ΔDBH observations obtained from Step 1 using the simulated values of the predictor variables obtained from Step 2; the resulting parameter estimates were recorded.
4. Distributions of model parameter estimates were obtained via 250 repetitions of Steps 1–3.

## Uncertainty Estimation

Estimates of the uncertainty in ΔDBH and DBH predictions and in derived BA variables were based on Monte Carlo simulations. The essence of the simulation procedures, explained in detail below, is to initialize plot and tree conditions using the annualized database of values, add random variability where appropriate to mimic uncertainty, use the models to predict annual DBH growth, record estimates at fixed time intervals, and repeat the process a large number of times to create a distribution of estimates.

Two approaches to evaluating uncertainty were used. The ACCUMULATING approach produces DBH predictions for each of 10 consecutive years by sequentially predicting ΔDBH using the models and adding the prediction to previous year's DBH to obtain current year's DBH. Annual estimates of plot BA, mean plot BA, and the standard error of mean plot BA are obtained and are designated the MODEL10 estimates. Uncertainty in estimates obtained with this approach represent the accumulated uncertainty in DBH predictions over the 10-year prediction interval.

The second approach is designated the ANNUAL approach and is intended to mimic the annual inventory system being implemented by the FIA program of the USDA Forest Service. The sampling design for this system features an equal probability grid of field plots which has been systematically divided into five interpenetrating, non-overlapping panels. Each year the plots in a single panel are measured with panels selected on a 5-year rotating basis. To mimic the annual inventory procedures, the plots included for these analyses were ordered with respect to their plot numbers and distributed among five equal-sized panels by systematically assigning every fifth plot to the same panel. Because FIA plot numbers had been assigned sequentially on the basis of the geographic locations of the plots, the panel assignments approximated the systematic, interpenetrating feature of the annual inventory sampling design. Annual inventory estimates of mean plot BA and the standard error of mean plot BA were obtained using three methods: (10 the SAMPLE20 estimates were based on measurements for the current year's 20-percent panel of plots; (2) the MOVING estimates were based on the most recent measurements for all plots; and (3) the UPDATE estimates were based on measurements for the current year's 20-percent panel of plots and updated information obtained using the growth models for the four panels of plots measured in previous years.

Estimates of the uncertainty in ΔDBH and DBH predictions and estimates of plot BA for the ACCUMULATING approach and in estimates of mean plot BA and the standard error of mean plot BA with both approaches were obtained using a 4-step Monte Carlo procedure:

1. Year 0:
   a. Measurement of all plots was simulated by adding the year 0 values of DBH, CR, CC, and PC from the annualized database and simulated measurement errors and sampling variability obtained by randomly selecting values from the distributions previously described; simulated DBH observations were recorded for each tree.

   b. Simulated values of BA and BAL were obtained from the simulated DBH observations by calculating BA for each plot and BAL for each tree on each plot; plot BA, mean plot BA, and the standard error of mean plot BA were calculated and recorded.

   c. A set of model parameter estimates for each species was randomly selected from the distributions previously constructed.

2. Subsequent years:
   a. ACCUMULATING approach. Simulated observations of ΔDBH for all trees were obtained as the sums of previous year's DBHs, predicted ΔDBHs, and residuals randomly selected from Gaussian distributions with zero mean and standard deviations obtained using [4] and predicted ΔDBHs; the simulated observation of DBH and the difference between current and previous years' simulated DBH observations were recorded for each tree.

   b. ANNUAL approach.
      (i) For panels selected for measurement, field measurement was simulated for all plots by replacing values for each tree with values from the annualized database for the appropriate year and adding measurement errors and sampling variability randomly selected from the appropriate distributions.

      (ii) For panels not selected for remeasurement, an updated value for DBH for each tree was obtained as the sum of previous year's DBH, predicted ΔDBH, and a residual randomly selected from a Gaussian distribution with zero mean and standard deviation obtained from [4] and predicted ΔDBH.

   c. For each of the four estimation methods, BA was calculated for each plot, BAL was calculated for each tree on each plot, and mean plot BA and the standard error of mean plot BA were calculated; plot BA, mean plot BA, and the standard error of mean plot BA were recorded for all four methods.

3. Step 2 was repeated 10 times to obtain predictions and estimates for all four methods for years 1-11.

4. Steps 1-3 were repeated 250 times to obtain distributions of DBH and ΔDBH predictions, plot BA estimates, and estimates of mean plot BA and the standard error of mean plot BA for each method for each year.

## ANALYSES

### Standards of Comparison

The standards of comparison for evaluating bias and the contribution of uncertainty in model predictions to the uncertainty in estimates of mean plot BA were the annual estimates of mean plot BA and the standard errors of mean plot BA obtained from the annualized database values. For comparison purposes, these estimates represent a current year sample of the entire geographic area under consideration and are regarded as being without measurement error. Estimates based on these values use 100 percent of the sample plots and are designated the SAMPLE100 estimates. Because the DBH values on which the SAMPLE100 estimates are based are regarded as having no uncertainty, any uncertainty in the SAMPLE100 estimates is due simply to sampling variability of trees on plots and BA estimates among plots.

### ACCUMULATING Approach

Uncertainty in ΔDBH and DBH predictions for individual trees, estimates of plot BA, and MODEL10 estimates of mean plot BA was quantified using the distributions of simulated estimates. Bias in the MODEL10 estimates of mean plot BA and the standard error of mean plot BA is evaluated by comparing these estimates to the comparable SAMPLE100 estimates. Differences between the medians of the distributions of MODEL10 estimates of the standard error of mean plot BA and the SAMPLE100 estimates quantify the effects of uncertainty in model predictions of DBH on the uncertainty of mean plot BA.

### ANNUAL Approach

Bias and uncertainty in the annual inventory estimates of mean plot BA and estimates of the standard error of mean plot BA were evaluated using the medians of the distributions of simulated estimates. Comparisons of median estimates of mean plot BA for the SAMPLE20, MOVING, and UPDATE methods to the annual SAMPLE100 estimates of mean plot BA provide the bias check. Comparisons of the medians of distributions of estimates of the standard error of mean plot BA for the UPDATE method to the SAMPLE100 estimates reveals the effects of uncertainty in model predictions on annual inventory estimates of mean plot BA.

## RESULTS

The adequacy of the 250 simulations was checked by evaluating the stability of coefficients of variation for the annual MODEL10 estimates of plot BA. For all plots, these coefficients of variation had stabilized by 100-150 simulations and were virtually unchanged for the final 50 simulations.

### ACCUMULATING Predictions and Estimates

Histograms of coefficients of variation for 10-year DBH and 10-year DBH predictions indicate that although the median coefficient of variation for ΔDBH was relatively large, approximately 0.20, the median for DBH was small, approximately 0.02 (fig. 1). Thus, 10-year DBH may be predicted quite precisely, even though 10-year ΔDBH is difficult to predict precisely. This result is attributed to two factors: first, as a component of 10-year DBH predictions, ΔDBH is relatively small compared to the other component, initial DBH; and second, the larger component, initial DBH, has little uncertainty, because DBH measurement error is small.

Bias in the MODEL10 estimates of mean plot BA was evaluated by comparing the medians of the distributions of the MODEL10 estimates of mean plot BA to the SAMPLE100 estimates (table 1). The Wilcoxon Signed Ranks test (Conover 1980) detected no statistically significant differences ($\alpha$=0.05) between the MODEL10 medians and the SAMPLE100 estimates. This result is consistent with observations that the medians of the MODEL10 estimates are in close proximity to the SAMPLE100 estimates and that they fall within a 2-standard error confidence interval around the SAMPLE100 estimates (fig. 2).

The medians of the distributions of the MODEL10 estimates of the standard error of mean plot BA were only slightly larger than the SAMPLE100 estimates. This result suggests that uncertainty in model predictions of ΔDBH has only a slight negative impact on the uncertainty in estimates of mean plot BA (table 1).

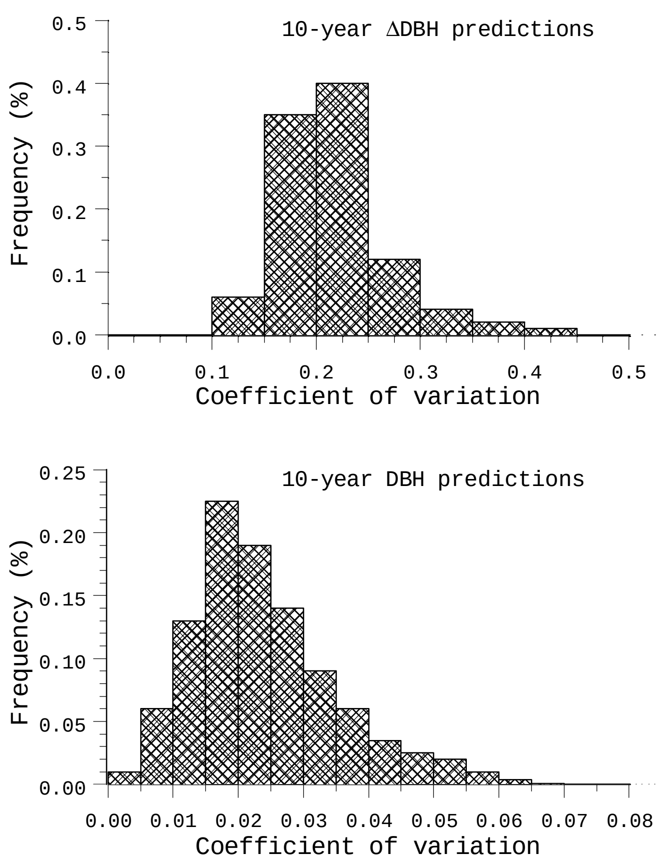


Figure 1—Distributions of simulated △DBH and DBH predictions.

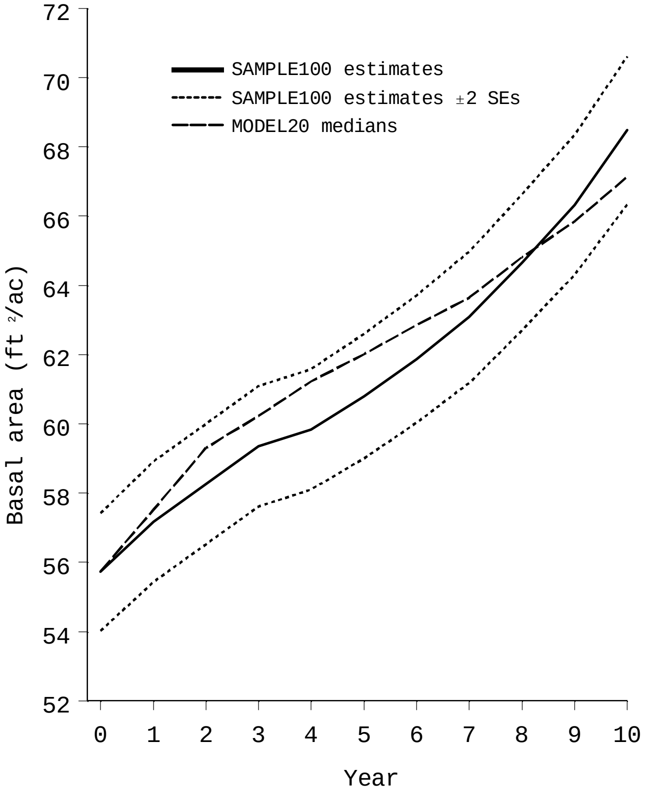


Figure 2—-Annual BA means obtained from DBH predictions.

**ANNUAL Approach**

Bias in the annual inventory estimates of mean plot BA was evaluated by comparing the medians of the distributions of the SAMPLE20, MOVING, and UPDATE estimates to the SAMPLE100 estimates (fig. 3, table 1). The medians of the SAMPLE20 estimates deviated considerably from the SAMPLE100 estimates due to the SAMPLE20 small sample size, while the medians of the MOVING estimates exhibited consistent bias due to the trend in the SAMPLE100 estimates. The medians of the distributions of the UPDATE estimates track the SAMPLE100 estimates quite closely, a result confirmed by the failure of the Wilcoxon Signed Ranks test to detect statistically significant differences (α=0.05).

The medians of the distributions of the UPDATE estimates of the standard error of mean plot BA were only slightly larger than the SAMPLE100 estimates, again indicating that uncertainty in model predictions of DBH has only a slight negative impact on the uncertainty of annual inventory estimates of mean plot BA.

**Table 1—Mean plot basal area estimates**

| | SAMPLE100 | | MODEL10 | | SAMPLE20 | | MOVING | | UPDATE | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Mean | SE | Mean | SE | Mean | SE | Mean | SE | Mean | SE |
| | | | | | *ft$^2$/ac* | | | | | |
| 0 | 55.72 | 0.85 | 55.74 | 0.86 | 55.72 | 0.85 | 55.72 | 0.85 | 55.72 | 0.85 |
| 1 | 57.17 | 0.87 | 57.51 | 0.87 | 60.12 | 2.02 | 56.22 | 0.86 | 57.79 | 0.88 |
| 2 | 58.26 | 0.87 | 58.87 | 0.88 | 59.57 | 1.97 | 57.37 | 0.86 | 59.20 | 0.88 |
| 3 | 59.35 | 0.87 | 60.23 | 0.88 | 57.04 | 1.94 | 57.59 | 0.87 | 60.61 | 0.88 |
| 4 | 59.84 | 0.87 | 61.23 | 0.89 | 61.01 | 1.92 | 58.50 | 0.87 | 61.08 | 0.89 |
| 5 | 60.80 | 0.90 | 62.01 | 0.91 | 59.45 | 1.85 | 59.44 | 0.87 | 61.98 | 0.91 |
| 6 | 61.87 | 0.92 | 62.85 | 0.93 | 64.33 | 2.16 | 60.28 | 0.89 | 62.87 | 0.93 |
| 7 | 63.08 | 0.95 | 63.65 | 0.95 | 63.04 | 2.16 | 60.97 | 0.90 | 63.97 | 0.96 |
| 8 | 64.66 | 0.98 | 64.81 | 0.97 | 63.14 | 2.24 | 62.19 | 0.93 | 65.45 | 0.99 |
| 9 | 66.33 | 1.01 | 65.85 | 1.00 | 68.92 | 2.26 | 63.78 | 0.96 | 66.94 | 1.02 |
| 10 | 68.48 | 1.07 | 67.14 | 1.07 | 67.60 | 2.24 | 65.50 | 0.99 | 68.93 | 1.07 |

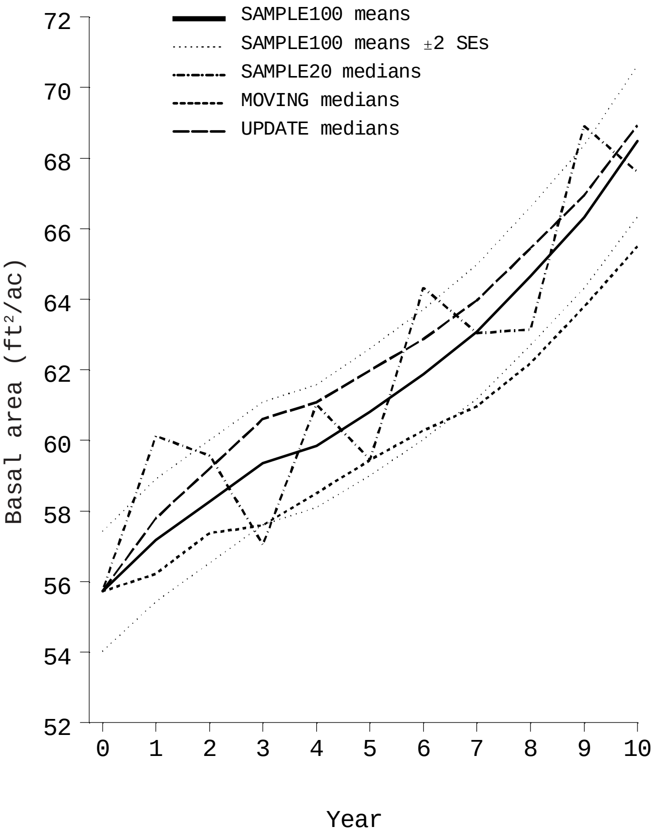


Figure 3—Annual BA means obtained from annual inventory system.

## CONCLUSIONS

Two conclusions emerge from this study. First, even though ΔDBH is relatively difficult to predict precisely, 10-year predictions of DBH were quite precise. This conclusion is partially attributed to the observation that 10-year ΔDBH is generally a relatively small component of 10-year DBH. The second conclusion is that the uncertainty associated with model-based updating technique had only a slight negative impact on the uncertainty of 10-year estimates of plot BA, and 10-year and annual inventory estimates of mean plot BA. For the mean plot BA estimates, this conclusion is partially attributed to the observation that DBH prediction uncertainty is relatively small compared to natural variability among estimates of plot BA. Acknowledgment is made, however, that a complete updating system also requires techniques for predicting the survival, regeneration, and removal of trees, components that are not considered in this study. Nevertheless, the study demonstrates that sufficiently unbiased and precise updates of DBH may be obtained.

## REFERENCES

**Conover, W.J.** 1980. Practical nonparametric statistics, 2nd ed. New York: John Wiley.

**Lessard, V.C.; McRoberts, R.E.; Holdaway, M.R.** Submitted. Diameter growth models using Minnesota FIA data. Forest Science.

**McRoberts, R.E.; Hahn, J.T.; Heft, G.J.; Van Cleve, J.** 1994. Variation in forest inventory estimates. Canadian Journal of Forest Research. 24: 1766–1770.

**McRoberts, R.E.** 1999. Joint annual forest inventory and monitoring system. Journal of Forestry. 97(2): 21–26.

**Miles, P.D.; Chen, C.M.; Leatherberry, E.C.** 1995. Minnesota forest statistics, 1990, revised. USDA Forest Service, Resour. Bull. NC–158. St. Paul, MN: U.S. Department of Agriculture, North Central Forest Experiment Station. 138 p.

**Myers, C.C.; Beers, T.W.** 1971. Point sampling and plot sampling compared for forest inventory. Res. Bull. 877. Aug., 1977. Lafayette, IN: Purdue University Agriculture Experiment Station.

**Nichols, N.S.; Gregoire, T.G.; Zedaker, S.M.** 1991. The reliability of tree crown classification. Canadian Journal of Forest Research. 21: 698–701.

**Spencer, J.S.** 1982. The fourth Minnesota forest inventory: timber volumes and projections of timber supply. USDA Forest Service, Res. Bull NC–57. St. Paul, MN: U.S. Department of Agriculture, North Central Forest Experiment Station.

**Van Deusen, P.C.; Dell, T.R.; Thomas, C.E.** 1986. Volume estimation from permanent horizontal points. Forest Science. 32(2): 415–422.

# A COMPARISON OF SEVERAL TECHNIQUES FOR ESTIMATING THE AVERAGE VOLUME PER ACRE FOR MULTIPANEL DATA WITH MISSING PANELS[1]

Dave Gartner and Gregory A. Reams[2]

**Abstract**—As Forest Inventory and Analysis changes from a periodic survey to a multipanel annual survey, a transition will occur where only some of the panels have been resurveyed. Several estimation techniques use data from the periodic survey in addition to the data from the partially completed multipanel data. These estimation techniques were compared using data from two periodic surveys from Georgia. The comparison criteria were based on (1) an estimated mean within the confidence interval derived from using the complete multipanel data set and (2) a small, estimated standard error that does not underestimate the complete data standard error. Multiple imputation matching performed best; the double sampling ratio estimator also performed well. Two methods—single imputation using group means and single imputation using matched stands—both underestimated the standard error. Replacing the missing observations with growth model predictions using SETWIGS caused an overestimation of the mean.

## INTRODUCTION

The USDA Forest Service, Forest Inventory and Analysis (FIA) Units have been conducting surveys of commercial forest land in the continental United States since the 1930s. Traditionally, FIA has conducted surveys on a State level with a cycle from 6 to 15 years with a mode of about 10 years in the South. Prior to a tightening of the supply and demand relationship for wood fiber in the South, the 10-year cycle was considered timely enough (Reams and others 1999).

With the growing demand for wood products from the South, the need for more current inventory information has become apparent. To meet this need, the Agricultural Research, Extension, and Education Reform Act (Public Law 105–185): (The Farm Bill) of 1998 mandated FIA to implement an annual inventory system Nationwide.

Southern FIA is changing from single-panel (periodic) whole-State surveys to an interpenetrating five-panel annual survey (Reams and Van Deusen 1999). The latter design divides the large periodic survey into five repeated smaller samples, called panels (Reams and Van Deusen 1999). By providing information about the variations between years, the separate annual samples are able to estimate annual and secular trends.

The new annual five-panel design will give rise to new estimation techniques. The new official FIA estimate will be a moving average using the annual survey data (Reams and others 1999). The moving average is operationally convenient, requires a minimum of assumptions, and is basically design-based as opposed to model-based.

To understand how the moving average will be implemented, consider the following situation: (1) the last full periodic survey has been completed; (2) starting immediately afterwards, the five-panel annual system has been implemented; and (3) three panels have been completed and now an estimate of live standing volume per acre for Georgia is needed. The official FIA estimate will be the average, using the annual survey data from the plots in panels one through three and the closeout periodic survey data for plots in panels four and five. Note that plots from panels four and five have yet to be measured under the annual system; therefore, the plot attributes are at least 3 years old.

Some users of the annual survey data suggested using statistical modeling techniques to update the data values. Some of these techniques replace the missing data in the unsurveyed panels with estimates from the surveyed panels. In the statistical literature, this replacement of missing data with modeled data is called imputation (Rubin 1987). After imputing data values for old or unmeasured plots, it would be tempting to analyze a simulated-complete data set as a complete data set. However, this approach tends to understate the true variance in the estimates (Little and Smith 1987, Van Deusen 1997).

This study compares the performance of several techniques. In addition to imputation, the double sampling ratio estimator was used as a comparison. Because multiple imputation is conservative in its estimate of the variance (Rubin 1987), the variance estimate for double sampling is expected to be lower than for multiple imputation.

## METHODS

### Data

We simulated the end of the third panel, having access to the data from the first three panels and the last periodic survey. The variable in the comparison is the statewide average volume of live trees in cubic feet per acre. To compare predicted values with observed values, we used the 1988 and 1996 periodic surveys from Georgia. To simulate conditions at the end of year three, we deleted the

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Mathematical Statistician and Supervisory Mathematical Statistican, USDA Forest Service, Southern Research Station, 2730 Savannah Highway, Charleston, SC 29414, and USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804, respectively.

stand volume data from 40 percent of the 1996 plots. We used only unit, county, plot, forest type, and stand origin from 1996 plots with deleted stand volume data.

## Operational Information Assumptions

We assumed that the regional FIA units know which areas have been harvested and the volumes of any remnant stands. In the simulation, we coded as “cut” all stands with trees that were marked as cut in the 1996 survey. We put these cut stands in a separate data set and did not delete any of their 1996 volume data. The prediction methods could have been modified to handle harvested stands by including a probability of being harvested and a prediction of harvesting intensity, but we decided against this approach.

## Data Preparation

Two major changes in the plot sampling protocols occurred between the 1988 survey and the 1996 survey. One was the change from using variable-radius plots to fixed-radius plots. The second was the handling of plots that contained more than one stand. During the 1988 survey, if any of the subplots fell into a different stand than the central subplot, that subplot was rotated until it fell within the same stand as the central subplot. During the 1996 survey, if any subplot fell into a different stand or stands, the subplots were not rotated but the different stands were given different codes, called condition codes. To make sure the 1988 data used to predict 1996 volumes matched the observed 1996 stands, only pure stands were used. If the trees measured in 1988 appeared in more than one 1996 condition code, we removed the plot from the data set. If a 1996 condition code did not have any 1988 trees, we removed the condition code from the data set.

After we removed these data, 3,749 plots remained. We calculated the live tree volumes in cubic feet per acre for each plot for both survey years. We then placed the 1,194 stands with cut trees in a separate data set. We simulated the two unsurveyed panels by deleting the 1996 volume data for 40 percent (1,020 out of 2,555) of the remaining stands. To suit the estimation techniques being run by forest type, we placed the forest types with fewer than seven plots in the 60 percent of the data that represent the three surveyed panels.

## Estimation Techniques

**Three-panel method**—The first estimation technique uses only the 1996 volume data for the cut stands and the three surveyed panels. We calculate the means and standard errors for both groups, and then combine them, weighting the uncut stands to include the number of stands with missing volumes.

**Single imputation group means**—Adding information on the forest types yields the second estimation technique. With single imputation group means, the missing volumes are replaced with the average observed volume for that forest type. We then recombine the cut and uncut data sets and use standard estimation procedures.

**Single imputation matching**—Adding the information on the 1988 volumes yields two more types of estimation techniques: single imputation matching, and multiple imputation matching. For single imputation matching, we find stands with 1988 volumes and forest types that match those of the stands with missing 1996 volumes. Once we find a matching stand, we replace the missing 1996 volume data with the data from the matched stand. Then we recombine the cut and uncut data sets and use standard estimation procedures.

**Multiple imputation matching**—Multiple imputation (Rubin 1987) matching differs from single imputation matching, in that a set of possible donor plots is sought for each missing value. A separate donor stand is then randomly chosen for each missing value from its donor pool. We repeat this process of randomly choosing donor plots several times, and combine the results from the repetitions.

For each imputed data set, we calculate the statistic of interest (mean live tree volume per acre), denoted as $\hat{Q}_{*l}$. The variance of $\hat{Q}_{*l}$ is denoted as $U_{*l}$. In this case, $U_{*l}$ is the standard error. The function for the estimated mean is

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

where m is the number of repetitions of the imputation process. The estimator for the variance of $\bar{Q}_m$ has two components. The first component is the average of the variances of this mean:

$$\bar{U}_m = \sum_{l=1}^{m} U_{*l} / m. \tag{2}$$

The second component of this variance estimator is the variance of the $\hat{Q}_{*l}$’s:

$$B_m = \sum_{l=1}^{m} (\hat{Q}_{*l} - \bar{Q}_m)^2 / (m-1). \tag{3}$$

These two components are combined in the following manner:

$$T_m = \bar{U}_m + (1 + m^{-1}) B_m. \tag{4}$$

When standard errors are mentioned in the results for multiple imputation techniques, we use the square root of $T_m$. The estimated overall mean has a t distribution with mean $\bar{Q}_m$ and standard error of the square root of $T_m$. The degrees of freedom according to Rubin (1987) is

$$\nu_m = (m-1)\left[1+\frac{\overline{U}}{(1+m^{-1})B_m}\right]^2 \tag{5}$$

This degrees of freedom has been given a modifier for possible small sample sizes (Barnard and Rubin 1999). The modifier is

$$\hat{\nu}_{obs} = (1-\gamma)\,\nu_0(\nu_0+1)/(\nu_0+3), \tag{6}$$

where $\gamma = (1+m^{-1})B_m/T_m$ and $\nu_0$ is the degrees of freedom of the full sample if no data values are missing. The final degrees of freedom is

$$\nu_m^* = \left[\frac{1}{\nu_m}+\frac{1}{\hat{\nu}_{obs}}\right]^{-1}. \tag{7}$$

The main advantages of multiple imputation over single imputation are that the variance caused by the process of randomly choosing donor plots is empirically estimated (eq. 3) and is explicitly included in the estimate of the overall variance.

**Multiple imputation modeling**—The assumption that the 1996 volume is an approximately linear function of 1988 volume for uncut stands yields another estimation technique. Multiple imputation modeling estimates the parameters from a linear regression of 1998 volume on 1996 volume. We modify these parameters by adding a random error term, determined by decomposing the parameters' variance-covariance matrix. Using a Cholesky decomposition turns the variance-covariance matrix into a set of variances for independent normal variables. We then multiply random normal variates by these variances and add them to the parameter estimates. We use these modified parameters to estimate the missing values. We calculate an imputed standard deviation by randomly generating a Chi-square variable and multiplying it by the observed standard deviation. We generate standard normal deviates, multiply them by this imputation standard deviation, and add them to the estimates for the missing values. We repeat this process several times and analyze the results in the same manner as for the multiple imputation matching data. Thankfully, the current multiple imputation software does all of these computations. As with the multiple imputation matching method, the objectives of repeating the process are (1) to empirically estimate the variance of the mean due to randomization, and (2) to incorporate this variance into the total variance for the estimator.

**Single imputation growth model**—Growth models use the 1988 tree-level information, such as species, diameters, and expansion factors, along with plot site index. We simply replace the missing volumes with the growth model predictions. For this study we used the growth model, SETWIGS (Bolton and Meldahl 1990).

**Multiple imputation using growth model predictions**—We could incorporate growth model predictions into multiple imputation efforts in two different ways. The first way would be to replace the missing data with the growth model projections. According to Rubin (1987), the proper method for replacing the missing data with the growth model predictions is the same method used for the linear regression predictions in multiple imputation modeling, including decomposing the parameter variance-covariance matrix and imputing new parameters and standard errors. Unfortunately, the current multiple imputation software will not calculate these values. We decided not to use this method because of the effort it required.

The second method of incorporating growth model predictions into multiple imputation is using the growth model predictions as the covariate. We used multiple imputation matching and multiple imputation modeling techniques by replacing the 1988 volume information in the earlier multiple imputations with the growth model predicted volumes.

**Double sampling ratio estimator**—We also used the classical sampling statistical technique called double sampling using a ratio estimator (Cochran 1977). Double sampling occurs when a sample is taken and the value of one variable (X) is observed. Then a subsample of the first sample is taken and the value of the variable of interest (Y) is observed. The estimated average for Y on the whole sample using a ratio estimator ( $\bar{y}_R$ ) is the average of X for the whole sample ( $\bar{x}'$ ), times the ratio of the average of Y for the subsample ( $\bar{y}$ ) divided by the average of X for the subsample ( $\bar{x}$ ):

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}}\bar{x}'. \tag{8}$$

The variance of this estimator is given by equation 9:

$$V(\bar{y}_R) = \frac{1}{n'}s_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)s_d{}^2, \tag{9}$$

where $n$ is the number of observations in the subsample, $n'$ is the number of observations in the full sample, and

$$s_d{}^2 = \sum_{i=1}^{n}\left(y_i - \frac{\bar{y}}{\bar{x}}x_i\right)^2/(n-1). \tag{10}$$

Note this estimator is not the same as regression through the origin. In this instance, we used Y as the 1996 volume for the uncut stands and the growth model-predicted volumes for X. We then combined the estimates and variances for the cut and uncut stands.

## Solas Software

We ran all multiple imputations using Solas software (1999). The multiple imputation matching techniques followed the propensity score method, which uses a logistic regression equation to predict the propensity of an X value to correspond with a missing Y value. A donor pool of observations is created from a local neighborhood of propensity values. If only one covariate is used, the predicted propensity is a monotonic function of the covariate, and the neighborhood of predicted propensities is the same as the neighborhood of the covariate values. However, this condition will not necessarily be true if more than one covariate is used. Because Solas creates a separate equation for each level of a grouping variable, adding grouping variables, such as forest type, will still allow the propensity scoring method to act as a matching method.

Using multiple grouping variables causes a problem for Solas. Solas can run only 30 groups at a time. With more than 30 forest types, we had to break the data set into several parts. We imputed each part separately, then merged them together again. Solas also runs out of memory and has trouble with large data sets. We hope the new SAS multiple imputation procedure will have fewer limitations.

We also ran the single imputation matching using Solas. Instead of limiting the donor pool to just the plot with the next larger and the next smaller 1988 plot volumes, Solas required that the donor pool include at least the next two larger and two smaller plot volumes. In keeping with single imputation, we picked only one value per observation with missing data.

## RESULTS

Because the multiple imputation techniques use randomization, we ran all multiple imputation techniques five times to estimate the variability caused by the randomization. We report this variability for the estimated statistics (tables 1 to 3).

### Means

Figure 1 shows the relationship between 1988 and 1996 volumes for the stands that were not cut. Several stands along the 1988 volume axis show that criteria for determining cut stands did not catch all of the stands that lost volume. The volume losses were probably the result of natural disturbances as opposed to harvesting.

The full 1996 data have a mean of 1,569.96 ft$^3$ per ac (table 1) and a standard error of 22.18 (table 2). Because the multiple imputation methods were run five times each, the overall means and standard deviations of the estimated means appear on table 1. All but one of the estimation techniques gave means within 1.15 standard errors of the full data mean. Replacing the missing 1996 volume data with the SETWIGS growth model projections provided a mean of 1,771.63 ft$^3$ per ac.

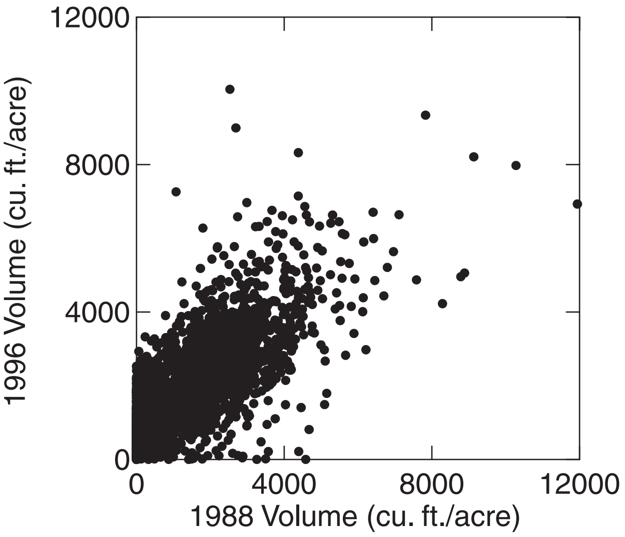


Figure 1—Observed 1988 stand volumes versus observed 1996 stand volumes.

Figure 2 shows the relationship between the volumes predicted by SETWIGS and the 1996 observed volumes for uncut stands. The line on the graph shows where the predicted volume equals the observed volume. Most of the points fall below the line, showing that the SETWIGS-predicted volumes were larger than the observed volumes. This estimated mean is about nine standard errors above the mean found by using all of the data.

**Table 1—Estimated mean stand volume by estimation technique**

| Estimation technique | Means |
|---|---|
| All of the data | 1,569.96 |
| Three surveyed panels only | 1,568.60 |
| Group mean imputation | 1,577.01 |
| Single imputation matching: volume 1988 | 1,556.53 |
| Multiple imputation matching: volume 1988[a] | Mean 1,571.57, std. dev. 1.89 |
| Multiple imputation modeling: volume 1988[a] | Mean 1,570.77, std. dev. 17.91 |
| Single imputation: SETWIGS | 1,771.63 |
| Multiple imputation matching: SETWIGS[a] | Mean 1,574.68, std. dev. 4.14 |
| Multiple imputation modeling: SETWIGS[a] | Mean 1,575.01, std. dev. 9.61 |
| Double sampling ratio estimator: SETWIGS | 1,577.15 |

[a] Multiple imputation techniques were run five times. The reported results are the mean and standard deviation of the five runs.

**Table 2—Standard errors of the various estimation techniques**

| Estimation technique | Standard error |
|---|---|
| All of the data | 22.18 |
| Three surveyed panels only | 25.23 |
| Single imputation group means | 19.45 |
| Single imputation matching: volume 1988 | 21.90 |
| Multiple imputation matching: volume 1988[a] | Mean 23.86, std. dev. 0.34 |
| Multiple imputation modeling: volume 1988[a] | Mean 30.80, std. dev. 3.16 |
| Single imputation SETWIGS | 25.43 |
| Multiple imputation matching: SETWIGS[a] | Mean 23.43, std. dev. 0.59 |
| Multiple imputation modeling: SETWIGS[a] | Mean 23.97, std. dev. 1.38 |
| Double sampling ratio estimator: SETWIGS | 23.15 |
| All of the data: stratified sample | 20.16 |

[a] Multiple imputation techniques were run five times. The reported results are the means of the standard errors and their standard deviations.

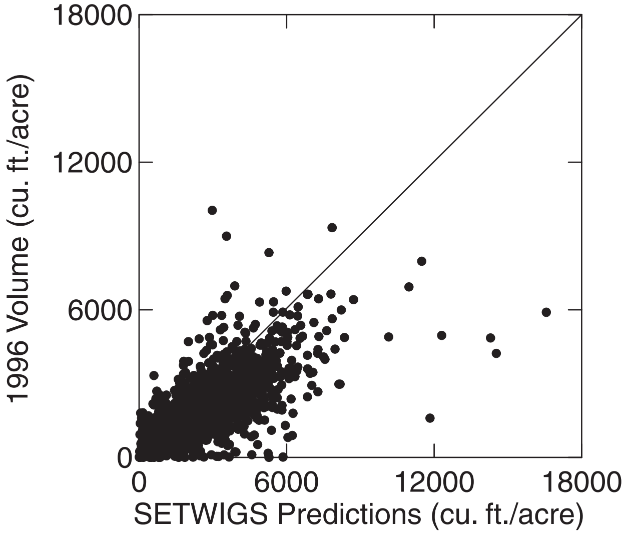


Figure 2—SETWIGS predicted volumes versus observed 1996 stand volumes.

## Standard Errors

The estimated standard errors are shown in table 2. As with the estimated means, for the multiple imputation methods, the means and standard deviations of the five estimated standard errors are reported in table 2. The standard errors for the single imputation group means and the single imputation matching are smaller than the standard error found when using the full data set. The double sampling ratio estimator and the multiple imputations have larger standard errors than using the full data set. Generally, the matching techniques outperformed the modeling techniques, and SETWIGS predictions outperformed 1988 volume.

## Mean Square Error

The mean square error is the bias squared plus the standard error squared. The mean square error for each method appears in table 3. As with the estimated means, for the multiple imputation methods, the means and standard deviations of the five mean square errors are shown in table 3. Only the single imputation group means

**Table 3—Mean square errors of the various estimation techniques**

| Estimation technique | Mean square error |
|---|---|
| All of the data | 491.95 |
| Three surveyed panels only | 638.40 |
| Single imputation group means | 428.00 |
| Single imputation matching: volume 1988 | 659.97 |
| Multiple imputation matching: volume 1988[a] | Mean 581.69, std. dev. 16.23 |
| Multiple imputation modeling: volume 1988[a] | Mean 1361.04, std. dev. 503.69 |
| Single imputation: SETWIGS | 41317.47 |
| Multiple imputation matching: SETWIGS[a] | Mean 590.03, std. dev. 66.46 |
| Multiple imputation modeling: SETWIGS[a] | Mean 686.78, std. dev. 100.88 |
| Double sampling ratio estimator: SETWIGS | 587.62 |
| All of the data: stratified sample | 406.43 |

[a] Multiple imputation techniques were run five times. The reported results are the means of the mean square errors and their standard deviations.

has lower mean square errors than using all of the data. As with the standard errors, all of the multiple imputation techniques had higher mean square errors, with the matching techniques performing better than the linear modeling techniques.

## DISCUSSION

The single imputation group means technique removes the variation between the missing volumes and their means by forest type. Removing this variation causes the standard error and the mean square error to be underestimated. Single imputation matching limits the variation in a similar manner. With a given set of original stand conditions, there is a range of possible ending conditions. Single imputation matching limits the variation less than does single imputation group means and, therefore, underestimates the standard error less than group means imputation. However, it still underestimates the standard error and the mean square error.

The standard error for double sampling ratio estimator is very close to the standard errors for the multiple imputation matching methods. The double sampling ratio estimator required splitting the plots into harvested and unharvested strata, while the multiple imputation methods did not stratify the data. Therefore, the multiple imputation matching methods actually performed slightly better than the double sampling ratio estimator.

Creating an inventory estimate for the harvested stands is more complex, especially for the double sampling ratio estimator and the single imputation growth model method. To use either of these methods on the harvested plots would require either the ability to remotely sense all harvests each year, or the creation of probability-of-harvest models and a method of allocating partial harvests to individual trees. The southern FIA unit currently does not have the budget or the infrastructure to be able to remotely sense harvesting on an annual basis. Some work has been done on probability of harvest models for stands, but no work has been done on methods of allocating partial harvests to individual trees. The multiple imputation methods would not require using remotely sensed information on harvesting or the probability of harvest models. Currently, about 20 percent of the stands are harvested (either clearcut or partial) within a 5-year cycle.

All of the imputation techniques predict plot level data and then calculate overall means and standard errors. The double sample ratio estimator, however, is not an imputation technique because it does not calculate values for the missing observations. Therefore the double sampling ratio estimator may not be suitable for variables that are difficult to model. The number of snags per acre, amount of fallen woody debris, and ownership are examples of such variables. Some tables, such as the diameter distribution tables, may be sensitive to the differences between model predictions and observed data, and may not be fit well by the double sample ratio estimator.

The multiple imputation techniques may have an additional advantage. While each table requires a separate run using double sampling ratio estimator, properly constructed multiple imputation data sets can be used for all tables simultaneously.

## ACKNOWLEDGMENTS

We thank Michael Thompson for his assistance with the Georgia data and the volume equations.

## REFERENCES

**Barnard, J.; Rubin, D.** 1999. Small-sample degrees of freedom with multiple imputation. Biometrika. 86: 948–955.

**Bolton, R.K.; Meldahl, R.S.** 1990. Design and development of a multipurpose forest projection system for southern forests. Alabama Agricultural Experiment Station Bulletin 603. Auburn, AL: Auburn University. 51 p.

**Cochran, W.G.** 1977. Sampling techniques. New York: John Wiley. 427 p.

**Little, R.J.A.; Smith, P.J.** 1987. Editing and imputation for quantitative survey data. Journal of the American Statistical Association. 82: 58–68.

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

**Statistical Solutions, Ltd.** 1999. SOLAS for missing data analysis 2.0 [User reference]. Cork, Ireland: Statistical Solutions, Ltd. 67 p.

**Reams, G.A.; Roesch, F.A.; Cost, N.D.** 1999. Annual forest inventory: Cornerstone of sustainability in the South. Journal of Forestry. 97(12): 21–26.

**Reams, G.A.; Van Deusen, P.C.** 1999. The southern annual forest inventory system. Journal of Agricultural, Biological, and Environmental Statistics. 4(4): 346–360.

**Van Deusen, P.C.** 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27: 379–384.

# FIA ESTIMATION IN THE NEW MILLENNIUM[1]

**Francis A. Roesch**[2]

**Abstract**—In the new millennium, Forest Inventory and Analysis (FIA) will deliver most of its database information directly to the users over the Internet. This assumption indicates the need for a GIS-based estimation system to support the information delivery system. Presumably, as the data set evolves, it will free FIA and the users from exclusive estimation within political boundaries.

A data set of basal area measurements from a survey unit in Georgia is used to simulate one that might have been obtained had an annual inventory been conducted over a 5-year time interval. The simulated data set was used to investigate various estimators and any potential spatial correlation of basal area. The presence of spatial correlation, coupled with a desire to fulfill user needs to obtain estimates over individually defined elements of the spatial-temporal cube, forms the basis for an argument that a real-time GIS-based estimation system should be developed as the main information delivery vehicle for FIA.

## INTRODUCTION

As we approach the next millennium, it is apparent that we cannot consider how we might improve Forest Inventory and Analysis (FIA) estimation without first asking: "What are the major products of FIA likely to be?" That is, through which routes will we deliver the bulk of our inventory information? Most likely, we will deliver most of our information directly from the database over the Internet, not in the paper reports that have historically taken about 2 years to publish. To use the Internet effectively and efficiently, we must build an estimation system to adequately support the delivery of information that is more sensitive to the needs of its users.

Insights into the needs of these users can come from myriad sources, but none so compelling as the reports of the two Blue Ribbon Panels, BRP I (Anonymous 1992) and BRP II (Anonymous 1998). These panels were formed specifically to provide suggestions for improving the FIA program. For instance, a concern over the potential misuse of FIA data resulted in the following statement from BRP I: "To maintain the credibility of the program, FIA, working together with experienced biometricians, must issue clear direction on the scientifically valid uses of FIA data without creating disincentives to innovation and advancement of technology" (Anonymous 1992).

The best way to communicate scientifically valid uses of the information is to develop a system that can provide estimates in as many usable forms as possible. In this manner, FIA will provide scientifically defensible mechanisms from which to make estimates. FIA may still challenge inferences drawn from the estimates, but if the estimates themselves are sound, the scientific community can debate the validity of various resulting inferences.

The second Blue Ribbon Panel reiterated and expanded the recommendations of the first in one specific recommendation:

"Better analysis is necessary for improving customer service. More analysis of FIA data would be useful in improving and increasing customer service. While some FIA customers have the capability and inclination to analyze raw data themselves, other customers rely on outside sources to summarize and analyze the data for them" (Anonymous 1998).

In addition, the second Blue Ribbon Panel charged FIA to "Produce the most current resource data possible."

The overwhelming consensus among panel members was that timeliness of resource data is of paramount importance:

"Strengthening of Forest Service research and expertise in Geographic Information Systems (GIS), and collaboration with other agencies, could deliver immediate benefits. We urge the Forest Service to:

"Reallocate funding within the Forest Service in order to reach the goal of timely resource data established in the first Blue Ribbon Panel report. Fully integrate GIS technology into the inventory process. Aggressively support and promote the annual inventory systems being established in the North Central and Southern FIA units. Based upon results from these efforts, establish a model for annual inventory to be adopted nationwide" (Anonymous 1998).

To fully comprehend the needs of the users, we must first identify those users. FIA users include State foresters, university researchers, National Forest System employees, Forest Service researchers, military bases, other government and State agencies, forest industry, forestry consultants, and members of conservation and environmental groups. Their needs are as diverse as the groups themselves.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Mathematical Statistician, USDA Forest Service, Southern Research Station, Asheville, NC 28804.

The term “drill-down technology” refers to a database feature that allows a user to view increasing levels of detail as scale increases. It is used extensively in mapping software and GIS packages to deliver the appropriate level of information at varying scales. For example, if one were interested in regional wood supply information, providing individual tree-level data would not be very helpful. On the other hand, region-level information provides little more than background to a State forester interested in a particular county’s available forest resource. A resourceful user can derive region-level information from tree level data from the entire region. However, users are generally supportive of programs that provide information in the most useful forms and at the most appropriate scales. It is apparent that the most efficient aggregator of FIA data into appropriately scaled information will usually be FIA itself. Exceptions do occur with a few special-needs users.

Given the richness of the data that will be available from the annual inventory design, and the power of existing GIS systems, the user of FIA data should not be constrained by boundaries in space or time that have been predefined by FIA. However, given user-defined spatial and temporal constraints, FIA should endeavor to provide estimates in real time. These estimators should be available for as diverse a set of needs as the data will allow. Certainly, the data support investigations into forest amenities and commodities at a wide range of scales, but they can also provide insights into the contributions and effects of forests in wide-ranging areas of interest. Some related areas include studies of pollution, watersheds, and even human behavior.

To increase the usefulness of our information, we must incorporate all of the available improvements in user-interface tools. As a minimum standard, the user should not have to know any variant of Structured Query Language (SQL). This requires that we provide the estimation system in a user-friendly GIS environment.

To ensure the greatest utility of the data, FIA must provide an estimation system robust to an unpredictable and uncontrollable set of click events. This will compel FIA statisticians to reach as deeply into their estimation toolbox as any single previous effort has ever required. A completely different approach will be required if the user is permitted to define areas of interest—say by digitization or by map overlays—and time periods of interest rather than be required to work within strictly defined boundaries in space and time. Given the plethora of information available in the data set, a truly robust system would often have to use estimators that “reach out” to external data and other information sources for support, rather than to rely solely on the FIA data collected within the user-defined, spatial-temporal limits. A GIS-based estimation system has to provide the “best” estimators at any scale of interest within the estimation range. For most attributes, the most intensive scale in the estimation range for FIA data includes areas the size of a large county. However, the relationship of available information to the area delineated varies by the size of the area relative to the sample, the variable of interest, and the period of interest. Large areas require only the usual sample estimates when sample sizes within the area and period are adequate, while small areas require the use of supplementary information from outside of the area or period.

## METHODS

Before an estimation system can be incorporated into a production system, its individual components, as well as the relationships between those components, must be thoroughly tested. This leaves us with the problem of testing a large, potentially complex, estimation system prior to the availability of the data. The approach we used was to manufacture a reasonably believable 5-year series of data by projecting data from a single year backward and forward 1 and 2 years. The data from FIA’s Survey Unit 1 in Georgia, collected in 1989 and 1996, were used to establish individual-tree basal area projection equations, mortality and harvest probabilities, and proportions by forest type, dominant species, and age class. These functions were then applied to the 1996 tree-level data to project it backward 1 and 2 years and forward 1 and 2 years, simulating tree data for 5 consecutive years on 2,353 plots. The survey unit consists of 35 counties, which were grouped into 5 contiguous 7-county groups for part of this study. This data set was considered to represent the “truth” for each of the years 1994 through 1998. Figure 1 graphs the “true” mean basal areas per acre. We define the “current truth” as the state of this simulated population in 1998.

The sample plots for the FIA Annual Inventory sample design are located in a systematic triangular grid consisting of five interpenetrating panels. One panel is measured each year for five consecutive years, after which the panel measurement sequence reinitiates. If panel 1 was measured in 1998, it will also be measured in 2003, 2008, and so on. Panel 2 would be measured in 1999, 2004, 2009, and every five years thereafter.

To mimic the systematic FIA Annual Inventory design, spatial coordinates of the plots were used to assign plots to panels, a panel being a single year’s measurement. Therefore, the simulated FIA Annual Inventory sample consisted of approximately one-fifth of the plots for each year.

A preliminary study investigated specific applications of two general methods for combining the multiyear data from the FIA annual inventory design to form current estimates for small areas. The two general methods are (1) the simple moving average estimator (MAE), and (2) a globally defined mixed estimator (ME) applied locally. Two variations of the mixed estimator method (ME1 and ME2) are compared to each other as well as to the assumed default estimator (MAE). Assume that one and only one full series of observations is available so that all five panels have been measured once. “Current” is defined as the measurement time of the last panel (panel 5).

MAE pools the latest five panels measured, under the assumption that no time trend exists at the observed scale. As some variables of interest will violate this assumption

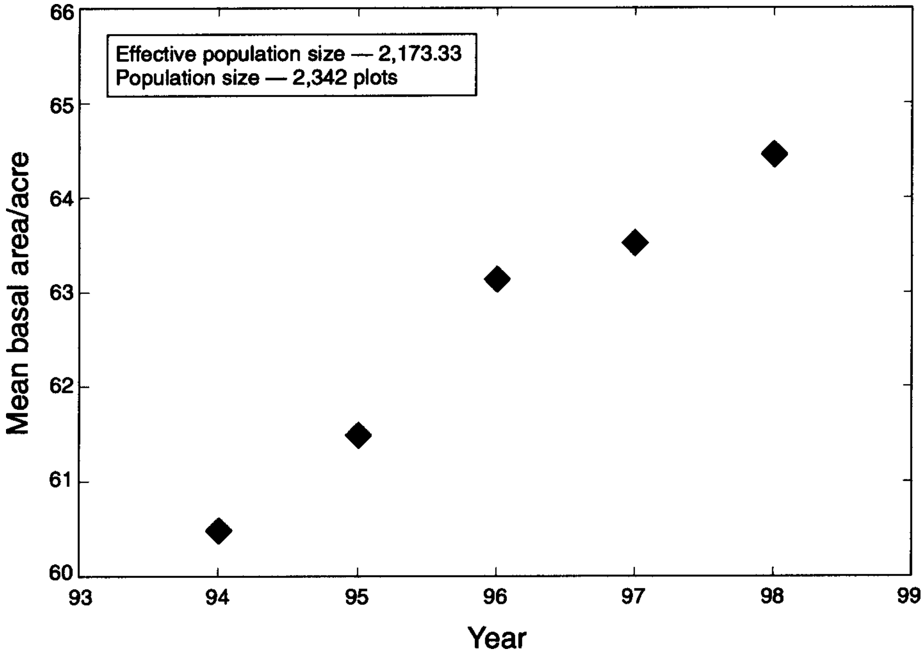


Figure 1—Survey unit "true" mean basal area per acre by year.

over the measurement interval, it is important to determine valid applications of this simple estimator.

Modeling an existing temporal trend becomes important when the objective is to estimate the time-specific value of some forest attribute, e.g. current volume or basal area per acre. When a temporal trend does in fact exist, MAE will have the tendency to mask the very trends that the FIA annual inventory design was intended to evaluate. Therefore, we explored the mixed estimator because it can recognize and efficiently utilize the time-series nature of the five-panel sample.

If we seek the estimate for a variable at a specific time, let:

$X_{ijt}$ = the per-acre value observed at plot $i$ in county *j*

($i = 1,..., n_j$, *j* = 1,...,*J*), and time *t* (*t* = 1,...,5),

$A_{ijt}$ = the area in acres sampled at plot $i$ in county *j*

($i = 1,..., n_j$, *j* = 1,...,*J*), and time *t* (*t* = 1,...,5), and

$A_P$ = the fixed plot area.

When no time trend is present, the sample area weighted mean for the five-panel series provides the best estimator of a per-acre value (*V*):

$$\hat{V}_{MAE} = \frac{1}{A}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{A_{ijt}}{A_P}X_{ijt} \qquad (1)$$

where:

$$A = \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{i=1}^{n_j} A_{ijt}$$

We tested two variations of mixed estimation for current basal area. Each variation applies global (survey unit) results of the mixed estimation methodology to subareas within the survey unit, under the assumption that the sample will often be too small for a direct application of mixed estimation to the subareas. In both variations, we used mixed estimation at the survey unit level to choose from the three simple models discussed by Van Deusen (1999), and to find the maximum likelihood estimate of the weighting parameter *p*. The models were (1) a straight line with a slope of zero, (2) a straight line of any slope, and (3) a quadratic. In the first variation (ME1), we fit the chosen model and level of *p* at the lower levels (i.e. county and county group). In the second variation (ME2), we fit the chosen model at the survey unit level to predict an overall $\hat{\beta}$ (a Tx1 vector described below, where T is the number of years in the sample, usually equal to 5). This leads directly to a simple updating vector $\mathbf{U}$, found by multiplying the inverse of each element of $\hat{\beta}$ by the fifth element of $\hat{\beta}$. Then:

$$\hat{V}_{ME2} = (1'\mathbf{A}_T)^{-1}\left(\left(DIAGRV(\mathbf{A}_T)\right)\mathbf{V}_T\right)'\mathbf{U}$$

where:

$\mathbf{A}_T$ = a Tx1 vector of total area sampled at each time,

$\mathbf{V}_T$ = a Tx1 vector of basal area estimates for each time,

$\mathbf{1}$ = a Tx1 vector of ones, and

$\mathrm{DIAGRV}(\mathbf{A}_T)$ = a function that places a Tx1 vector $\mathbf{A}_T$ into the diagonal of a TxT matrix of zeroes.

We then evaluated the estimators for how well they predicted the "true" county level and county group level basal areas for 1998 from the 1994 to 1998 sample, under a squared error loss function. Initially, we conducted a case study yielding a unique solution for the moving average estimator (MAE) and each variation of mixed estimation (ME1 and ME2). The squared error calculated for these methods is simply the mean of the squared difference of each estimate by county and county group from the truth for that county or county group.

Finally, we performed a simulation, assuming that spatial correlation between plots was unimportant. The plots were randomly rearranged 1,000 times and then grouped into simulated, approximately equally sized, "counties." We varied the number of counties from 10 to 50 to see what effects sample size would have on the ranking of the estimation approaches. When the plots were grouped into 10 counties, there was an average of 235.3 plots in each county, (actually 235 plots in 7 counties and 236 plots in 3 counties). At the other extreme, when there were 50 counties, there were 47 plots in 47 counties and 48 plots in 3 counties. We calculated the mean difference and mean squared difference from the "truth" over the 1,000 random arrangements of the 2,353 plots. We defined the "truth" as the population mean of each simulated county at time 5.

The simulation results led to the suspicion that the assumption of spatial independence between plots was weak. Therefore, in an attempt to detect spatial trends, we performed median polishes of the "true" population plot data for 1998 aggregated at five different scales of a square grid (50, 40, 30, 20, and 10 miles on a side). We conducted the median polishes in the cardinal directions (north-south and east-west). At two scales, a strong north-south trend was indicated. The cell sizes for the first of these scales were slightly larger than the average county size (a square grid with 30 miles on a side), resulting in 31 filled cells. The second scale was 20 miles on a side, resulting in 57 filled cells. The results for 50, 40, and 10-mile grids are not presented because they did not show any spatial trends. Subsequent to the median polishes, we calculated the variograms for the 30 and 20-mile grids of both the original data and the residuals.

## RESULTS

For the case study, table 1 shows the mean difference from the truth over all counties and county groups for MAE, ME1, ME2, and the mean of panel 5 (P5M). Table 2 shows the corresponding mean squared differences. The panel 5 mean is included because panel 5 is the portion of the sample that observes only the population partition of interest (that is, tree basal areas during 1998). In the case study, the mean difference is not a true measure of model bias, but can be an indication of model bias. Note that two of the estimators have roughly the same mean difference at both the county and county group levels, leading us to suspect that the respective levels may reflect the true level of bias in these estimators. Of these two, MAE shows the largest absolute difference. Due to the increasing trend in the variable of interest, all values for the moving average were low. The magnitude of the absolute mean difference is close to zero for ME2. When going from the county to the county group level, the large reduction in magnitude of absolute mean difference for the other two estimators appears to be more a result of decreasing variance than of bias. Of course, because P5M is design unbiased and does not rely on a time dependent model, we know that this is the case for P5M.

**Table 1—Mean difference—case study**

| Estimator | County | County group |
|---|---:|---:|
| Moving average estimator | -2.026 | -1.919 |
| Panel 5 mean | 2.293 | .024 |
| Mixed estimator, variation 1 | -2.367 | .027 |
| Mixed estimator, variation 2 | .078 | .159 |

**Table 2—Mean squared difference—case study**

| Estimator | County | County group |
|---|---:|---:|
| Moving average estimator | 12.586 | 4.305 |
| Panel 5 mean | 314.065 | 8.963 |
| Mixed estimator, variation 1 | 98.350 | 2.470 |
| Mixed estimator, variation 2 | 9.128 | .513 |

In table 2, ME2 shows the lowest mean squared differences overall. In addition, ME1 has a higher variance than MAE and ME2 at the county level, because the sample sizes were too small at the county level to fit the model. Two observations support this statement. First, ME1 behaves better at the county group level than at the county level. Second, ME2, in which the model was fit at the survey unit level and then applied at the lower levels, works well even at the county level.

The second part of the study, the simulation in which we randomly rearranged the plots, has led to unexpected, albeit explainable results. The top graph in figure 2 shows the mean squared difference from the truth for the 1,000 random arrangements of the 2,353 plots after being grouped into 10 to 50 counties; the bottom graph gives the corresponding mean differences. Note that although the MAE of time 5 basal area still displays the expected bias, it now compares favorably, in terms of mean squared error, with ME2. ME2 can be expected to work best if the individual county basal areas at times 1 through 4 have the same values relative to the county basal areas at time 5 as occurs globally over the entire survey unit. In a heterogeneous population, this condition is more likely to occur if similar plots are spatially collocated. ME1 requires that the

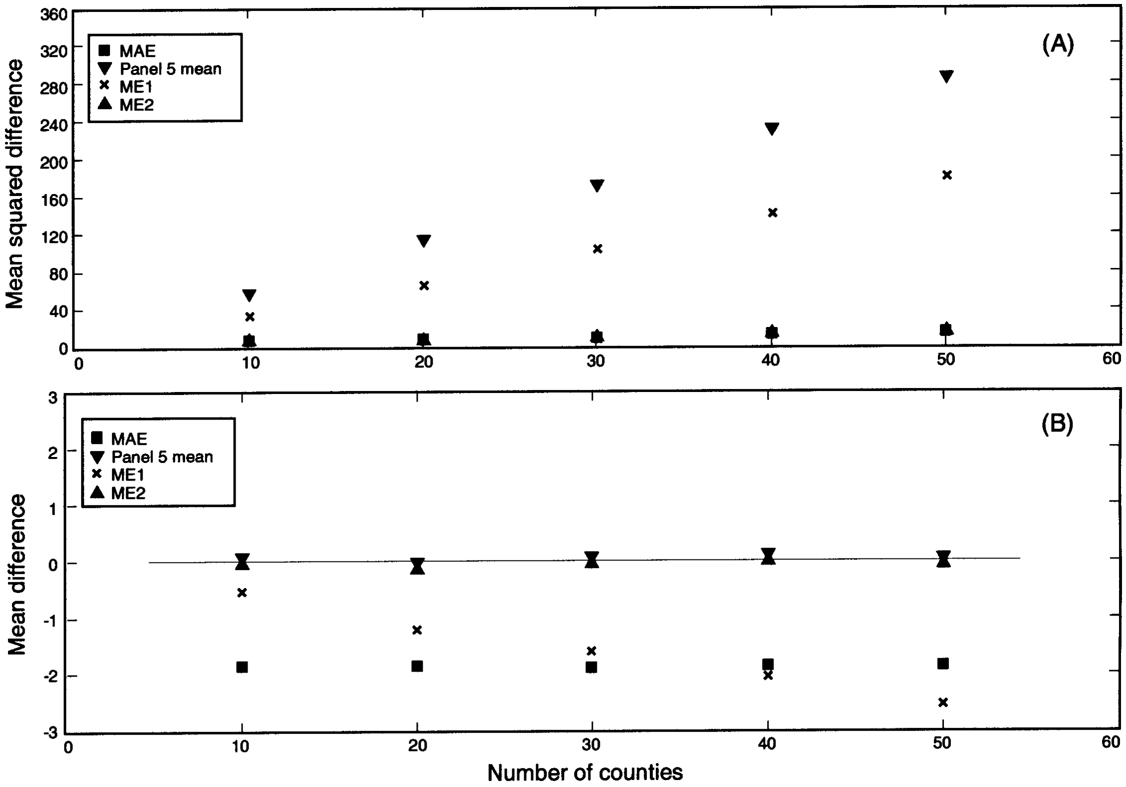


Figure 2—Mean squared difference (A) and mean difference (B) from the truth for 1,000 random arrangements of the 2,353 plots after being grouped into 10 to 50 counties; i.e., sample size per county is decreasing from left to right.

selected model be fit at the county level. This would be advantageous if plots within counties were more homogeneous than between counties, and if there were a sufficient number of plots in each county. These observations have led to a search for spatial trends in the data.

In the top graph, figure 3 gives the aggregated mean population basal areas for the data within grid cells of 20 miles on a side; in the bottom graph, it gives the residuals, row, column, and all effects following a median polish of this data. Tukey (1977) and Cressie (1991) explain the median polish (also known as median sweep). Figure 4 gives the corresponding information following a coarse mapping with 30-mile grid cells. In the bottom graph of both figures, the row effect (far right column, save for the "all" effect at the bottom) is a large, positive number at the top and a not-quite-as-large, negative number at the bottom. Although neither vector strictly decreases from top to bottom, a trend does appear likely.

The top graph of figure 5 shows the classical estimates, as well as the Cressie-Hawkins robust estimates (Cressie and Hawkins 1980), of the north-south variograms for the data in the top graph of figure 3. The bottom graph of figure 5 displays the corresponding estimates for the residuals in figure 3. Likewise, figure 6 provides the same estimates for the data in figure 4. Figure 5 illustrates the classic argument that the median polish removes spatial correlation from the data, as the estimated variograms of the residuals are decidedly flatter than those of the data. At first blush, figure 6 seems to give quite the opposite impression; that is, unless one ignores the values for the lag of six (equal to 180 miles). It is appropriate to ignore this lag since only a single observation supported it and one end of the interval happens to be in a row with only two observations. Ignoring the lag 6 values, we see that the plots for the residuals are slightly flatter than the plots for the data. In toto, figures 3 through 6 show that there is a north-south trend observable at scales greater than or equal to 100 miles.

## CONCLUSIONS

The FIA annual inventory design will provide a set of sample observations of forest attributes that is thoroughly diffused through space and time. This will allow estimation of forest attributes for an almost-infinite set of subdomains of interest. FIA cannot provide this extremely large set of potential estimates; however, FIA could and should provide a reasonable set of tools within an estimation system to users accessing the data over the Internet. Such a system would be more useful if it made reasoned use of data from outside the domain of interest (i.e. the space-time cube defined by a user) when that domain of interest is too small to contain enough observations for the usual sample-based estimators.

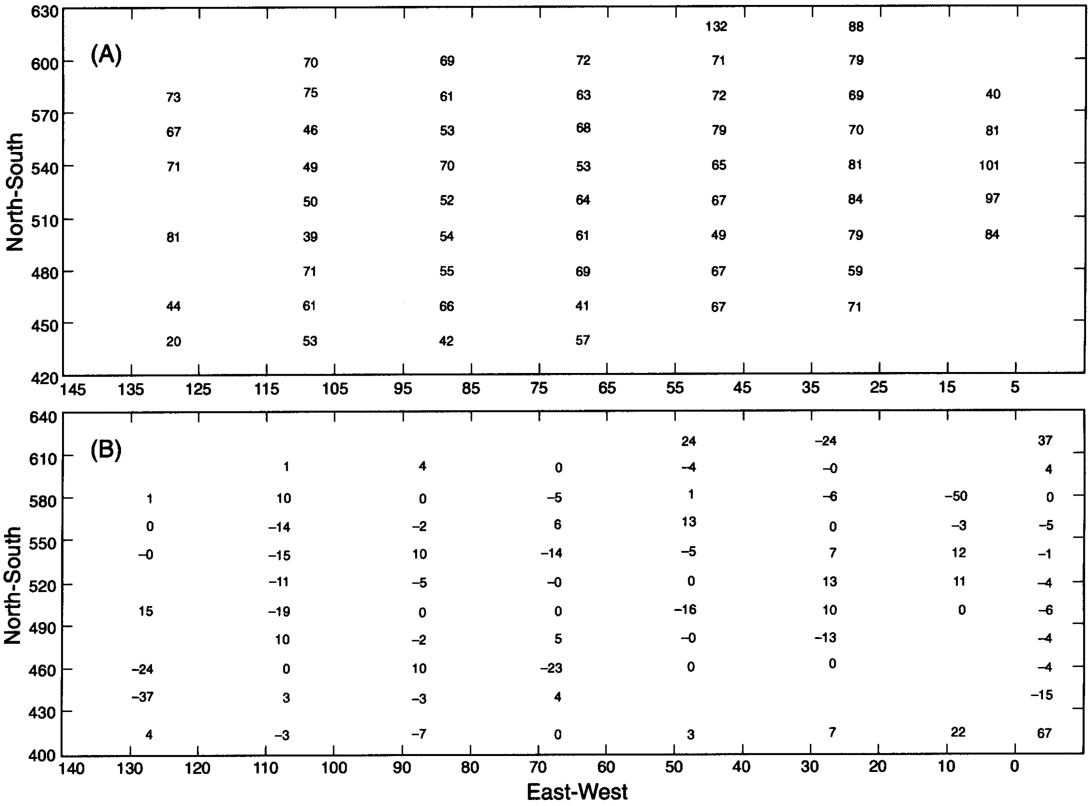


Figure 3—Aggregated mean basal area (A) of the “true” population simulated from Georgia, Survey Unit 1 Forest Inventory and Analysis data following a coarse mapping with a grid size of 20 miles on a side, plotted by an arbitrary coordinate system. The bottom graph (B) shows the overall effect (bottom right), the column effects (remainder of the bottom row), the row effects (remainder of the right column), and the residuals (remaining values) following a median polish of the data in the top graph.

This study examines methods of making estimates over a smaller domain than the sample within that domain will actually support. The methods that use outside information in different ways, MAE, ME1, and ME2, yield substantial improvement in terms of squared error loss over P5M. None of the alternative estimators, as applied to the small-areas, however, can be shown to be design unbiased. In the presence of increasing or decreasing trend, the alternatives to the simple moving average have the potential of being model unbiased. For basal area, and presumably all variables that are likely to exhibit trends over the 5-year measurement period, even simplistic approaches to modeling the trends can result in significant reductions in MSE over the simple moving average.

These estimators (MAE, ME1, and ME2) use the same information in different ways. That information comes only from the FIA annual inventory data, although 80 percent comes from outside of the domain of interest. Therefore, comparisons between the methods are direct. On the other hand, some methods that we have not discussed here benefit from a rich history of external growth and yield research. Mixed estimation, in general, represents a much lower investment in human resources both initially and in the long term than common industrial methods, which use growth and mortality equations to update plot data. This latter approach would be difficult for FIA to use because appropriate growth models do not exist for many condition classes of interest, and those that do exist would have to undergo thorough testing for use in this context. In addition, to ensure that the forest populations are not moving away from those upon which the models were built, the growth model predictions would have to be constantly monitored.

There are at least two ways to view any differences between Part 1, the case study, and Part 2, in which the plots were randomly rearranged 1,000 times. Conducted over a broader range of conditions, the simulation, on the one hand, should be considered a more robust test of the behaviors of the respective estimators. On the other hand, the simulation disfavors estimators that draw strength from spatial correlation, if that correlation exists in real populations. Any spatial correlation inherent in the data remained intact in the case study but not in the simulation. The results support this second viewpoint on a number of fronts. For instance, the moving average estimator moved up in ranking during the simulation relative to the case study. Since spatial correlation would lead to stronger time trends within counties, and the moving average estimator would be at a disadvantage in the presence of a time trend, a simulation ignoring potential spatial correlation might

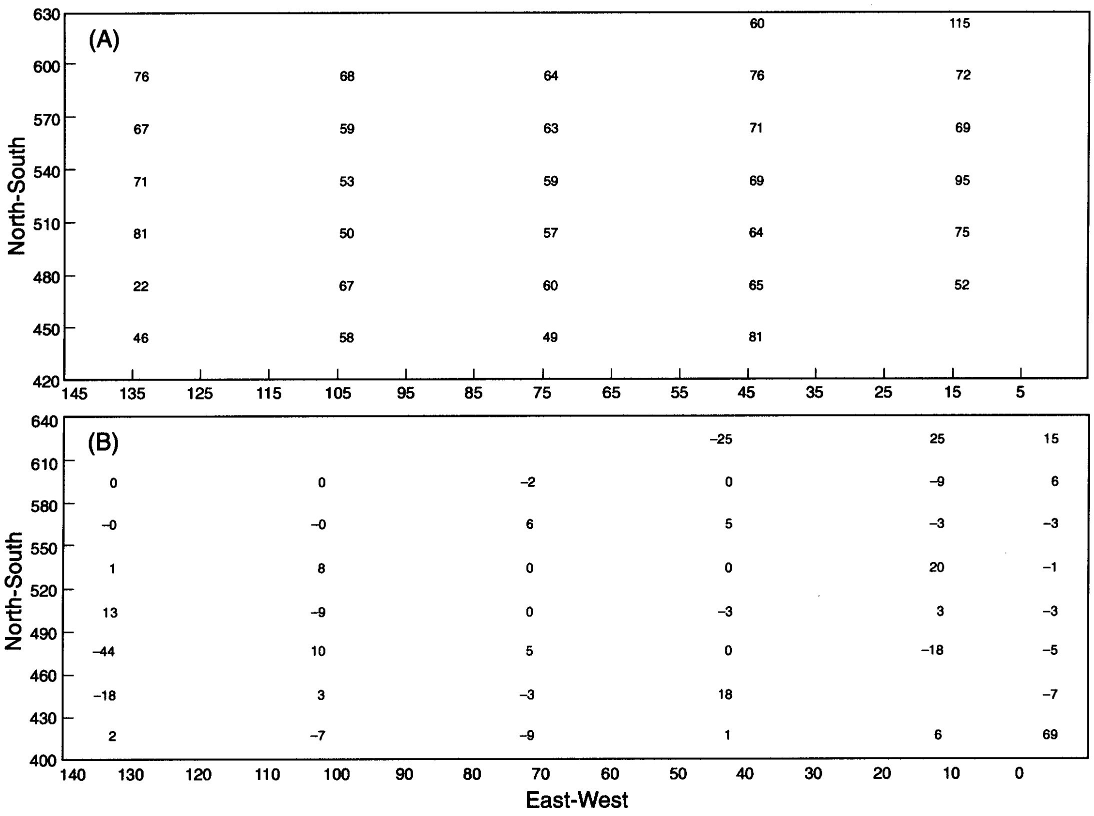


Figure 4—Aggregated mean basal area (A) of the “true” population simulated from Georgia, Survey Unit 1 Forest Inventory and Analysis data following a coarse mapping with a grid size of 30 miles on a side, plotted by an arbitrary coordinate system. The bottom graph (B) shows the overall effect (bottom right), the column effects (remainder of the bottom row), the row effects (remainder of the right column), and the residuals (remaining values) following a median polish of the data in the top graph.

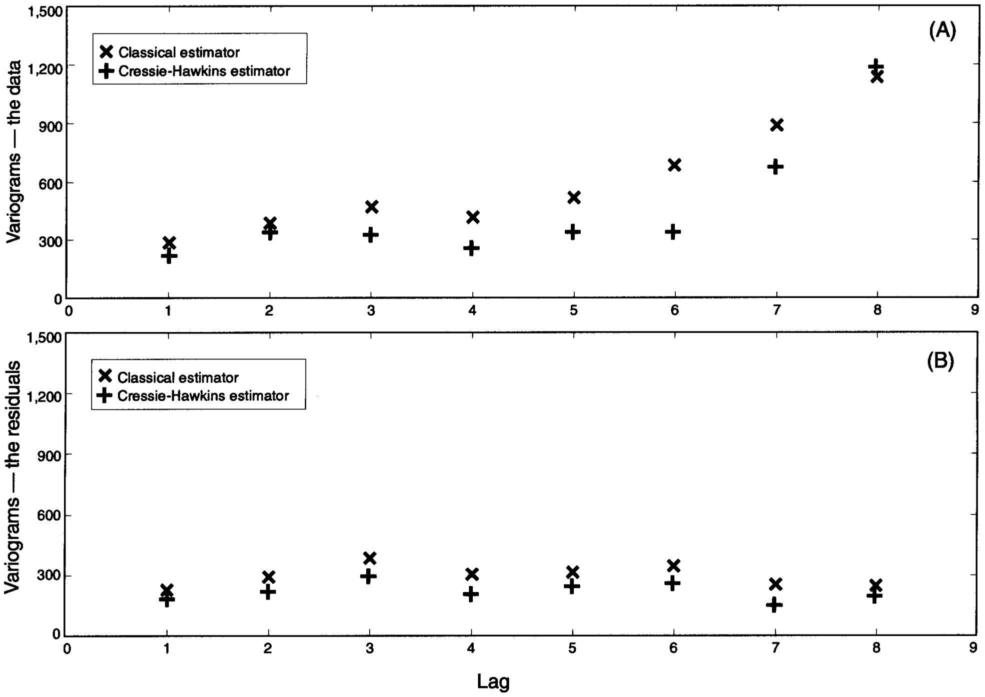


Figure 5—Classical and Cressie-Hawkins robust estimators of the north-south variograms for a grid size of 20 miles on a side, for the aggregated data (A) and the residuals (B) following a north-south, east-west median polish.

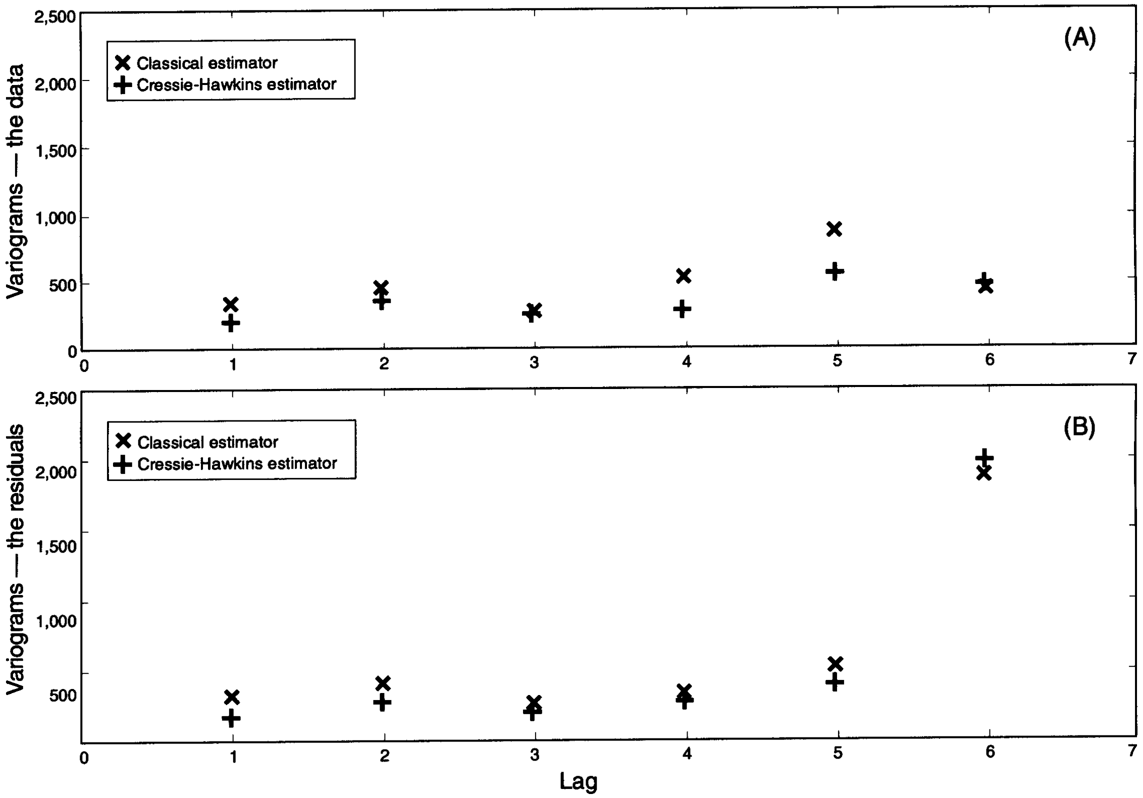


Figure 6—Classical and Cressie-Hawkins robust estimators of the north-south variograms for a grid size of 30 miles on a side, for the aggregated data (A) and the residuals (B) following a North-South, East-West median polish.

garble any time trend enough to favor the moving average estimator. Similarly, if a strong global time trend existed, in the presence of strong spatial correlation at the county level, the two applications of mixed estimation would benefit. Therefore, we should not be surprised if they fare better in the case study than in this particular simulation.

The spatial analysis established that the basal area data did contain spatial correlation at relevant scales. Other survey units, of a similar size and diversity, could also exhibit spatial trends for this and probably other variables. Therefore, modeling for both the potential spatial trends as well as the potential temporal trends within survey units could benefit small-area estimates. This gives further credence to the call to FIA for the development of a GIS based estimation system with the ability to adapt to user-defined areas and periods of interest.

## REFERENCES

**Anon**. 1992. Report of the blue ribbon panel on forest inventory and analysis, Washington, DC. Available from: Frank Roesch, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC.

**Anon.** 1998. Report of the second blue ribbon panel on forest inventory and analysis, Washington, DC. Available from: Frank Roesch, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC.

**Cressie, N.** 1991. Statistics for spatial data. New York: John Wiley. 900 p.

**Cressie, N.; Hawkins, D.** 1980. Robust estimation of the variogram, I. Journal of the International Association for Mathematical Geology. 12: 115–125.

**Tukey, J.** 1977. Exploratory data analysis. Reading, MA: Addison-Wesley. 688 p.

**Van Deusen, P.** 1999. Modeling trends with annual survey data. Canadian Journal of Forest Research. 29(12): 1824–1828.

# ALTERNATIVES TO THE MOVING AVERAGE[1]

**Paul C. Van Deusen[2]**

**Abstract**—There are many possible estimators that could be used with annual inventory data. The 5-year moving average has been selected as a default estimator to provide initial results for states having available annual inventory data. User objectives for these estimates are discussed. The characteristics of a moving average are outlined. It is shown that moving average characteristics don't always coincide with user objectives. Alternative estimators are proposed that may have more desirable characteristics than the simple moving average.

## INTRODUCTION

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service is shifting from doing periodic inventories in each state to implementing an annual forest inventory (USDA Forest Service 1999) where a percentage of the plots are measured each year. The design is intended to provide annual systematic coverage of each state and to generally provide the same amount of information each year. For the most part, the annual inventory uses the same plot system that existed under the periodic design, and one could argue that the annual system merely changes the timing of plot visits. In fact, the annual inventory is a response to changing user needs and therefore represents a major transition for FIA.

A widespread user desire for more timely data is arguably the driving force that led to the 1998 Farm Bill directive for annual inventories. This is documented in 2 Blue Ribbon Panel reports, BRP I and BRP II (American Forest Council 1992, American Forest and Paper Association 1998). BRP I called for shortening the cycle between periodic surveys from 10 to 5 years. This shortened cycle was never achieved and cycles averaged 10 years or more when BRP II convened in 1997. The BRP II call for an annual survey led to the 1998 Farm Bill legislative mandate for annual surveys.

## WHAT THE USER WANTS

Most users want timely data and timely estimates. They want estimates that reflect current values in accordance with the current data that an annual inventory provides. It follows that estimates of per acre values are needed for year t, where t can denote any year beginning with annual inventory implementation up through the current year. Likewise, estimates of change between any 2 years should be available. It goes without saying that the user also wants current estimates of area by forest type, but that is a subject for another paper.

These needs can be quantified by letting the implementation year for a state be year 1 and the current year be T. Call the true per acre value $\ _t$. Therefore the user wants

- Current per acre estimates, $\hat{\ }_t$, for years t=1,...,T.
- The related variance estimate V($\hat{\ }_t$), for t=1,...,T.
- Annual change estimates, $\hat{\ }_t - \hat{\ }_{t-k}$ for k=1,...,t–1
- Estimated variance of change, V($\hat{\ }_t - \hat{\ }_{t-k}$).

## WHAT THE MOVING AVERAGE ESTIMATES

The 5-year moving average is equivalent to taking all plot measurements from the last 5 years in a state and averaging them together. For years t-4 through t this can be written as

$$MA_{t-4,t} = \sum_{j=t-4}^{t} w_j \ \bar{y}_j$$

where $\bar{y}_j$ is the average of all plot values measured in year *j*, and $w_j$ is a weight such that $\sum w_j = 1$. The plan for the annual inventory is to assign plots to panels and to measure 1 panel per year. Therefore, $\bar{y}_j$ can also be called the panel mean. The weight, $w_j$, ensures that each panel is weighted according to the proportion of the total plots it contains. With an exact 20 percent sample, $w_j$ =0.2.

The panel mean is unbiased for the true underlying value, $\ _t$, and we can write

$$\bar{y}_t = \ _t + e_t \qquad (1)$$

where $e_t$ is a random error term. It follows that the expected value of the moving average is

$$E(MA_{t-4,t}) = \sum_{j=t-4}^{t} w_j \ \ _j \qquad (2)$$

Therefore $MA_{t-4,t}$ estimates the true average over the last 5 years and is not an unbiased estimate of the current value, $\ _t$. This isn't what most users want, but it is similar

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Principal Research Scientist, National Council for Air and Stream Improvement, 600 Suffolk Street, Lowell, MA 01854.

to what was done under the old periodic design. It often took 3 or more years to complete the full periodic survey in a state, and state estimates were obtained by averaging all plots together as if they had been measured in the same year. This estimate was then assumed to represent the value at the final measurement year. Based on this precedent, one could conclude that the 5-year moving average is "good enough" even though it isn't estimating the current year value.

The variance of the moving average is easy to derive as

$$V(MA_{t-4,t}) = \sum_{j=t-4}^{t} w_j^2 \; V(\bar{y}_j) \tag{3}$$

where $V(\bar{y}_j) = {}^{2}_{j}/n_j$ , ${}^{2}_{j}$ is estimated from the between plot variance within the panel and $n_j$ is the number of plots in the panel measured in year *j*. Therefore, the expected value and the variance of the moving average are well-defined, and both are easy to estimate.

## ESTIMATING CHANGE WITH THE MOVING AVERAGE

Change and trend are more important to many FIA users than current status. FIA is committed to producing official state-level reports every 5 years, but users will not wait for 10 years to assess trend. Since the moving average is currently considered to be the default estimator it makes sense to look at the difference between 2 moving average estimates. Suppose we are at year 6 of the annual survey and want an estimate of change since year 5. The difference between the year 6 and year 5 moving average is

$$MA_{2,6} - MA_{1,5} \begin{aligned} &= \frac{1}{5}\left((\bar{y}_2+\bar{y}_3+\bar{y}_4+\bar{y}_5+\bar{y}_6)-(\bar{y}_1+\bar{y}_2+\bar{y}_3+\bar{y}_4+\bar{y}_5)\right) \\ &= \frac{1}{5}\left(\bar{y}_6-\bar{y}_1\right) \end{aligned} \tag{6}$$

Equation (6) shows that taking the difference of consecutive moving averages cancels out much of the data. Remember that the goal of the annual survey is to cycle through the plots every 5 years, so the plots measured in year 1 will generally be remeasured in year 6. Therefore, equation (6) shows that simple 5-year moving average change estimates give the average of the 5-year change in the current panel. The other 4 panels (80 percent of the plots) measured over the past 5 years are ignored. This is clearly an undesirable situation and less variable estimators could be constructed by using more of the data.

The variance of the moving average change estimator (equation 6) is

$$V(MA_{2,6} - MA_{1,5}) = \frac{1}{25}\left(V(\bar{y}_1) + V(\bar{y}_6) - 2COV(\bar{y}_1, \bar{y}_6)\right)$$

so the covariance between remeasured plots will reduce the overall variance. However, this doesn't justify ignoring 80 percent of the plots.

## ALTERNATIVES TO THE MOVING AVERAGE

Alternatives to the moving average are needed because (1) the moving average is not unbiased for current status and (2) moving average annual-change estimates ignore 80 percent of the plots under a 5-panel annual inventory design. There are many alternatives that could be considered. Here the focus will be on mixed-estimation methods (Theil 1971), but multiple imputation (Rubin 1987) and double sampling for regression will also be briefly discussed.

### Multiple Imputation

Multiple imputation (Van Deusen 1997, Reams and Van Deusen 1999, Roesch and Reams 1999) uses the intuitively appealing approach of filling in values for unmeasured plots and then applying standard complete data analysis methods. Imputation can be performed by database lookup (hotdeck methods), with regression estimates, or with more elaborate modeling efforts. Single imputation is a special case where only one possible value is imputed for each missing value. Single imputation usually requires complex procedures to properly estimate variance. This makes it tempting to treat imputed values as if they are real which will lead to under-estimating the variance. Multiple imputation requires the imputer to incorporate variability into the imputations, which leads to a simplified variance estimation process for the analyst.

Multiple imputation can work for variables that are difficult to model but are amenable to database lookup. Examples of such variables include: number of snags, Red-cockaded Woodpecker (*Picoides borealis*) nests, or disturbance status. A disadvantage is that multiple datasets must be stored (say m) and each analysis must be repeated m times. Typical users might find this confusing, so multiply imputed datasets will probably not become an official FIA product in the near future.

### Double Sampling for Regression

Double sampling for regression (DSR) can be viewed as a single imputation procedure. Intuitively, single imputation methods should place different weights on imputed values and real data. DSR (Cochran 1977, Fairweather and Turner 1983, and Hansen 1990) does this by incorporating predictions via the following regression equation

$$\hat{}_t = \bar{y}_t + a\left(\bar{X} - \bar{x}\right)$$

where $\bar{y}_t$ is the mean from the year t panel, a is a regression coefficient, $\bar{X}$ represents concomitant information from the 4 panels not measured in year t, and $\bar{x}$ is concomitant information from the year t panel. Usually one refers to a large and a small sample with DSR, where the small sample includes the hard-to-measure variable, y, and the easy-to-measure variable, x. Only the easy-to-measure variable is measured in the large sample. For the annual inventory application, the current panel is the small sample and the other 4 panels constitute the large sample.

A necessary assumption is that the large and small samples represent the same population, and therefore $E(\bar{X}) = E(\bar{x})$. However, this may not be true with the annual inventory. The values for $\bar{x}$ from the current panel must come from measurements made 5 years earlier, whereas $\bar{X}$ comes from measurements made 4, 3, 2, and 1 years earlier on the other 4 panels. Therefore, there are systematic differences between the small and large sample x's, and it is likely that $E(\bar{X}) \neq E(\bar{x})$. Regardless, double sampling for regression could be used as a single imputation technique, but some validation studies should be conducted first. Also, variance of DSR change estimates would be difficult to derive such that auto correlation is correctly handled.

## Mixed Estimation Methods

Mixed estimation (Theil 1971) offers a flexible time series approach that leads to model-unbiased estimates of current status, change estimates over any time interval, and variance estimates. There are numerous variations that can be considered (Van Deusen 1996, 1999) and a subset of the possibilities is presented here.

Generally, a mixed estimator is defined by an observation equation and a transition equation, where the transition equation is analogous to the Bayesian prior distribution. Although mixed estimation has a Bayesian flavor, it is a cross between Bayesian and frequentist approaches. The observation equation used here is

$$\bar{y}_t = \quad {}_t + e_t \tag{7}$$

where $e_t$ is an independent random error with mean 0 and variance $\sigma_t^2 / n_t$. Consider the following three transition equations

$$ {}_t - \quad {}_{t-1} = v_t \tag{8a}$$

$$ {}_t - 2 \quad {}_{t-1} + \quad {}_{t-2} = v_t \tag{8b}$$

$$ {}_t - 3 \quad {}_{t-1} + 3 \quad {}_{t-2} - \quad {}_{t-3} = v_t \tag{8c}$$

where $v_t$ is an independent random error with variance $p\sigma_t^2/n_t$ and $p$ is a parameter that is estimated from the data. As $p$ gets larger, the influence of the transition equation diminishes and the mixed estimator approaches the mean for each panel. Each transition equation leads to a mixed estimator with somewhat different characteristics. Likewise, each equation represents a different prior assumption about how ${}_t$ is related to ${}_{t-1}$. The transition equations (8a-c) constrain the first, second, and third differences of the *s* and lead to progressively smoother estimates of trend. The transition equations also state that past values give an indication of current values. This seems eminently plausible, since the forest won't change much from 1 year to the next, barring catastrophe. Transition equation (8b) represents an intermediate smoothness constraint and would make a reasonable choice for FIA purposes.

The estimation process is best described using matrix notation. It follows that there is no particular reason to use only the most recent 5 years of data. The equations stay the same regardless of how much data are used, and the estimates will usually improve with more data. The matrix estimation equations for years 1 through T are

$$\hat{\ } = \left[ \sum{}^{-1} + \frac{1}{p} R' \quad {}^{-1} R \right]^{-1} \sum{}^{-1} \bar{Y} \tag{9a),}$$

and

$$V\left(\hat{\ }\right) = \left[ \sum{}^{-1} + \frac{1}{p} R' \quad {}^{-1} R \right]^{-1} \tag{9b}$$

where $\hat{\ } = \lfloor {}_{1,} \ {}_{2}, \ldots, \ {}_{T} \rfloor$ , $\sum{}^{-1} = DIAG\left( {}^2_1/n_{1,} \ {}^2_2/n_2, \ldots, \ {}^2_T/n_T \right)$ , R depends on the selected transition equation, contains transition variances on the diagonal, and $\bar{Y} = [\bar{y}_{1,} \bar{y}_2, \ldots, \bar{y}_T]$. More details can be found in Van Deusen (1999).

The important point here is that relatively simple equations are available to estimate current status and its variance. Trend estimates are also available, since equation (9a) provides estimates of ${}_1$ through ${}_T$. To estimate the the change from time t–k to t, simply use $\hat{\ }_t - \hat{\ }_{t-k}$. A complete covariance-matrix is available from (9b), which makes it possible to estimate the variance of change. Thus, mixed estimators provide very general capabilities for estimating status and trend.

## SUMMARY

FIA is replacing the periodic inventory with an annual inventory. Even though data will be acquired annually, FIA plans to produce official estimates for each state every 5 years using a 5-year moving average. The 5-year moving average has been selected as the default estimator for the annual survey, in part because it seems easy to understand and compute. Users want FIA procedures that are statistically valid, not unnecessarily complicated, and that meet their needs. The MA is statistically valid and easy to implement, but it doesn't fully meet user needs. In particular, the MA does not estimate current status at time t. Regardless, it is similar to what was done under the old periodic design and might be an adequate approximation of current status.

Users also want estimates of trend between any 2 years, say t and t–k. In particular, they should be able to obtain change estimates between the current year and the previous year. It was shown in equation (6) that the difference between consecutive 5-year moving averages gives an estimate of the average annual growth over the last 5 years using only 20 percent of the plots. Therefore, this cannot be the best trend estimator available.

Mixed estimators were discussed that do provide estimates of current status and trend between any 2 years. These estimators are more complex than the MA, but will give users a wider array of estimates. For a few years following annual inventory implementation, the moving average may be sufficient. In the long run, FIA should give serious thought to finding alternatives to the MA.

## REFERENCES

**American Forest Council**. 1992. The report of the first blue ribbon panel on forest inventory and analysis. Washington, DC.

**American Forest and Paper Association**. 1998. The report of the second blue ribbon panel on forest inventory and analysis. Washington, DC.

**Reams, G.A.; Van Deusen, P.C.** 1999. The southern annual forest inventory system. Journal of Agricultural Biological & Environmental Statistics. 4(4): 346–360.

**Roesch, F.A.; Reams, G.A.** 1999. Analytical alternatives for an annual inventory system. Journal of Forestry. 12: 33–37.

**Rubin, D.B.** 1987. Multiple imputation for nonresponse in surveys. Wiley.

**USDA Forest Service**. 1999. A strategic plan for forest inventory and monitoring. Washington, DC.

**Van Deusen, P.C**. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27: 379–384.

**Van Deusen, P.C**. 1997. Alternative sampling designs and estimators for annual surveys. In: Hansen, M.; Fairweather, S., eds. Proceedings of international conference on the inventory and monitoring of forested ecosystems. Boise, ID; 1998 August; USDA Forest Service General Technical Report.

# Information Management

*Moderator:*

**Paul Van Deusen**
National Council for Air and Stream Improvement

# THE NEW SOUTHERN FIA DATA COMPILATION SYSTEM[1]

**V. Clark Baldwin, Jr., and Larry Royer[2]**

**Abstract**—In general, the major national Forest Inventory and Analysis annual inventory emphasis has been on data-base design and not on data processing and calculation of various new attributes. Two key programming techniques required for efficient data processing are indexing and modularization. The Southern Research Station Compilation System utilizes modular and indexing techniques applied with standard Oracle tools. We present the unit's approach and describe the challenges encountered as a guide for others embarking on the same complex computational journey.

## INTRODUCTION

The new southern Forest Inventory and Analysis (FIA) Inventory Compilation System includes data processing activities from field collection to processed data output. For convenience, the process is divided into three phases: (1) data collection, (2) calculation of derived data, and (3) output of the compiled data. Although relatively straightforward and manageable, the computing process is complex because of the changes in sampling design, addition or redefinition of variables measured, and new or additional outputs required in the shift from periodic to annual inventories.

The complexities begin at the field collection phase, where the data recorder software must be capable of managing input from various sampling scenarios. The first scenario is the remeasurement of the periodic/prism sample with the initial inventory of the annual/mapped sample overlaid on the prism/periodic sample (USDA 1967, 1998). The second scenario is the remeasurement of the annual/mapped sample (USDA 1998). The third scenario is the remeasurement of a fixed-area sample (e.g., 1998 Kentucky plots) with a new initial annual/mapped sample at a different location (USDA 1999). At the present time, nine States are downloading previous inventory data, collecting new data, and transmitting data. While the States are actively converting data from periodic to annual inventories in different stages and ways, data from no two States are similarly compiled. Sophisticated software downloads and formats data from a variety of sources, such as flat-files (Anonymous 2000) and Oracle™ database tables (Koch and Loney 1997), into a common Microsoft™ (MS) Access format (Anonymous 2000). The field crews operating in the remeasurement of the annual/mapped mode can query the Oracle database tables through any remote Internet connection and build the county's previous inventory data interactively on the personal data recorder. Crews operating in the other modes download preformatted historic data to the personal data recorder. They can transmit and capture data through a dial-up connection to a server in Starkville, MS, which loads the data into a set of Oracle relational production database tables.

The amount of data flowing concurrently from nine States is part of the "tidal wave of data" referred to during this conference. From 28,101 sample plots, 21,161 were submitted via data recorder; and 6,940 were submitted by other means. The number of trees from these plots totaled 605,281. Data flowing so quickly into the system caused a logjam. To accommodate this massive data input, we have designed a compilation system and data-flow method to clear the logjam and ensure efficient data processing.

The Southern Compilation System had to overcome many challenges during system development; these involve algorithm development and programming and required the creation of immediate solutions. The difficulties centered in the areas of (1) computer system and Oracle software performance, (2) input data, (3) change accommodation, (4) area reconciliation, and (5) moving-average estimation.

## THE SOUTHERN COMPILATION SYSTEM

### Concepts

The data compilation system must process current and previous data for all of the scenarios described above. The Oracle database system accomplishes these complicated tasks using a relational database format. The three major groups of relational tables within our Compilation System are (1) Production Tables, (2) Regional Tables, and (3) National Tables.

**Indexing—**The Compilation System hierarchy was designed with primary index keys that link the database tables to allow quick and easy access to any element. The primary index key, PIX_ID, grows as the level of the table increases within the hierarchy tree. The primary index key is similar to a serial number and can be used to locate an item in any table, based upon its relation to any other table. Levels of processing are State, county, cycle, panel, plot, and individual tree. The following programming code example illustrates the process:

Production Table Prod_Plot plot level PIX_ID = 1300103141
where State = 13, County = 001, Plot = 031, Cycle = 4, and Panel = 1.
Regional Table Inventory tree level PIX_ID =
13001031413010 where State = 13, County = 001, Plot = 031, Cycle = 4, Panel = 1 , Subplot = 3, and Tree Number = 10.
National Table Tree level PIX_ID = 13001031413010 where
State = 13, County = 001, Plot = 031, Cycle = 4, Panel = 1, Subplot = 3, and Tree Number = 10.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester and Mensurationist, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804, respectively.

We establish an SQL query by joining the three tables, using a substring of the PIX_ID key. The following simple SQL statement can query the attributes of species code, diameter from the Tree table, past diameter at breast height (d.b.h.) from the Inventory table, and remeasurement period from the Plot table:

Select **Tree.spcd, Tree.dia, Inventory.Pastd.b.h., Plot.Remper** from **Tree, Inventory, Plot** where **Tree.pix_id = 13001031413010** and **Inventory.Pix_Id = Tree.Pix_Id** and **Plot.Pix_Id = substr(Tree.Pix_Id,1, 10)**

By using the PIX_ID primary index key, we can reduce paragraphs of Structured Query Language (SQL) code to a statement that can locate individual items, such as one tree. Without the PIX_ID primary index key, the query to select the exact same attributes would be:

Select **Tree.Spcd, Tree.Dia, Inventory.Pastd.b.h., Plot.Remper** From **Tree, Inventory, Plot** Where **(Tree.Statecd = 13** and **Tree.Countycd = 1** And **Tree.Plot = 31** And **Tree.Cycle = 4** And **Tree.Panel = 1** And **Tree.Subplot = 3** And **Tree.Tree = 7)** And **(Inventory.State = Tree.Statecd** And **Inventory.County = Tree.Countycd** And **Inventory.Location = Tree.Plot** And **Inventory.Cycle = Tree.Cycle** And **Inventory.Panel = Tree.Panel** And **Inventory.Point_Number = Tree.Subplot** And **Inventory.Tree_Number = Tree.Tree)** And **(Plot.Statecd = Tree.Statecd** And **Plot.Countycd = Tree.Countycd** And **Plot.Plot = Tree.Plot** And **Plot.Cycle = Tree.Cycle** And **Plot.Panel = Tree.Panel)**

In complexity and performance, the PIX_ID process is less complicated and executes faster than code written without this feature. Each tree has a unique PIX_ID that allows the isolation, tracking, and processing of any individual tree throughout the entire system.

**Modules—**The Compilation System uses modules and functions of Oracle PL/SQL code (Urman 1996) to break down the complex process scenarios into small tasks. To manage maintenance, debugging, and change, each module or function within the system was kept small and limited to one task. Thus, for example, if a volume equation's coefficients change, only the function for volume needs modification or replacement. For debugging a data problem, the data need only be run through an appropriate module or function for the problem to be isolated. Most modules and functions can be executed at any level of processing and in any order. An individual tree with a data problem may be reprocessed using the module or modules in question without reprocessing the whole data set. However, area data can be processed only at the county level due to the nature of the data. Some possible levels of processing are State, county, unit, cycle, panel, plot, and individual tree. In this semi-automated design mode, as a plot clears the internal edit, processing modules at the plot level begin for all of the data on the plot. When the last plot of a county clears the internal edit, the area processing modules are triggered to process the county data. This semi-automated mode provides that all of the data for a processing level will be complete when the last plot clears the internal edit.

The Compilation System modules are divided into four groups: (1) Loader, (2) Stocking, (3) Volume, and (4) Area (fig. 1). Loader modules were designed as a dynamic front-end engine to translate, format, and populate data into national and regional tables. When the data structures, definitions, and variables change within the input data received from the field, the Loader modules can accommodate the changes without affecting the other more static module groups. The National Field Manual will require major system design changes in the Loader modules, but the other modules will need only minor or no modifications. The Stocking group consists of modules that calculate trees per acre, stocking, forest type, and stand size, which use national algorithms. The Volume group modules calculate total cubic foot volume; cubic volume of the sawlog section; board foot volume; growth, removals, and mortality; and weight. The Area group modules calculate forest area, area factors, and remeasurement factors.

The Stocking group consists of four major modules: (1) trees per acre, (2) stocking, (3) forest type, and (4) stand size. The trees per-acre modules calculate both prism plot sample trees per acre and mapped/annual plot trees per acre. Two methods calculate prism trees per acre depending upon the size of the trees. For trees 1.0 to 4.9 inches in d.b.h.:

> Trees per acre = 300 / number of measured sample points.

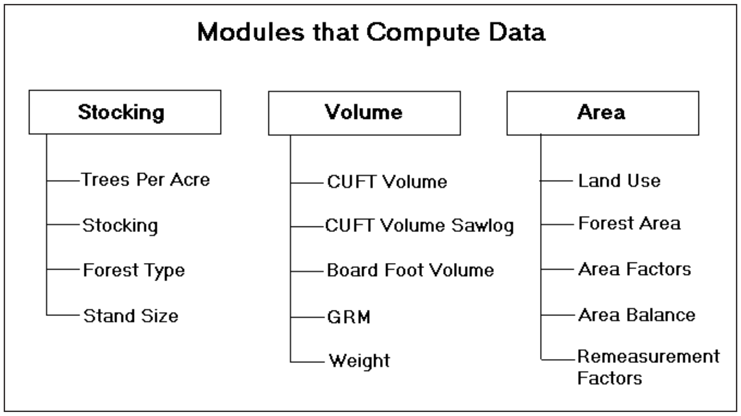


Figure 1—Computation modules of the Southern FIA Compilation System.

For trees = to 5.0 in. d.b.h.,

Trees per acre = 6,875.49354 / [(number of measured sample points) d.b.h.$^{2}$ ].

Two methods calculate the annual/mapped trees per acre for trees sampled in a microplot or subplot. For trees within a microplot:

Trees per acre = 43,560 / [(number of microplot points / 100) 581.07].

For trees within a subplot:

Trees per acre = 43,560 / [(number of subplot points / 100) 7,238.23].

A national team of scientists developed algorithms for the Stocking, Forest Type, and Stand Size modules in the Stocking group.

The Volume group consists of five major modules: (1) Cubic-Foot, (2) Sawlog Portion, (3) Board-Foot, (4) Growth, Removals, and Mortality (GRM), and (5) Weight. The volume and weight modules use standard volume equations generally of the form:

Volume or weight = Coeff A + Coeff B (d.b.h.$^{2}$ Height).

However, any acceptable equations may be used to calculate the tree volumes and weights.

The GRM module is more complex because it must deal with missing items such as diameter and height of cut and dead trees. The general model for GRM growth values is

Growth = (current volume - previous volume) / remeasurement period.

A regression model computes any missing d.b.h.

Predicted Current d.b.h. = (Coeff A ) Former Measured d.b.h. $^{\text{(Coeff B)EXP(Coeff C) Former Measured d.b.h.}}$.

This produces values for deriving annual radial increments for the area sampled.

By adjusting a height equation for site differences, we can predict the height of a cut or mortality tree from the predicted diameter. First, we predict the current height from equations of the form:

Predicted Current Height = Coeff A + Coeff B [Log10(Predicted Current d.b.h.)]$^{1/2}$.

Next, we predict a Former Height using equations developed from the same model, with Measured Former D.B.H. replacing Predicted Current D.B.H. We then determine a harmonic proportion by:

Proportion = Measured Former Height / Predicted Former Height.

Finally, we predict Current Height as:

Predicted Current Height = Predicted Current Height (Proportion).

The resulting height is a function of the original height of the tree as well as the diameter. This procedure reflects the influence of the tree site on the height prediction. After calculating a tree's missing variables we can then estimate the growth of removals or mortality. The trees are theoretically grown forward or shrunk backward by the appropriate number of years of growth.

The Area group consists of five major modules: (1) Land Use, (2) Forest Area, (3) Area Factors, (4) Area Balance, and (5) Remeasurement Factors.

The Land Use module loads the photo interpretation information. The Forest Area module interprets aerial photography, field calls, and intensification plot samples corrected for forest, nonforest, and water land-use types.

We divide the county forest area, nonforest area, and water area by the number of conditions within each classification to produce an area factor for each sample. The sample area factor is proportioned by the percentage of each land-use classification and assigned to sample conditions.

The Area Balance module adjusts the area factors at the condition level so that the rounded area factors will equal the enumerated acreage of the county. The Remeasurement Factor module calculates an area factor based upon the previous inventory forest acreage and forest sample plots for GRM expansion. The Southern Annual Inventory uses different procedures for calculating area factors (Reams and Van Deusen 1999), because the current inventory sample must be combined with the previous inventory sample.

## Challenges

The overall challenge was to design a system that would accommodate the various existing and planned situations. The system has to accommodate (1) multiple State inventories using different procedures to produce a common set of data, (2) a massive flow of field data, (3) major changes without substantially affecting the processing of data, and (4) a major sample design change due to transition from a periodic inventory to an annual inventory. The SRS Compilation System can accommodate all of these complexities and more.

One specific major challenge was the calculation of area. The national database structure requires that the Area Factors be applied at the plot level and not the condition level where they are calculated. This was particularly difficult for the southern State inventories because Area Factors are calculated for forest, nonforest, and water areas, especially when a sample contained mixed conditions between forest, nonforest and water. Two examples highlight area calculation challenges:

**Example 1**—Table 1 illustrates an Area Factor situation at the plot level. The first two rows demonstrate a problem that arises when individual condition-level Area Factors are combined into a single plot-level Area Factor. This sample plot contains two conditions: one is a nonforest land use of code 62 and the other a forest land use of code 20 (col. 1). There are separate Area Factors for nonforest and forest at the county level (col. 6), and these factors have different values (4,423.7331 acres and 6,016.7952 acres, respectively). To obtain the proportioned condition Area Factor for nonforest in this sample, we multiply the county nonforest Area Factor (4,423.7331 acres) by the condition proportion of nonforest (.75, col. 2), which equals 3,317.7998 acres (col. 3). We calculate the forest condition Area Factor (1,504.1998 acres) in the same manner. If we sum these individual condition factors with all of the other sample plot condition factors grouped by their respective land uses, they would equal their respective totals of area for the county. However, if we combine the two condition factors for different land uses *based upon different Area Factors* into a single plot-level Area Factor, they will never equal the correct county acreage for the respective land use. The values in Table 1 illustrate this. The single plot-level Area Factor, derived by summing the nonforest condition factor (3,317.7998 acres) with the forest condition factor (1,504.1998 acres) equals 4,821.9996 acres (col. 4). This value is the only Area Factor that the national database tables carry.

To calculate area for the respective land uses within a county based upon its sample plots, we multiply the single plot-level Area Factor (col. 4) by the condition proportion. For this example, the procedure produces the following results: the nonforest condition factor would equal the plot Area Factor (4,821.9996 acres) multiplied by the nonforest proportion (.75), which equals 3,616.497 acres (col. 5). The forest Area Factor would equal the plot Area Factor (4,821.9996 acres) multiplied by the forest condition proportion (.25), which equals 1205.4999 (col. 5). But now neither of these two values will sum to the correct acreage for their respective land uses at the county level. The nonforest condition factor should be 3,317.7998 acres, but has been calculated by the plot-level Area Factor method as 3,616.4970 acres. This results in 298.6972 too many acres for this condition. The forest condition factor should be 1,504.1998 acres, but has been calculated by the plot-level Area Factor method as 1,205.4999. This results in 298.6669 too few acres for this condition. Thus, since the county-level Area Factors for differing land uses will always be different values (col. 6), this method does not work when there are different land use conditions. On the other hand, this example also illustrates that when a sample plot has two different conditions that are in the same land use (rows 3 and 4), the plot-level Area Factor (col. 4) process does work because they are both calculated using the same county land-use Area Factor (col. 6).

The lower portion of table 1 demonstrates the solution. We must recalculate the condition proportion (col. 2) for each sample that contains more than one land-use condition (col. 1). We accomplish this recalculation by dividing the calculated land-use condition area factor (col. 3) by the county Area Factor (col. 6) and using that value as the condition proportion (col. 2: .688 and .312 for land uses 62 and 20, respectively) in the database record. We then proportion the condition Area Factors by these new values, resulting in 3,317.5857 and 1,504.4628 acres for Land Use conditions 62 and 20, respectively (col. 5). The sum of these new values (which bring along the influence of their original Land Use factor) is 4,821.9985 acres, a value very close to the column 4 value. When we proportion the plot Area Factors by the recalculated condition proportion (col. 5), the resulting values are also very close to the original condition area Land Use factors (col. 5). Note that since these values are not the exact calculated condition Land Use factors,

**Table 1—An example of a potential problem in calculating area factors**

| Land use | Condition proportion | Condition area factor | Plot area factor | Condition area factor proportioned | County area factor |
|---|---|---|---|---|---|
| | | | **Plot level area factor challenge** | | |
| 62 | .75 | 3,317.7998 | 4,821.9996 | 3,616.497 | 4,423.7331 |
| 20 | .25 | 1,504.1998 | 4,821.9996 | 1,205.4999 | 6,016.7952 |
| 20 | .75 | 2,691.846 | 3,591.9996 | 2,693.8845 | 3,591.8459 |
| 20 | .25 | 897.9615 | 3,591.846 | 897.9615 | 3,591.8459 |

Solution: Condition Proportion = Condition Area Factor / County Land Use Area Factor

CP62 = 3,317.998 / 4,423.7331 CP20 = 1,504.1998 / 6,016.7952

| Land use | Condition proportion | Condition area factor | Plot area factor | Condition area factor proportioned | County area factor |
|---|---|---|---|---|---|
| 62 | .69 | 3,317.7998 | 4,821.9996 | 3,327.1772 | 4,423.7331 |
| 20 | .31 | 1,504.1998 | 4,821.9996 | 1,543.0387 | 6,016.7952 |

there is a small difference in the total acreage when summed at the county level. Rounding of the recalculated condition proportion makes the difference larger.

**Example 2**—Another challenge has been the calculation of an Area Factor under the moving average estimation approach (Reams and Van Deusen 1999) for the annual forest inventory system. This concept combines the current panel completed with the rest of the sample population. In other words, if panel 1 contains 20 percent of the sample plots just inventoried, and the remaining 80 percent of a State's plots were inventoried during the last survey cycle, to produce statewide estimates we must combine the forest area calculation using the 20-percent sample with the older 80-percent sample. Two other complications offer challenges. There may be a transition from an old to a new photo interpretation methodology, and sometimes the previous compilation methodology used data in a flat-file format, whereas the current data resides within a relational database format.

First, we collected and reformatted all of the flat-file data and loaded it into the current database structure. We calculated a new estimate of forest area and computed a new set of Area Factors for the entire sample. Then another problem surfaced. The 20-percent current sample population level volumes had been calculated using current tree data and current Area Factors. The 80-percent previous sample population-level volumes had been calculated using previous data and previous Area Factors. To put the entire sample into the same context, we had to calculate the 80-percent sample population level-volumes using the current set of Area Factors.

To accomplish that task, we reduced plot volumes to individual tree-level volumes, recalculated trees-per-acre for each tree, and then recalculated population-level volumes using the current Area Factors. Either of two alternative procedures could be utilized: (1) reformat the 20-percent current sample into a flat-file format and combine those data with the 80-percent previous data, recalculating the population volumes; or (2) reformat the previous 80-percent sample into the National database format and then combine the samples. We selected the first method, reasoning that tested and trusted table-building software could build very accurate tables based on the combined data. If there were discrepancies, that method would point to the compilation procedures and not the table-building procedures, thus quickly revealing any existing problems.

Unfortunately, there were discrepancies. As is common in developmental work, thorough testing of the compilation system output revealed that the first procedure did not produce acceptable results. It, thus, became necessary to pursue development using the second procedure.

## CONCLUSION

The indexing and modularization techniques are two key procedures in the new Southern FIA Compilation System that make complex compilation situations manageable. The development examples presented, showing a success and an initial failure, represent just a few of the many challenges we encountered with the new system. So far, the life cycle of the system's processing phase has required 6 months of planning and design, 9 months of initial application programming by one application developer, and 7 months of testing, debugging, and modifications. Unit personnel have concentrated their effort for more than a year on algorithm development, programming, testing, and documentation for all modules, and on computer system design, development, implementation, and maintenance. A conservative estimate of time already spent on this project is about 12,000 person-hours. Final development, testing, and debugging are currently in progress using actual data in a production-type mode of operation. The system will be operational in 2001.

## LITERATURE CITED

**Anon.** 2000. Computer desktop encyclopedia [CD-ROM]. Vers. 14.1. Point Pleasant, PA: The Computer Language Company, Inc., December 2000.

**Arner, S.L.; Woudenberg, S.; Water, S. [and others].** National algorithms for determining stocking class, stand size class, and forest type for forest inventory and analysis plots. [Documentation available from the Northeastern Research Station, Forest Inventory and Analysis, 11 Campus Blvd., Newton Square, PA 19073].

**Gildin, R.** 2000. October 17, 2000. Keynote presentation presented at the second annual forest inventory and analysis symposium; 2000 October 17–18; Salt Lake City, UT. Unpublished presentation. On file with: R. Guldin, Forest Service, U.S. Department of Agriculture, Sidney R. Yates Federal Building, 201 14th Street, SW at Independence Ave., SW, 1 NW Yates, Washington, DC 20250.

**Koch, G.; Loney, K.** 1997. Oracle 8, the complete reference. Berkely, CA: Osborne McGraw-Hill: 9.

**Kreines, D.C.** 2000. Oracle SQL, the essential reference. 1st ed. Sebastopol, CA: O'Reilly and Associates.

**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.

**Urman, S.** 1996. Oracle PL/SQL programming. Berkeley, CA: Osborne McGraw-Hill.

**U.S. Department of Agriculture, Forest Service.** 1967. Handbook. Washington, DC.

**U.S. Department of Agriculture, Forest Service.** 1998. Field instructions for the southern forest inventory, A: remeasurement of prism. Version 3. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.

**U.S. Department of Agriculture, Forest Service.** 1998. Field instructions for the southern forest inventory, B: remeasurement of prism. Version 2. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.

**U.S. Department of Agriculture, Forest Service.** 1999. Field instructions for the southern forest inventory, Kentucky—remeasurement of fixed-radius plots. Version 3. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.

# ON FIA VARIABLES FOR ECOLOGICAL USE[1]

David C. Chojnacky[2]

**Abstract—**The Forest Inventory and Analysis (FIA) program collects or calculates over 300 variables for its national network of permanent forest plots. However, considerable ecological analysis can be done with only a few key variables. Two examples—Mexican spotted owl habitat in New Mexico and down deadwood in Maine—are used to illustrate the potential of FIA data for ecological use. These examples illustrate the importance of the variables (1) diameter at breast height, (2) tree species, and (3) live/dead/cut tree status for compiling estimates and confidence intervals within FIA's sample design. Priority variables are suggested for constructing an ecological database with FIA data.

## INTRODUCTION

The Forest Inventory and Analysis (FIA) program collects or calculates over 300 variables for its national network of permanent forest plots (FIA 2001). Many of these variables are collected specifically to assess the U.S. timber supply. However, ecological analysis of forest structure is also possible with these data.

The FIA variables can be classified into two broad groups of "attribute" and "category" variables (Chojnacky 1996). Attribute variables are generally continuous variables based upon direct field measurements (or functions of direct measurements). Examples include diameter at breast height (dbh), tree age, height, volume, biomass, and so forth. Category variables are generally discrete classifications of observed phenomena or social/political groupings. Examples include county, vegetation class, geographic location, tree species, forest type, land uses, and many other observations of forest structure and use impacts.

Of the two variable types, the continuous attribute variables are most flexible because these can usually be summarized in raw form, combined with other variables in calculations, used in models to produce estimates, and fit theory for confidence interval computation within FIA's double sampling for stratification design (Chojnacky 1998). On the other hand, discrete categorical variables are generally not compiled as statistical endpoints. More often, categorical variables classify summations of attribute variables or classify forest area.

The difference between "attribute" and "category" can be fuzzy for some calculated variables that are functions of both variable types, but the point of distinction hinges on the end product. A variable can be considered an "attribute" if it is possible to total it in some meaningful manner and estimate a variance within FIA's sample design. This definition can be tricky for a variable such as percent understory cover because a "total cover" estimate is not meaningful, but a ratio estimator for total cover divided by total area (for any size area) is a nice attribute variable with a variance (Chojnacky 1998, p. 13).

Categorical variables are not necessarily undesirable, but they are somewhat limited and often difficult to define for multiple uses. For example, forest type (Hansen and others 1992) and habitat type (Pfister and others 1977) are two categorical variables for plot-level vegetation description. Forest-type classifies from a timber stocking perspective, and habitat-type classifies from a climax vegetation perspective, but neither necessarily gives an accurate description of present tree cover by species, and they do not collapse uniformly in some hierarchical fashion. Ability to logically collapse categorical variables is crucial because FIA data applications inevitably have too few plots for some category, which requires category grouping.

An example of a well-defined, flexible, categorical variable is taxonomic tree species. Species distinction is supported by a wealth of information, including taxonomic nomenclature, genetics, growth rates and forms, specific gravity, shade tolerance, nutrient requirements, and so forth. Species has been traditionally used by FIA to tally regional timber statistics on volume or numbers of trees, but FIA defers to forest-type all plot-level forest classifications because of the problem of multiple species per plot. However, this is likely more a matter of traditional convenience from earlier days of limited computing power than of practical necessity. For example, the continuous variable, basal area, could easily be ranked according to predominate species or species group to obtain a flexible plot classifications tailored to many different needs (Chojnacky and Woudenberg 1994).

This paper illustrates, by example, ecological analyses done with only a few variables. Estimates are based on attribute (continuous) variables but also use categorical (discrete) variables that can be easily and meaningfully regrouped. The two ecological examples include Mexican spotted owl habitat calculated for a national forest in New Mexico and down deadwood estimated for Maine.

## METHODS

The data for both examples were previously analyzed in other studies. For the owl example, habitat data were

[1] Paper presented at the Second Annual Forest Inventory and Analyis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, USDA Forest Service, 201 14th Street, SW, Yates Building, Washington, DC 20250.

compiled for a test on using FIA data to monitor Mexican spotted owl habitat as defined by a recovery plan for the owl's "threatened" status (Chojnacky and Dick 2000). Included were 464 FIA plots spanning 1.3 million ha of forest and wilderness in New Mexico's Gila National Forest from the 1994 inventory. FIA attribute variables used were dbh, tree height, and number of trees per plot. FIA category variables included tree species, live/dead tree status, tree dominance class, forest type, and habitat type. The latter three variables were needed for an elaborate "forest cover" algorithm defined by the recovery plan. (Forest cover probably could have been defined more simply, but the previous study had to be done in compliance to recovery plan guidelines.) Also used were several sample design variables for field plots sizes, phase 1 and phase 2 samples sizes, and stratum identifications and areas. These were needed to compute confidence intervals and expand the estimates to forest totals.

Data for the down deadwood (DDW; also called coarse wood debris) example were compiled from 1,842 plots that were re-measured in Maine's 1995 inventory (Chojnacky and Heath [In preparation]). Down deadwood data are not currently available in the FIA database but they were collected in Maine from transects overlaid on FIA plots. Down deadwood is important for assessing carbon stocks for global warming concerns, habitat for numerous organisms, nutrient cycling, and soil movement (fig. 1). The purpose of the deadwood study was to predict DDW from other routinely collected FIA variables. Results showed a subsampling scheme and simple model as reasonable. For this approach, dbh and DDW were the only attribute variables used. FIA category variables included tree species, live/dead/cut tree status, ownership, stand size class, and forest type. Because Maine's 32 forest types were unmanageable for simple compilation, predominate basal area by species was used to collapse the forest types into six groups. Also used were the sample design variables for sample sizes and stratum areas.

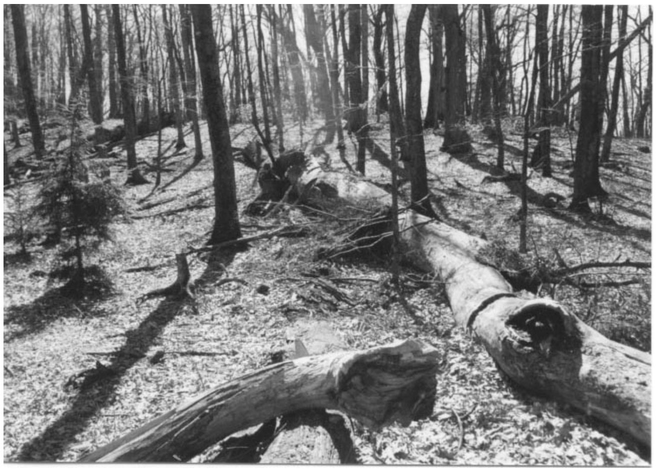


Figure 1—Down deadwood or coarse woody debris found in eastern hardwood forest.

## RESULTS AND DISCUSSION

### Mexican Spotted Owl

An actual amount of Mexican spotted owl habitat was not calculated from the FIA data. Instead, area was calculated for a plausible habitat scenario defined from the bird's habitat needs (Chojnacky and Dick 2000). The main purpose of the scenario was to assess the FIA sample intensity for detecting change. Results of the 95 percent confidence intervals for mixed-conifer and ponderosa pine forest cover ranged from ±20 to ±35 percent (fig. 2). The small amount of pine-oak was more variable at ±50 to ±73 percent.

The results for the Gila National Forest were encouraging because the planned use of the method was for combining several national forests into recovery units, which would increase sample size and further reduce confidence intervals. The habitat scenario needs testing against actual owl demographic data, but there seems sufficient power in the FIA data to monitor modest changes in habitat area.

The FIA variables needed to monitor owl habitat included dbh, height, tree species, live/dead tree status, and several other plot-level and tree classifications for computing forest cover. The list could be shortened to the first four variables if forest cover were recomputed from species and basal area (or trees per area or other dbh-based density metrics).

### Down Deadwood

The other example on down deadwood in Maine illustrates DDW estimated from a subsample. The FIA program has recently combined with the Forest Health Monitoring national network of plots, which are sampled at about 1/16th the intensity of the FIA grid. These plots are being called the third phase (P3) of FIA's sample design, with first phase (P1) being the remote sensing for stratification points and the second phase (P2) being the full sample of FIA field plots.

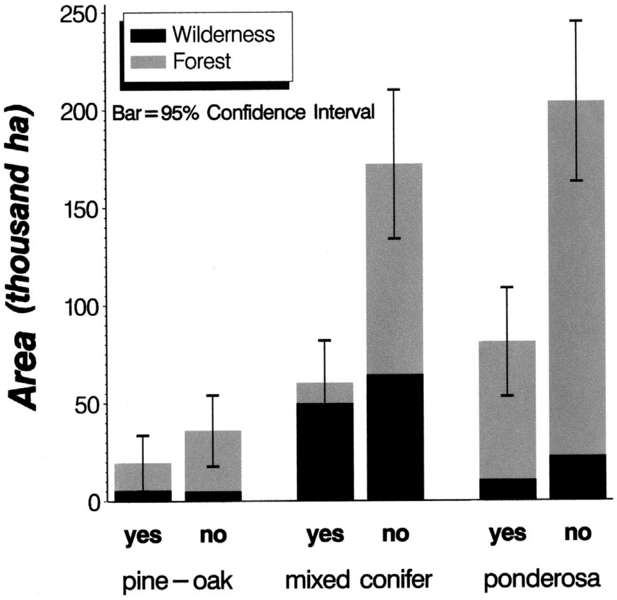


Figure 2—Area of hypothetical Mexican spotted owl habitat for Gila National Forest, 1994 FIA inventory. The scenario illustrates habitat can be determined from FIA data within about 20 to 35 percent of the estimates for mixed conifer and ponderosa pine forest cover.

A method for including DDW in an FIA assessment is to model DDW from the P3 plots and then apply the model to the more intense P2 design. A regression model $(R^2 = 0.20,\ n = 135)$ fit to P3-plot data is:

$$\begin{aligned} DDW = {} & 6.2092 - 3.3003 * d_i - 1.3542 * d_{cr} - 2.2594 * X_1 + \\ & 0.3527 * X_1^2 + 0.9124 * X_2 + 0.3487 * X_3 \end{aligned}$$

where

$$\begin{aligned} DDW &= \text{down deadwood greater than 7.6 cm diameter (Mg/ha)} \\ d_i &= \text{1 if forest industry ownership, 0 otherwise} \qquad (1) \\ d_{cr} &= \text{1 if conifer forest type, 0 otherwise} \\ X_1 &= \text{FIA stand size class (seedling, poletimber, sawtimber)} \\ X_2 &= \text{basal area of dead trees 7.6 - cm dbh and larger } (m^2/ha) \\ X_3 &= \text{basal area of recently cut trees 7.6 - cm dbh and larger } (m^2/ha) \end{aligned}$$

Application of the P3-based model to FIA P2 data compared favorably (fig. 3). There was little statistical difference between methods because the 95 percent confidence intervals overlapped for the total and for all species groups, except spruce-fir.

These results illustrate another aspect of the FIA design where subsampling can be used to add in new attribute variables such as DDW without the expense of collecting the variable on every plot. The results could also have been summarized directly in a three-phase sample design instead of using the regression model, but the details of this application for FIA data have not yet been worked out.

As for the owl example, the key FIA variables include dbh, tree species, and live/dead/cut tree status. Additional classifications for ownership and stand size class were used to construct the P3-model, but stand size class could have been redefined from basal area and numbers of trees.

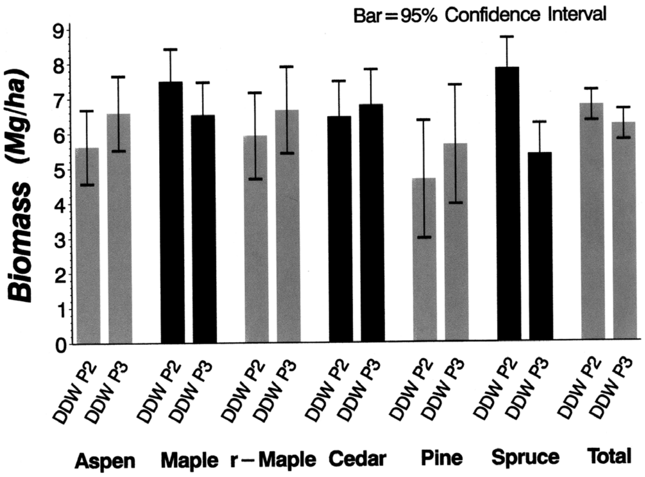


Figure 3—Down deadwood (DDW) biomass estimated from FIA plots (DDW P2) for Maine's 1995 inventory compares well to a model-based estimate (DDW P3) from a 1/16th subsample. The model was developed from P3 plots but applied to P2 plots for forest type groups (aspen/birch, maple/beech/birch, red maple/other hardwoods, cedar/hemlock, pine/oak, spruce/fir).

## CONCLUSION

Of all the FIA variables, (1) dbh, (2) tree species, and (3) live/dead/cut tree status seem most important for ecological use. As shown in the examples, these can describe forest structure for two different uses. By adding in tree age, height, a few social/political variables (such as county, ownership, measurement period, and geographic coordinates), and the sample design variables, one has foundations for a strong ecological database.

Because FIA already collects these variables to some degree, it would be fairly easy to construct an ecological subset of the FIA database. Key items should include the following:

### Attribute variables

- Diameter for all trees at a consistent height above groundline regardless of where the tree was actually measured. (Because diameter is so important, it would also be desirable to have conversion capability between groundline and breast height measurement points for all species. Groundline diameter makes much biological sense because it is at the interface between roots and bole.)

- Meta-data explaining age measurement, and future development of field procedures that consistently subsample tree ages in a statistical design.

- Meta-data explaining height measurement and development of field procedures that consistently subsample tree heights in a statistical design.

### Category variables

- Live/dead/cut tree status code that also includes a decay class for dead.

- Tree species and all other discrete observations and classifications that are consistently recorded throughout the entire FIA program, such as State, county, measurement period, ownership, tree damage, slope, aspect, and so forth.

### Sample design variables

- Trees-per-area expansion factors (plot sizes) for every live, dead, cut, and missing tree within a plot regardless of size or classification.

- Phase 1 sample sizes and strata information necessary to calculate a variance for any estimate.

This list is meant to emphasize priorities but not include exhaustive detail. Needed is consistently available information for calculating stand structure metrics and variances from tree-level data. FIA has recently done a good job of including noncommercial tree species and all forestlands without regard to timber utility into its inventories. However, many tree-level and plot-level compilation procedures still include gaps when considering all trees. Creating an ecologically oriented subset of FIA data would simplify data access for users who need to carefully account for every for every live, dead, cut, or missing tree regardless of value judgments.

## ACKNOWLEDGMENTS

I thank John Caspersen, Princeton University; Andy Gillespie, Washington Office FIA; and Gretchen Moisen, Rocky Mountain Research Station FIA, for reviewing this manuscript.

## REFERENCES

**Chojnacky, D.C.** 1996. Forest Inventory and Analysis (FIA) variables: indicators of ecological integrity? In: Aguirre Bravo, C., ed. North American workshop on monitoring for ecological assessment of terrestrial and aquatic ecosystems; 1995 September 18–22; Montecillo, Texcoco, Mexico. Gen. Tech. Rep. RM–GTR–284. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station: 247–258.

**Chojnacky, D.C.** 1998. Double sampling for stratification: a forest inventory application in the Interior West. Res. Pap. RMRS–RP–7. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 15 p.

**Chojnacky, D.C.; Dick, D.L.** 2000. Evaluating FIA forest inventory data for monitoring Mexican spotted owl habitat: Gila National Forest example. Western Journal of Applied Forestry. 15(4): 195–199.

**Chojnacky, D.C.; Heath, L.S.** [In preparation]. Estimating down deadwood from FIA variables in Maine. In: Proceedings, Advances in terrestrial ecosystem carbon inventory, measurements, and monitoring; 2000 October 3–5; Raleigh, NC.

**Chojnacky, D.C.; Woudenberg, S.W.** 1994. Toward an ecological approach to inventorying cedar-hemlock-white pine in the Inland Northwest: barriers and opportunities. In: Proceedings, Interior cedar-hemlock-white pine forests: ecology and management; 1993 March 2–4; Spokane, WA. Pullman, WA: Washington State University, Department of Natural Resources: 9–16.

**FIA.** 2001. Homepage of U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis. http://fia.fs.fed.us/. (4 January 2001)

**Hansen, M.H.; Frieswyk, T.; Glover, J.F.; Kelly, J.F.** 1992. The eastwide forest inventory database: users manual. Gen. Tech. Rep. NC–GTR–151. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station. 48 p.

**Pfister, R.D.; Kovalchik, B.L.; Arno, S.F.; Presby, R.C.** 1977. Forest habitat types of Montana. Gen. Tech. Rep. INT–34. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. 174 p.

# Analysis and Reporting

*Moderator:*

**Ron McRoberts**
USDA Forest Service
North Central Research Station

# ESTIMATION BASED ON THE FIRST CYCLE OF THE ANNUAL FOREST INVENTORY SYSTEM: METHODS, PRELIMINARY RESULTS, AND OBSERVATIONS[1]

**Mark H. Hansen, Gary J. Brand, Daniel G. Wendt, and Ronald E. McRoberts[2]**

**Abstract**—The first year of annual FIA data collection in the North Central region was completed for 1999 in Indiana, Iowa, Minnesota, and Missouri. Estimates of timberland area, total growing-stock volume and growing-stock volume per acre are presented. These estimates are based on data from 1 year, collected at the base Federal inventory intensity, a lower intensity sample than previous periodic inventories conducted in these States. In the North Central region, plots are measured on a 5-year cycle (20 percent of the plots measured each year) at a base intensity of one plot per 5,937 ac. These first-year estimates, obtained from the 20 percent sample, are based on a sample intensity of one plot per 29,685 ac and, thus, have greater sampling variability than previous periodic inventories conducted by FIA.

## INTRODUCTION

The Forest Inventory and Analysis (FIA) program at North Central Research Station, USDA Forest Service, has completed the first year of statewide annual forest inventories in four States: Indiana (IN), Iowa (IA), Minnesota (MN), and Missouri (MO). Ground plot measurements and quality control checks were made by a combination of NCFIA field personnel, cooperating State employees, and private forestry consultants under contract to cooperating state agencies. Plots in these states were measured between September 1998, and April 2000.

These first year plot data, together with a thematic GIS layer based on a classification of Landsat TM data for stratification purposes (Vogelmann and others 1998), were used to produce estimates and sampling errors for standard forest resources variables. This paper presents results from the inventory, compares those results to estimates from previous periodic inventories, and discusses implications of these comparisons.

## DATA AND ESTIMATION

NCFIA sampled a total of 5,240 systematic plot locations in the four-State study area using the National FIA plot design, a cluster of four 1/24th ac fixed area subplots. Of these 5,240 plot locations, 1,467 were found to contain some forest land. Table 1 compares the number of plots observed in each state under the first year of the annual inventory system to the number of plots in the most recent periodic inventory. It is important to keep in mind that there was a change in plot design in IA, MN, and MO between the last periodic inventory and 1999. Under the old plot design, a plot was tallied as a forest plot only if a forest condition existed at plot center. Under the new plot design, a plot is tallied as a forest plot if any portion of the plot falls in a forest condition. This change in plot design increases the percentage of forest plots in the total sample without a change in total forest area, especially in areas with a great deal of forest/nonforest interface such as IA and MO.

**Table 1—Number of plots observed in the first year (1999) of the annual inventory and number of plots observed in the most recent periodic inventory in the North Central Region**

| | Number of observed plots | | | | Date of most recent periodic inventory |
|---|---|---|---|---|---|
| | 1999 annual inventory | | Most recent periodic inventory | | |
| | Total | Forest | Total | Forest | |
| Iowa | 1,202 | 110 | 12,767 | 713 | 1998 |
| Indiana | 769 | 174 | 6,402 | 1,605 | 1990 |
| Minnesota | 1,801 | 578 | 43,955 | 13,507 | 1990 |
| Missouri | 1,468 | 605 | 17,259 | 5,072 | 1989 |
| All States | 5,240 | 1,467 | 80,383 | 20,897 | |

[1] Paper presented at the Second Annual Forest and Inventory Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, Research Forester, Supervisory Computer Specialist, and Mathematical Statistician, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108, respectively.

The intensity of the plots from a single year of the annual inventory is lower than any periodic inventory conducted by NCFIA. This first year sample comprises 20 percent of the base sampling intensity of the full 5-year inventory. When the complete 5-year cycle of plots has been measured, estimates will be based on all the plots measured over the 5-year cycle (five times the number available for this first year). In addition, some states have provided resources to intensify the inventory by measuring additional plots. In this first year, we have not included any intensified plots in the analyses.

The ground plot locations sampled in the first year are a combination of plot locations from the previous periodic inventory and plots in new locations. The hexagon/panel system (Brand and others 2000) determined the plot locations that were measured. This system divides the entire conterminous 48 States into 5,937-ac hexagons and selects a measurement location within each hexagon. When one or more locations from the previous inventory exist within a hexagon, one is selected for remeasurement, otherwise a new location is selected. In three States (IN, MN, and MO) the previous periodic inventory used a sample design in which the status of some plots that were determined to be undisturbed between the two previous inventories were updated using models (Miner and others 1988) rather than a field and remeasurement. For example, when the 1990 MN inventory was conducted, all plots measured in the 1977 inventory plots were examined using aerial photography. The conditions on a portion of the plots that were found to be undisturbed were projected forward to 1990 using a forest growth model (adjusted using information from actual remeasured data), and the resulting projected plot data were used as observations in the estimation process. In the plot location selection procedures for the annual inventory system, both the remeasured and projected plot locations were candidates for measurement. Thus, three kinds of plots were measured in the 1999 annual inventory:

1. New plots at locations never previously measured,
2. Remeasured plots from the previous periodic inventory, and
3. Remeasured plots that were not measured in the immediate previous periodic inventory.

Table 2 summarizes the numbers of these various kinds of plots for each State.

In addition, various ground plot designs were used in different States so that not every remeasurement consists of the remeasurement of the same plot design. IN used the current standard FIA plot design for its previous inventory, however, the other States used a 10-point cluster of variable radius plots. The change in plot design and the mix of new and remeasurement plots limits our ability to estimate the components of change attributes requiring actual plot remeasurements (e.g. growth, removals, and mortality) and also increases the sampling error associated with estimates of total change (e.g. change in forest area and change in total growing-stock volume). Only after 10 years under the annual inventory system will all the plots used to produce estimates be remeasurement plots with observations 5 years apart using a standard plot design, assuming we maintain this system and plot design.

The stratified random sampling estimator, with stratification after the selection of the sample (post-stratification), was used to produce all resource estimates. This estimator is a special case of the double sampling for stratification

**Table 2—Number of plot locations by plot type and State**

| | Number of observed plots | | | |
|---|---|---|---|---|
| States | New plots | Plots remeasured from the previous periodic inventory | Plots remeasured from and older periodic inventory | All plot types |
| | **All plots (includes forest and nonforest plot locations** | | | |
| Iowa | 240 | 962 | 0 | 1,202 |
| Indiana | 42 | 296 | 431 | 769 |
| Minnesota | 103 | 928 | 770 | 1,801 |
| Missouri | 160 | 920 | 388 | 1,468 |
| All States | 545 | 3,106 | 1,589 | 5,240 |
| | **Forest plot locations only** | | | |
| Iowa | 19 | 91 | 0 | 110 |
| Indiana | 13 | 97 | 64 | 174 |
| Minnesota | 84 | 382 | 112 | 578 |
| Missouri | 87 | 427 | 91 | 605 |
| All States | 203 | 997 | 267 | 1,467 |

estimator that has been used in NCFIA estimation for over 30 years. With the use of satellite imagery and computer classification in place of photo plot sampling and human interpretation, we obtained known strata areas (rather than estimates of strata areas), eliminating one source of sampling error in our estimates. National Land Cover Data (NLCD) available from the Multi-Resolution Land Characteristics Consortium (MRLC) (http://www.epa.gov/mrlc/) are used to obtain strata areas. This classification was conducted without any knowledge of the location, classification or other characteristics of FIA ground plots and is truly independent of the ground sample. Independence of the strata area estimates and the ground plot classification (assumed in previous inventories that used double sampling for stratification) came into question when both the strata area estimates and classification of plot locations into strata was done by the same photo interpreters. Some bias that could not be corrected was likely introduced into estimates under that system. Details of the procedures used to produce these estimates are presented in another paper in this report (Hansen 2001).

## RESULTS

Estimates of timberland area, total growing-stock volume and average growing-stock volume on timberland for each state are shown in figures 1 through 3. Estimates from the first year of the annual inventory and those from all periodic statewide inventories since 1965 are shown for comparison. In these figures, the solid lines indicate the estimates and the dashed lines indicate the estimates plus or minus one standard error and thus delineate a 67 percent confidence interval around those estimates.

Estimates of timberland area increased in three of the four States (MO, IA, and MN) from the estimates made at the time of the previous inventories: 1989 for MO, 1990 for IA, and 1990 for MN. Data for these inventories were collected over several years and the aerial photography used for stratification was taken prior to the field work, usually several years prior. The date (year) of a periodic inventory is a reporting date assigned to the entire inventory and reflects the year when the last of the field plot measurements were taken, although the measurements are usually made over several years. Based on the inventory date, the estimated annual change in timberland area between inventories was 0.54 percent in MO, 0.32 percent in MN, and 0.37 percent in IA. These increases continue trends that were observed between the prior two inventories in these three States.

The estimates for IN are contrary to the trends that were seen in the past and show a decrease in timberland area. The last periodic inventory in IN was 1998, just one year prior to the first year of the annual inventory. An estimated 5.65 percent decrease in timberland area is indicated between that inventory and the first year of the annual inventory system. The sampling error on the 1999 estimate of the timberland in IN is fairly high (4.56 percent), and the sampling error on the 1998 estimate is considerably less (1.59 percent). As mentioned previously, the 1998 inventory was based on considerably more observations, and the 1999 estimate is based on a remeasurement of only 11 percent of these plots (727 of the 6,402 plots measured in 1998) plus the addition of 42 new plots. Also, the 1998 estimate was based on a different classified image for stratification. In detailed analysis of the data, only 3 of 727 (0.41 percent) remeasurement plots in the 1999 annual data indicated an observed change from timberland to another condition from the 1998 periodic inventory. One of these three plots was measured in 1998 and observed to be timberland at that time. When this plot was remeasured in 1999, the area had been cleared of most trees and construction of a park was underway. The other two were field visited in the 1986 inventory and found to be timberland, classified undisturbed in 1998 (based on aerial photo interpretation), and not remeasured in 1998. In 1999, when these two plots were remeasured, they were observed to have changed to a nonforest land classification. One plot

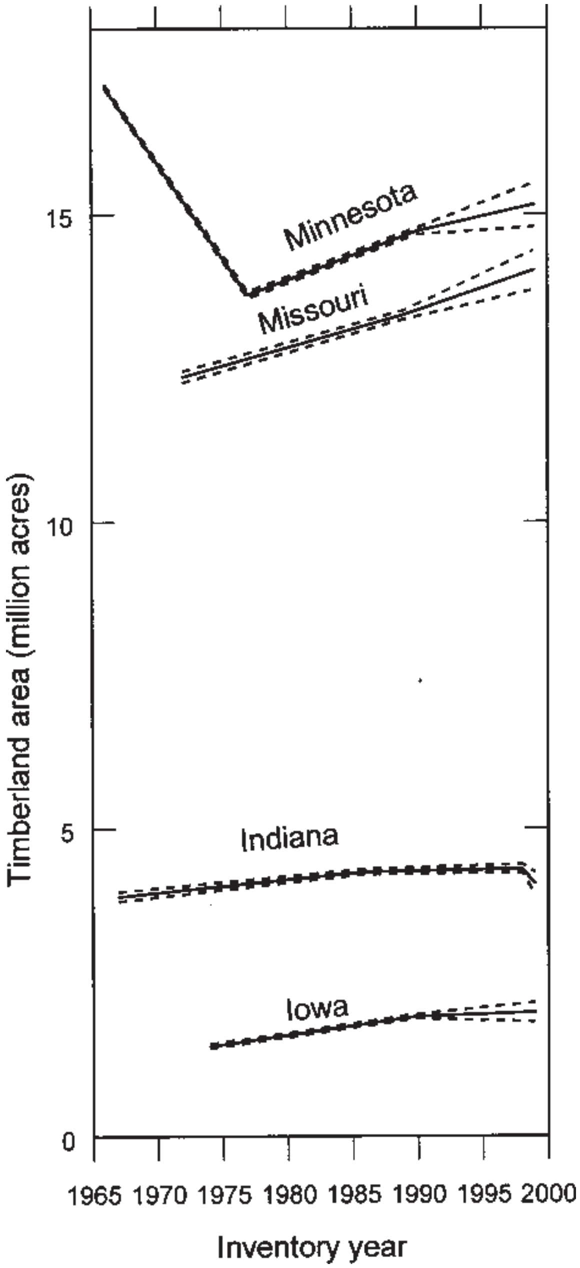


Figure 1—Timberland area estimates for four states, 1965 to 1999. Solid lines indicate the estimates and the dashed lines indicate the estimates plus or minus one standard error.

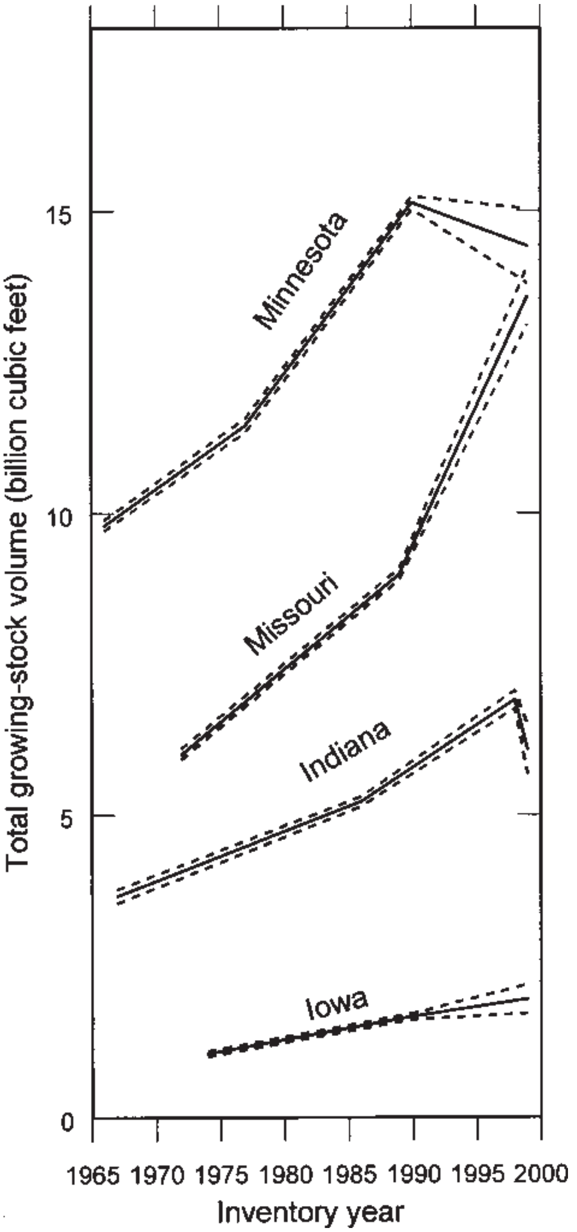


Figure 2—Total growing stock volume estimates for four States, 1965 to 1999. Solid lines indicate the estimates and the dashed lines indicate the estimates plus or minus one standard error.

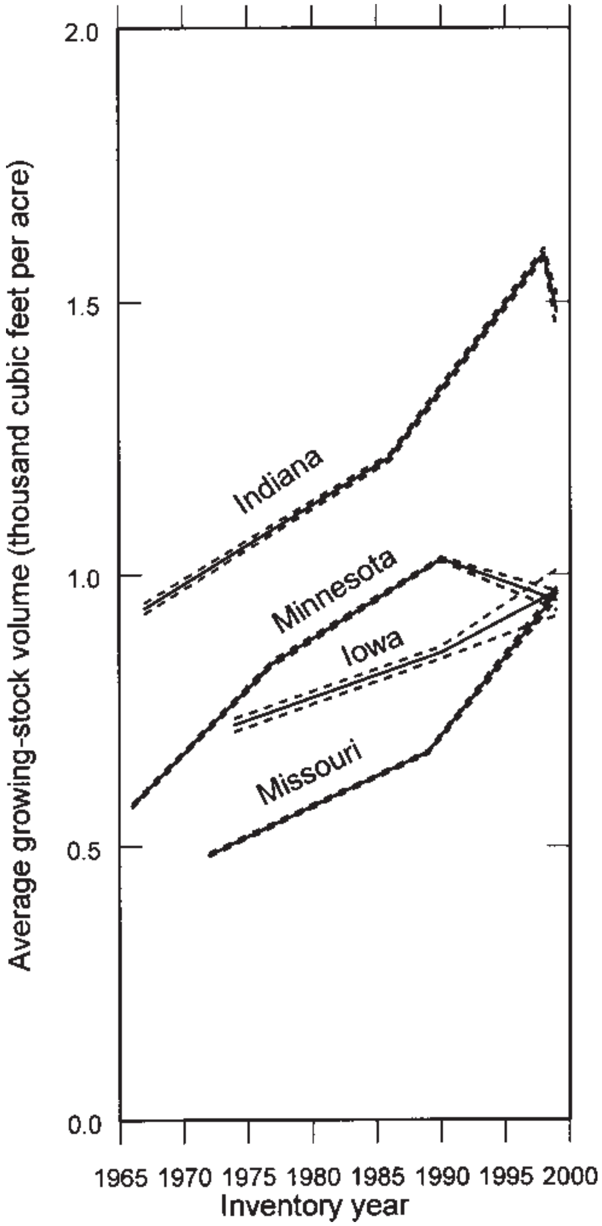


Figure 3—Average growing stock volume per acre estimates for four states, 1965 to 1999. Solid lines indicate the estimates and the dashed lines indicate the estimates plus or minus one standard error.

was in a lowland area where flooding associated with beaver activity had killed the trees, and the other was a farm woodlot where tree mortality associated with cattle grazing resulted in a reclassification of the plot to nonforest. In both of these cases it is unclear when the change to nonforest occurred. This analysis of the IN remeasurement plots indicates that the decrease in estimated timberland area in IN in 1999 from that reported in the 1998 periodic inventory may be a result of the high sampling error associated with the 1999 estimates rather than a loss of timberland between the two inventories.

The total growing-stock volume estimates presented in figure 2 show a large increase in MO from 1989 (over 5 percent per year), a smaller increase in IA from 1990 (slightly less than 2 percent per year) and net decreases in MN (0.5 percent per year) and IN (12 percent per year). Total growing-stock volume estimates are dependent on estimates of timberland area and volume per acre, of which both have associated sampling errors. When combined, the sampling errors for total growing-stock volumes are fairly high. Figure 3 shows the growing-stock volume per acre estimates for each State. As with the timberland area estimates in IN, there is a decrease in the estimated growing-stock volume per acre (6.7 percent from 1998 to 1999). This decrease, combined with the estimated decrease in timberland area results in a large decrease in estimated total volume. Again, high sampling errors most likely account for a part of this estimated decrease. As additional data become available, the precision of these estimates will improve; and it will be possible to associate changes in volume to the various factors that can cause these changes (growth, mortality, harvesting, and land-use change). Although sampling errors are fairly high, it does appear that the trend of increasing growing-stock volumes in MN and IN that have been observed over the past two periodic inventories may have ended. Additional analysis and data are needed to confirm these findings and to study the causes and impacts of these changes.

Other publications are available that present more detailed estimates for individual States (Schmidt 2000a, 2000b). Additional analysis of the data is being conducted and will be presented in future publications. Also, data will be made available via the internet at http://www.fs.fed.us/research/databases.htm.

## DISCUSSION

One purpose of this paper was to demonstrate that estimates from the first year of the annual forest inventory could be made shortly after the completion of data collection. FIA has made a commitment to complete inventories on an annual basis and make final data and estimates from these inventories available when sufficient data is available for detailed analysis. Estimates from the first four states inventoried under this new system were presented at a symposium within six months of the final field measurements. As we gain more experience and improve data collection and processing systems the time to produce and report estimates will decrease. As the speed of reporting increases, the amount of detailed analysis that we can provide with the estimates decreases. To date, results are reported with a minimum of analysis. As annual inventories are implemented and replace periodic inventories as the source of the best available information on the forest resources of a state, users will find it possible and necessary to perform more of the detailed data analysis themselves.

During the transition period to annual inventories, sampling errors for estimates are high and detailed breakdowns of the estimates are questionable. Our plans for making raw data and estimates available to our users over the next few years will be determined by the number of years of annual inventory data collected in a state. Following the first year, state-level estimates such as those presented here will be published. After 2 years, nine State-level tables of area, volume, and number of trees will be published. Following the third year the full suite of 24 standard FIA core tables will be published. After 4 years, the core tables will be revised and web access to the plot data and the FIA table generation program will be available. Following the fifth year of data collection, a full analytical report will be prepared. Periodic analysis of the data will be conducted once the system is in full operation. Over the transition some States will have few remeasurement plots (Kansas, Nebraska, North Dakota, and South Dakota) and estimates of change will not be produced until sufficient remeasurement data is available. In these States tables related to current conditions will be produced following this schedule.

## REFERENCES

**Brand, G.J.; Nelson, M.D.; Wendt, D.G.; Nimerfro, K.K.** 2000. The hexagon/panel system for selecting FIA plots under an annual inventory. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C., eds. Proceedings of the 1st annual Forest Inventory and Analysis symposium; 1999 November 2–3; San Antonio, TX. Gen. Tech. Rep. NC–213. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 8–13.

**Hansen, M.H.** 2001. Remote sensing precision requirements for FIA estimation. In: Reams, G.A.; McRoberts, R.E.; Van Deusen, P.C., eds. Proceedings of the second 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: 43–51.

**Hansen, M.H.; Wendt, D.G.** 2000. Using classified landsat thematic mapper data for stratification in a statewide forest inventory. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C., eds. Proceedings of the 1st annual Forest Inventory and Analysis Symposium; 1999 November 2–3; San Antonio, TX. Gen. Tech. Rep. NC–213. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 20–27.

**Miner, C.L.; Walters, N.R.; Belli, M.L.** 1988. Guide to the TWIGS program for the North Central United States. Gen. Tech. Rep. GTR–NC–125. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 105 p.

**Schmidt, T.L.** 2000a. Minnesota's forest resources in 1999, Res. Note. RN–NC–376. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 8 p.

**Schmidt, T.L.** 2000b. Missouri's Forest Resources in 1999, Res. Note. RN–NC–375. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 8 p.

**Vogelmann, J.E.; Sohl, T.; Campbell; P.V.; Shaw, D.M.** 1998. Regional land cover characterization using landsat thematic mapper data and ancillary data sources. Environmental Monitoring and Assessment. 51: 415–428.

# ANALYSIS AND REPORTING NEEDS FOR ANNUAL FOREST INVENTORIES IN THE SOUTH[1]

**Michael T. Thompson**[2]

**Abstract**—The annual forest inventory system essentially eliminates the concept of the periodic inventory. Therefore, the reporting requirements and needs for individual States will change as the South makes the transition to full annual inventories. The Farm Bill and the Forest Service Strategic Plan require both annual reports and 5-year comprehensive, analytical reports. The Southern Research Station Forest Inventory and Analysis unit proposes that the annual reports be brief research notes, supplemented by available hard copy and electronic tabular data. The annual reports will summarize estimates derived from a moving average of five inventory panels. Based on the preceding 5 years of data, succinct analytical reports will focus on status and change in forest resources—with color maps and graphics—and nonproduct-related issues. A statistical report published prior to the analytical report can contain all tabular data, detailed discussions of inventory methods, sampling errors, and definitions. Where the previous full inventory is outdated and cannot be included in moving-average estimates, there will still be a need for interim reports in States that have initiated annual plot measurements.

## INTRODUCTION

The annual forest inventory system implemented nationwide by Forest Inventory and Analysis (FIA) essentially eliminates a periodic survey cycle. As a result, reporting requirements must change to accommodate the annual nature of releasing inventory estimates. As all reporting will still be at the State level, individual States in various stages of transitioning to full annual inventories will need to adjust their reporting. In this paper, I will discuss FIA reporting requirements and needs for (1) States still inventoried under a periodic inventory, (2) States in transition to full annual inventories, and (3) States in full annual inventory mode. It will explain how FIA in the South will meet the reporting objectives outlined by the Farm Bill and the Forest Service Strategic Plan.

## REPORTING REQUIREMENTS FOR PERIODIC INVENTORIES

For States with a full periodic inventory, the two required reports are statistical reports (regional and State) and analytical reports. Statistical reports that follow the format in Tennessee's recent periodic inventory will add an expanded methods section that addresses differences between the current and the previous inventory procedures. A number of changes in procedure, such as sample designs, volume computation, and merchantability standards, will affect the assessment of change between inventory periods. Reports will clearly identify these changes and, when appropriate, highlight problems that affect trend statistics.

Analytical reports for periodic inventories will be different if the inventory was conducted before or after the merger of the Southern and Southeastern FIA units. States using the format of the previous separate FIA units include Arkansas (1995), Florida (1995), Mississippi (1994), Oklahoma (1993), and Texas (1992). Analytical reports for States inventoried during or after the merger will make a transition toward the new model for analytical reports.

## INVENTORY ESTIMATES IN ANNUAL INVENTORIES

The 1998 Farm Bill mandates annual collection of field data for 20 percent of all FIA plots in each Southern State. Several Forest Inventory and Monitoring (FIM) workshops have discussed accurate time-series techniques to estimate current conditions at the scale of a State or sub-State region (Survey Unit). Participants agreed that annual measurement of 20 percent of the FIA plots in each State provides the most current information, but also reduces the sample size to an unacceptable level of accuracy to most users of the data. Therefore, the recommendation is that States in transition from a periodic to an annual inventory use a moving average that includes the last periodic inventory if the data are not too old. The data are considered too old if the period is greater than approximately 3 years.

The assumption is that FIA will measure a total of $n$ plots over a period of $m$ years. We will divide $n$ plots into $m$ approximately equal panels. The proposed target for the South is $m = 5$, and $n/5$ = 20 percent of the plots measured each year.

The inventory statistics in an annual report will use the moving average, averaged over the last $m$ years. When interpreting the results, a user should consider this data an unbiased estimate of conditions. Some trend analyses will have a time-bias; this is not expected to be great if $m = 5$. Actually, the moving average is a familiar concept for those States that required several years to inventory. For example, for the 1997 periodic inventory, Georgia took plot measurements from the fall of 1995 to the spring of 1998.

## REPORTING REQUIREMENTS FOR ANNUAL REPORTS

The Farm Bill and the Forest Service Strategic Plan require that States compile and release data in both hard copy and electronic format once per year on a predetermined schedule. FIA considered several options to meet these

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Resource Analyst, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804.

objectives. Electronic format supplemented with simple hard copy would deliver a product more quickly, but would give little credit to analysts and the Research Work Unit. Another option was for State agencies to publish annual estimates, but not all States would publish results equally. Therefore, the primary means of reporting annual inventory results should be through a brief Forest Service publication, such as a research note, supplemented by tabular data available in hard copy and electronic form. Analysts should pursue collaborative reporting efforts between State agencies and the Forest Service.

## REPORTING REQUIREMENTS FOR ANALYTICAL REPORTS

The Farm Bill and Forest Service Strategic Plan require that every 5 years the Forest Service produce a complete State analytical report that includes core information covering the current status of the forest resource based upon the previous 5 years of data. More specifically, inventory statistics should use the Moving Average where $n$ is the number of plots measured under the panel system over a period of $m = 5$ years. Also recommended are trends in forest status and condition, timber products output information for each State, and projections for key resource attributes over the next 20 years. To meet these requirements, one comprehensive report could cover all aspects of the forest resource and associated inventory data. Past reports in the South fit this category, but they were timber oriented and lengthy. The advantage is that everything is available in one publication; the major drawback is the long development and publication process required, which is detrimental to the timely objective of the Farm Bill.

The proposed approach is to develop analytical, succinct, 5-year reports focused on the general status and change in the forest resources of each State. A wider audience can be reached with a shorter report providing color maps and graphics, and by more inclusive, nonproduct-related language whenever possible. The analytical report should contain 20 to 25 pages of text, graphics, and maps. The development and publication of a statistical style report should occur prior to the development of the analytical report. The statistical report would contain all of the tabular data, detailed discussion of inventory methods, sampling errors, and definitions. These sections would be excluded in the analytical reports. The major advantage to developing 5-year reports is a short turnaround period, enabling timely reports. Analysts will have more time to write scientific reports on forest resources in each State.

## REPORTING REQUIREMENTS FOR INTERIM REPORTS

Before the end of the 5-year cycle, States that have initiated annual plot measurements, but whose previous full inventory is too old, will generate interim reports. Analysts can process two or three panels of inventory data (40 to 60 percent of the full periodic inventory plot total) and then update the estimate annually. For example, Virginia, whose previous inventory was 1992, will soon complete three panels, or 60 percent, of the total periodic inventory. After this estimate, an interim status report will be published on forest resources.

## REPORTING FOR GEORGIA'S INVENTORY

Georgia's inventory serves as an example of the reporting process. The design-based estimate used for annual surveys will be the moving average. The State will combine plot measurements from the previous periodic inventory with plots measured in the annual inventory.

The field work for Georgia's 1997 periodic inventory began in November of 1995 and ended in April 1998; all five Survey Unit reports and the State Statistical Report have appeared. The annual inventory effort began immediately after the periodic inventory, and the field work for panel 1 (approximately 20 percent of the periodic inventory plot total) was completed in June of 1999. For Georgia, then, the State-wide measurement interval between a plot measured in panel 1 and the same plot measured in the periodic inventory averaged 1.8 years.

The FIA moving-average estimate for Georgia will combine about 80 percent of the plot network from the 1997 periodic inventory and the 20 percent of plots selected for panel 1.

Let us suppose that the reporting attribute of interest is the average per-acre volume of all live standing trees 5.0 inches in diameter at breast height (d.b.h.) and larger on timberland area in the State of Georgia. The inventory sample to measure live-tree volume was a fixed-radius plot cluster of four points spaced 120 feet apart. Each point served as the center of a 1/24-acre circular subplot used to sample trees 5.0 inches d.b.h. and larger. These sample plots were established without regard to land use or forest cover. Forest and nonforest condition classes were delineated and recorded. FIA defines condition classes by six attributes: land use, forest type, stand origin, stand size, stand density, and major ownership category (Thompson 1998). FIA assigns trees to their respective condition classes.

To determine the average per-acre volume of live trees 5.0 inches d.b.h. and larger, a weighted mean was computed. The equation for mean per-acre live volume is,

$$\bar{y} = \sum_{i=1}^{w} (y_i w_i) / \sum w_i \qquad (1)$$

where, $y_i$ = per-acre volume of live trees 5.0 inches d.b.h. and larger on timberland area sampled by plot $i$ ,

$w_i$ = proportion of total area sampled on plot that sampled timberland area, and

$\sum w_i$= sum of the proportions of plots that sampled timberland area.

The statistics for computing the standard error were determined using PROC MEANS with the WEIGHT statement and VARDEF=WDF option (SAS Institute 2000).

The Statewide inventory per-acre estimate of all live trees 5.0 inches d.b.h. and larger on timberland as determined by the moving-average estimate ($\bar{y}_m$) is,

$$\bar{y}_m = n_{2,3,4,5}/N(\bar{y}_{2,3,4,5}) + n_1/N(\bar{y}_1) \qquad (2)$$

where

$n_1$ = number of plots that sampled timberland in panel 1,

$n_{2,3,4,5}$ = number of plots that sampled timberland in panels 2,3,4, and 5,

$N = n_1 + n_{2,3,4,5}$,

$\overline{y}_1$ = mean live volume per acre on timberland for pane (equation 1), and

$\overline{y}_{2,3,4,5}$ = mean live volume per acre on timberland for panels 2,3,4, and 5 (eq. 1);

$\overline{y}_p$ = the mean determined by all plots in the periodic inventory that sampled timberland, representing the Statewide inventory per-acre estimate of all live trees 5.0 inches d.b.h. and larger.

Note that in equation 2, the plots in panels 2,3,4, and 5 have not been measured under the annual system, but are considered a valid component of the estimate since the most dated samples were measured in 1995. Figure 1 illustrates the average per-acre volume of live trees 5.0 inches d.b.h. and larger on timberland area for the State of Georgia for the three separate estimates $\overline{y}_1$, $\overline{y}_p$, and $\overline{y}_m$. The estimate in the annual report will be $\overline{y}_m$, and any discussions of change or trend from the previous inventory will be the difference between $\overline{y}_m$ and $\overline{y}_p$.

## DISCUSSION

For the immediate future, as the transition continues toward annual inventories, FIA in the South will deal with data availability and reporting. After implementation of the annual inventories, Southern States will use the simple $m$-year moving average for reporting of FIA information. The moving average is a reasonable and practical way to analyze and report data. In the future, when more complicated analysis will update inventory statistics, the moving average approach can evaluate whether other estimates are improving the results.

## LITERATURE CITED

**SAS Institute Inc.** 2000. SAS online documentation. Version 8.1. Cary, NC: SAS Institute Inc. [CD-ROM].

**Thompson, M.T.** 1998. Forest statistics for Georgia, 1997. Resour. Bull. SRS–36. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 92 p.

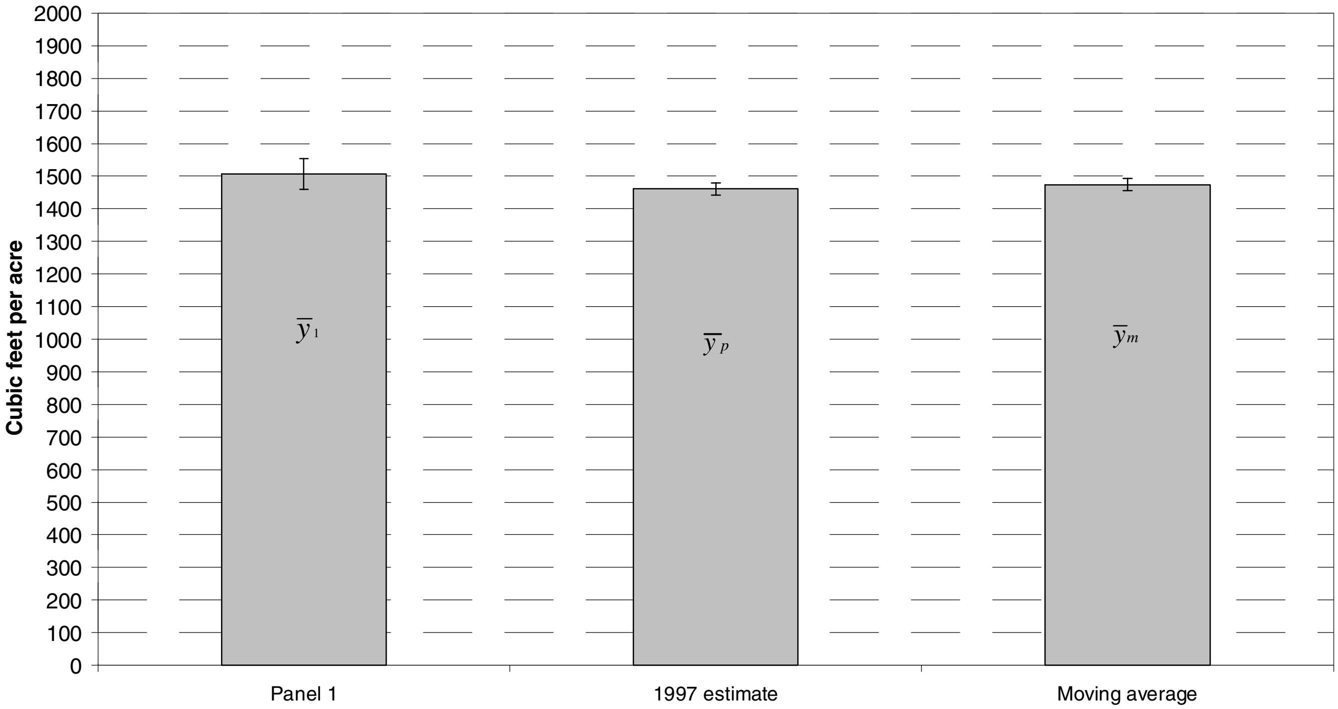


Figure 1—Three estimates of average per-acre volume of live trees 5.0 inches d.b.h. and larger on timberland area for the State of Georgia.

## Special Studies

*Moderator:*

**Greg Reams**
USDA Forest Service
Southern Research Station

# ASSESSING REGENERATION ADEQUACY IN PENNSYLVANIA'S FORESTS: A PILOT STUDY[1]

**William H. McWilliams, Susan L. King, and Charles T. Scott[2]**

**Abstract—**The USDA, Forest Service, Northeastern Research Station (NE), Forest Inventory and Analysis (FIA) unit began collecting forest inventory data on an annual basis in Pennsylvania starting this past field season. The forestry community of Pennsylvania has identified forest regeneration as a primary research issue for the inventory to address. New techniques for measuring and quantifying regeneration are needed because existing NE-FIA protocols and national FIA protocols will not provide the level of detail required. A pilot study is being conducted to determine a cost efficient method for measuring tree seedlings, shrubs, and competing vegetation. The study is expected to result in a recommended approach for full implementation next field season.

## INTRODUCTION

Along with resource sustainability, forest regeneration has emerged as a critical research question in Pennsylvania due to the paucity of tree seedlings found in the typical forest understory (McWilliams and others 1995). The USDA, Forest Service, Northeastern Research Station (NE), Forest Inventory and Analysis (FIA) unit began collecting forest inventory data on an annual basis in Pennsylvania starting this past field season. The annual inventory provides an opportunity for landscape-level assessment of regeneration adequacy across the range of forested ecosystems common in the State. Regeneration assessment in Pennsylvania will require development of new inventory techniques because existing NE-FIA protocols (USDA Forest Service, 2000) and national FIA protocols (USDA Forest Service 2000) do not provide the level of detail needed to address regeneration adequacy.

The diversity of plant communities common across Pennsylvania's forested landscape, high white-tailed deer (*Odocoileus virginianus* Boddaert*)* populations, competing vegetation, and other factors make this a challenging and costly research problem to address. A consortium of interested groups within the State has agreed that a comprehensive system for evaluation and monitoring of forest understories is needed. Specific research questions associated with this goal are:

- What are the abundance, composition, and quality of advance regeneration?
- What are the abundance, composition, and quality of regeneration following major disturbance?
- What are the extent and composition of competing vegetation?
- What is the status of regeneration of oak and other key species?

Because of the importance and complexity of this issue, a pilot study to assess regeneration measurement protocols is being conducted during this year's field season.

## PILOT STUDY

### Goals and Objectives

The regeneration pilot study is intended to provide a field test of regeneration sampling design and measurements. Objectives of the study are:

- Determine the minimum (or optimal) set of sample plots required to quantify the character of tree regenerationand competing regeneration,
- Develop a scientifically credible and peer-reviewed set of measurement protocols that address the range of research questions,
- Test the analytical framework for assessing tree regeneration and competing regeneration, and
- Provide a cost model for full implementation of the regeneration assessment.

The findings of the pilot study will be used in the annual inventory design for Pennsylvania and as a tool for designing and implementing regeneration studies in other northeastern states.

### Sample Design

The basic premise for designing the sample used in the pilot study was to collect as much data as possible within the footprint of the national sample design. For example, tree seedlings were tallied on four 6.8-foot microplots at each subplot, compared to one microplot that is used for the national sample design (fig. 1). The resulting dataset allows for analysis of standard errors associated with differing numbers of microplots. Data collection spanned the period between leaf-on and leaf-off conditions to allow a subjective evaluation of the field forester's ability to identify tree seedlings.

Regeneration and supporting data were collected at three levels: sample location, subplot, and microplot. Time data

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Foresters, USDA Forest Service, Northeastern Research Station, 11 Campus Blvd., Suite 200, Newtown Square, PA 19073.

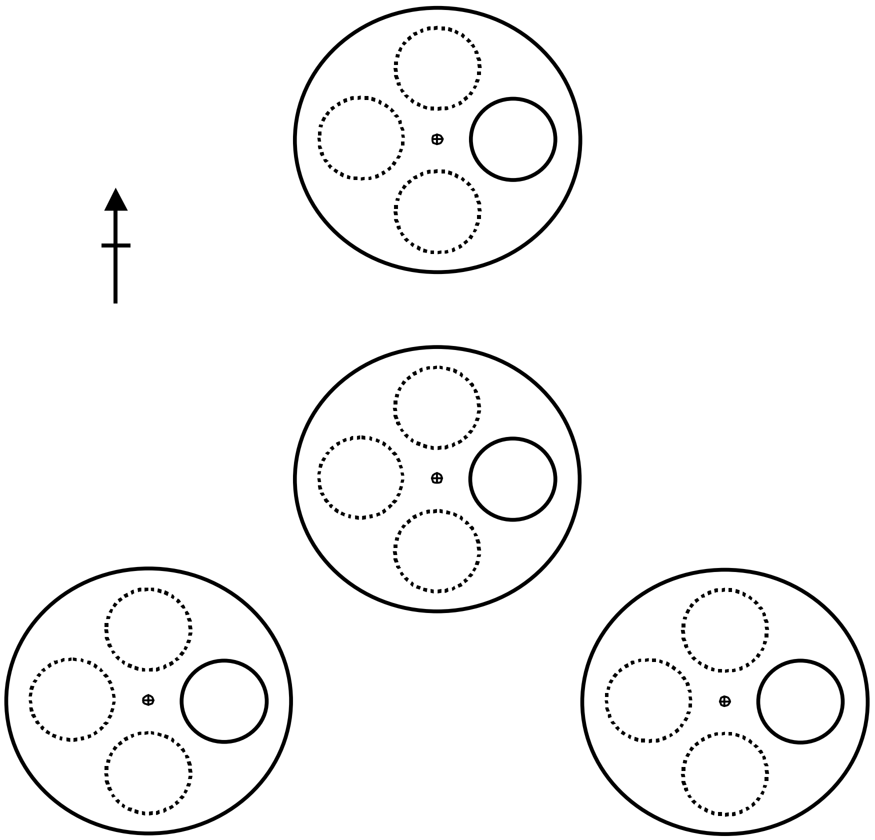


Figure 1—Sample location layout. The larger circles represent 24-foot radius subplots that are located 120 feet apart at azimuths 360, 120, and 240. The smaller circles represent 6.8-foot radius microplots that are located 12 feet from subplot center at azimuths 90, 180, 270, and 360. Microplots with solid lines are part of the national sample design. Microplots with dashed lines were added for the pilot study.

were recorded to provide estimates of the cost of collecting regeneration data and assist in determining the number of personnel needed to complete the measurements.

At the sample location level, a broad assessment of the impact of deer on understory vegetation was made. The assessment was based on general knowledge of local deer populations, the amount of available alternative food sources within one-square mile of the sample location (agricultural crops), and a list of understory species ranked by preference for browse by deer.

At the subplot level, a tally of vegetation that competes with tree seedling development was made. In the context of this study, competing vegetation is defined to include non-tree vegetation; such as shrubs, fern, grass, and other herbaceous vegetation that may inhibit the establishment and development of tree seedlings. Standard NE-FIA species codes were used for deciduous, evergreen, and dwarf shrubs and vines. Additional codes were needed for rhizomous ferns (Hayscented, New York, and Bracken), other ferns, grass, and other herbaceous vegetation. All competing vegetation data was tallied using 10-percent cover classes. The larger subplot was used to tally competing vegetation because of the patchy occurrence of these life forms (Marquis and others 1990).

Measurements of trees were collected on four 6.8-foot radius microplots located at cardinal directions within each subplot. The tree tally is intended to provide a sample representing future stand occupancy and composition. Conceptually, the size of the microplot represents the ground area occupied by a tree once it reaches a diameter of 5.0 inches, which represents a fully established tree of merchantable size (Marquis and others 1990). The sample of tree seedlings was limited to “established” seedlings. As such, seedlings that were less than 2.0-inches tall, had fewer than two normal sized leaves, or that bore cotyledons were not counted.

At each microplot, tree seedlings were tallied by species, seedling source (stump sprout and other), and eight height classes. A tally of saplings (trees from 1.0-inches to 4.9-inches in diameter) was made using a condensed set of the national FIA protocols that included diameter, crown ratio, and crown class. On microplots with trees 5.0-inches and larger, the species code of the most dominant large tree was recorded.

### Complicating Factors

Design of an analytical framework for assessing regeneration adequacy in Pennsylvania is complicated by many factors. A diverse mix of forest communities and conflicting habitat uses make it difficult to set universal regeneration guidelines. Fike (1999) describes 54 forest and woodland communities. Very often, communities are not distinct but are in transition to other community types making it difficult to assess the ability of understory vegetation to perpetuate existing or potential overstory compositional traits. For example, a mature overstory comprised of oak (*Quercus* sp.) and ash (*Fraxinus* sp.) species may contain a well-stocked understory of red maple (*Acer rubrum* L.). The high economic value of oak, black cherry (*Prunus serotina* Ehrh.), sugar maple (*Acer sacharrum* Marsh.), ash (*Fraxinus* sp.), and other species in relation to less valuable species common in forest understories; as well as the importance of wildlife, aesthetic, and other recreational activities and their related habitat requirements further confuse the issue.

Any classification of regeneration must consider site occupancy requirements of individual species, degree of establishment, and seedling source. The suite of species common in Pennsylvania exhibit a wide range of site occupancy requirements that depend on numbers of stems and stem height. For example, the probability of survival of a six-foot seedling is quite different from a six-inch seedling. In addition, the degree of establishment can be challenging to measure, especially for small seedlings. New research has shown that root collar diameter is a better indicator of oak seedling establishment than seedling height (Personal communication. Dr. Patrick Brose. 2000. Research Forester, Northeastern Research Station, Old Route 6, Irvine, PA 16329–0267). Lastly, information on whether seedlings originate as stump sprouts or from other sources is useful for gauging the quality of the prospective future stand.

The regeneration assessment should also perform well under a variety of stand conditions, from older closed canopy stands to heavily disturbed stands with a need for new stand establishment. This will allow analysis of both advance regeneration levels and post-disturbance regeneration success.

Competing vegetation in forest understories is another complicating factor. The regeneration assessment should include an estimate of the degree that competing vegetation interferes with the establishment and growth of tree seedlings. So far, the study objectives do not call for remeasurement of percent cover; however, it is likely that remeasurement will emerge as an objective in future discussions with clients for this information. Any remeasurement design will need to consider the work of van Hees and Mead (2000) who noted some limitations of remeasuring percent cover estimates over time.

Perhaps the most significant factor influencing regeneration in Pennsylvania is the large deer population in the State. Population data indicate that relative deer densities exceed thresholds for adequate seedling development across most of the State (Pennsylvania Game Commission. 2000. Unpublished deer population data available from the author).

### Analytical Features

The analysis will focus on classifying regeneration adequacy of forest conditions encountered at each sample location and estimation of numbers of tree seedlings per acre. The number and placement of microplots will be determined using standard errors. Standard errors will be examined for individual species and species groups for different numbers and placement of microplots. The configuration(s) that provides the smallest acceptable standard errors will be considered for implementation. Quality assurance methods will be developed following the choice of sampling protocols. Lastly, cost estimates will be developed using the time data.

There are a number of existing guidelines for evaluating regeneration and silvicultural research will likely provide improved measures over time. Because of this, the approach for classifying regeneration adequacy needs to be flexible enough to consider different schemes for evaluating the stocking of seedlings by height class. It is also useful to be able to examine a range of regeneration guidelines, rather than assuming a single metric will fit all needs. As a starting point, the data will be compiled using the framework from a previous study of regeneration using NE-FIA data (McWilliams and others 1995). The approach considers three species groups, two levels of acceptable seedling density, and a set of weights to be applied to height classes. Individual tree species are assigned to desirable, commercial, and woody groups using commonly accepted timber conventions. The two levels of acceptable seedling density, 25 (low) and 100 (high) seedlings per acre, cover the range of regeneration guidelines found in the literature (Leak 1980, Sander and others 1976, Marquis and Bjorkbonm 1982). To account for different seedling survival rates by height class, each seedling was weighted as follows:

| Height Class | Weight |
|---|---|
| 2 to 6 inches | 1 |
| 6 inches to 1 foot | 1 |
| 1 to 2 feet | 2 |
| 2 to 3 feet | 2 |
| 3 to 5 feet | 20 |
| 5 to 10 feet | 50 |
| 10 to 15 feet | 50 |
| Greater than 15 feet | 50 |

Any combination of weighted stems that meets or exceeds the minimum number required was considered adequately stocked. For example, a microplot is considered to meet the high-density requirement if it contains a minimum of two stems at least five feet tall for the species group of interest. Similarly, a microplot is stocked at the low-density level if twenty 6-inch and three 2-foot stems are encountered. Any forested condition at the sample location is considered adequately regenerated if at least 70 percent of the microplots are adequately stocked.

The species groups, acceptable seedling densities, and height-class weights can all be adjusted to fit differing viewpoints regarding what constitutes adequate regeneration. It may also turn out that different height-class weights will be needed for individual species or species groups. For example, one viewpoint may hold that American beech (*Fagus grandifolia*) is not a commercial species. The prevalence of beech in Pennsylvania's forest understories would likely have a significant impact on the population estimates of commercial-species regeneration adequacy.

Analysis of competing vegetation is relatively straightforward. Estimates of percent cover for shrubs, fern, grass, and other herbaceous vegetation will be used to support and explain regeneration stocking and composition.

The suite of regeneration indicators will be used to classify the overall status of regeneration across Pennsylvania and for important sub-regions, such as ecoregions, forest-type groups, and geographical units. The measurements should also provide sufficient data for developing relationships between overstory and understory composition. This will facilitate predictions of prospective composition of future forest communities in the State.

## NEXT STEPS

The data collection phase of the study has been completed. The immediate next step will be to complete the analysis of study data to address the specific objectives of the pilot study described above. Once a final set of measurement protocols is determined, the protocols and associated analytical framework will be circulated for peer review and final adjustments will be made. Full implementation of the regeneration measurements will proceed during this coming field season. Another opportunity for research is to explore auxiliary data sources for correlation of regeneration findings with other variables, such as soils, physiography, deer populations, and acid deposition.

## REFERENCES

**Fike, J.** 1999. Terrestrial and palustrine plant communities of Pennsylvania. 8140–bk-dcnr–1128. Harrisburg, PA: Commonwealth of Pennsylvania, Department of Conservation and Natural Resources, Bureau of Forestry. 86 p.

**Leak, W.B.** 1988. Effects of weed species on northern hardwood regeneration in New Hampshire. Northern Journal of Applied Forestry. 5: 237–240.

**Marquis, D.A.; Bjorkbom, J.C.** 1982. Guidelines for evaluating regeneration before and after clearcutting Allegheny Hardwoods. Research Note NE–307. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station. 4 p.

**Marquis, D.A.; Ernst, R.L.; Stout, S.L.** 1990. Prescribing silvicultural treatments in hardwood stands of the Alleghenies (revised). Gen. Tech. Rep. NE–96. Radnor, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 101 p.

**McWilliams, W.H.; Stout, S.L.; Bowersox, T.W.; McCormick, L.H.** 1995. Advance tree-seedling regeneration and herbaceous cover in Pennsylvania forests. Northern Journal of Applied Forestry. 12(4): 187–191.

**Sander, I.L.; Johnson, P.S.; Watt, R.F.** 1976. A guide for evaluating the adequacy of oak reproduction. Gen. Tech. Rep. NC–23. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station.

**van Hees, W.S.; Mead, B.R.** 2000. Ocular estimates of understory vegetation structure in a closed Picea/Betula papyrifera forest. Journal of Vegetation Science. 11:195–200.

**USDA Forest Service.** 2000. Forest inventory and analysis national core field guide, Volume 1: field data collection procedures for phase 2 plots, Version 1.4. Washington, DC: U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis. 171 p.

**USDA Forest Service.** 2000. Forest Inventory and analysis northeast field guide. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, Forest Inventory and Analysis. 249 p.

# COMPARISON OF STANDING VOLUME ESTIMATES USING OPTICAL DENDROMETERS[1]

Neil A. Clark, Stanley J. Zarnoch,
Alexander Clark III, and Gregory A. Reams[2]

**Abstract**—This study compared height and diameter measurements and volume estimates on 20 hardwood and 20 softwood stems using traditional optical dendrometers, an experimental camera instrument, and mechanical calipers. Multiple comparison tests showed significant differences among the means for lower stem diameters when the camera was used. There were no significant differences among the methods for volume or height.

## INTRODUCTION

Digital technology is being utilized more and more to facilitate the collection of forest inventory data. Satellites that scan the earth's surface on a periodic basis provide affordable data for various forms of regional analyses. Digital aerial cameras and positioning systems can be directed to capture more detailed information quickly and with little manual processing. Technologies are reducing data collection costs and changing the way we can model and analyze these data. Regardless of methodology and scale, at some point this macroscale data must be combined with a more detailed subsample on the characteristics of the individual trees that make up the regional forests. For metrics such as volume or biomass, current methods of collecting this individual stem data are cost prohibitive, and some metrics such as crown dimension are highly error prone. There is hope that the digital camera system can aid in the affordable and accurate collection of individual stem data.

Since the 1950s, cameras have been used in various ways to collect tree stem data (Bradshaw 1972; Crosby and others 1983; Juujärvi and others 1998). In 1998, the USDA Forest Service funded a study to examine the feasibility of using a digital camera to collect data from individual standing trees (Clark 1998). While the method was feasible, several improvements—including increased focal length, digital range, and inclination—were needed before it could be considered practical. The main advantages of using digital cameras over film cameras are the elimination of film and development costs and the capability of direct integration with digital image processing software. These things greatly reduce the costs of using camera systems, propelling them into competition with traditional methods. This paper will examine how use of the camera instrument compares to standard and felled-tree methods of measuring diameter, height, and volume.

## METHODS

Twenty hardwood and 20 softwood trees were selected from a mature, mixed oak-pine stand at the Bent Creek Experimental Forest near Asheville, NC. The trees were approximately distributed by 1-inch classes from 4 to 24 inches in diameter at breast height (d.b.h.). Diameter measurements were made at heights of 0, 0.5, 1, 2, 4.5, 17.3 ft, and every 4 ft from 8 ft to the top of the tree (excluding 16 ft). The locations of the measurement points along the tree boles were not marked, and thus were measured independently by each method. However, to minimize extraneous variability, a vertical paint mark was made on each stem to ensure that measurements were taken from the same side of the tree by each method. Occasionally, due to thick underbrush or tree crowns, some of the crew members taking optical measurements may have deviated from the directional control in order to observe the diameter at the desired height. Total tree height was also recorded along with live crown ratio. Truth data were collected by felling each stem, then measuring diameters with a mechanical caliper and heights with a nylon tape.

### The Standard Method

Four four-member crews collected diameter and height data using sectional aluminum poles for height determination and pentaprism calipers for diameters. In some cases, McClure pentaprisms were used on diameters exceeding 24 in., which was the limit of the pentaprism calipers. Five hardwood and five softwood trees spanning the diameter ranges were systematically assigned to each crew.

The standard method deviated from the previously described protocol in two ways. First, whenever a fork or other drastic diameter change occurred, a new base was established above the anomaly from which diameters were again collected in 4 ft intervals. Second, diameter tapes or mechanical calipers were sometimes used for the measurement of d.b.h. in order to verify the size class of the stem. In some instances, these measurements were recorded in lieu of the optical dendrometer measurements.

## The Camera Method

The camera-rangefinding instrument (fig. 1) used in this study was a prototype model built by Laser Atlanta, Inc., which incorporates a Panasonic GP-CX161, 480 x 720 output pixel, color, CCD (charge coupled device) video camera into their Advantage® CIL laser rangefinder (Clark 2000). Camera data were output to a Sony GV-D300

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, USDA Forest Service, Southern Research Station, 1650 Ramble Road, Blacksburg, VA 24060; Mathematical Statistician, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804; Forest Products Technologist, USDA Forest Service, Southern Research Station, 320 Green Street, Athens, GA 30602, Mathematical Statistician, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804, respectively.

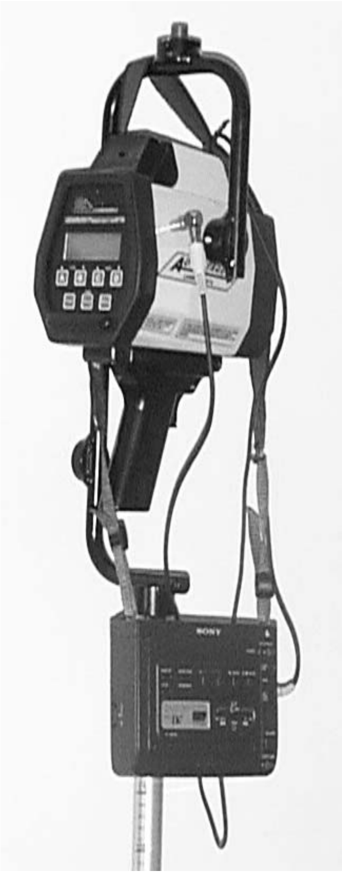


Figure 1—Camera rangefinder prototype used in this study.

portable digital video (DV) cassette recorder, and the range data were output to a memory card. Each mini DV cassette was capable of storing 60 minutes of video data. The 2Mb memory card was capable of storing approximately 36,000 range/bearing/pitch records. The instrument was set to record 30 frames per second of video data and 3 ranges per second (derived from 238 ranges measured per second).

Each stem was scanned with the camera-rangefinding instrument from two or more visible vantage points, with at least one of these points aligned with the vertical paint mark on the stem. Redundant data were collected from the additional vantage points to increase the probability that a given height was visible. In general, a close-range (10 to 30 ft) distance was used to view the lower portion of the stem and a >30 foot range for the upper portions in order to avoid severely acute perspective angles.

The camera data were post-processed back at the office, using customized software written in C++ for the Windows platform by Neil Clark. The following procedure was used to calculate the diameters:

1. Range data were filtered in order to determine the actual range to the stem and to filter out ranges from occluding objects.
2. Heights were calculated using ranges and inclination angles.
3. Video frames were manually correlated with the range data, extracted from the videotape, and saved as digital images.
4. Image coordinates representing the diameters were collected by on-screen digitizing and the resultant diameters were calculated, then output to a digital file.

Differences, defined as method measurements minus felled tree measurements, were analyzed for total height, volume calculated by Smalian's formula, and outside bark diameters at d.b.h., 17.3 ft, and in the clear bole and crown. The clear bole and crown diameters were determined using total height and live crown ratio. Graphical analysis and descriptive statistics were used to compare the methods. A randomized block design analysis was also performed to test for differences between the two instruments and the "true" value for each taxonomic division. Trees were considered blocks, which contained three treatments defined as Camera, Standard, and True. Overall significance tests were performed, least square means computed, and Bonferroni multiple comparisons performed using an experimentwise error rate of 0.05.

## RESULTS AND DISCUSSION

Figure 2 and table 1 show the results of diameter errors by category. Due to procedural errors, four softwood stems were not matched and were left out of the comparison for the camera method. A 17.3 ft diameter measurement from one stem was also excluded from comparison because it was missing in one of the data sets. Differing sampling methods also created different sample sizes among the other diameter categories; therefore, only diameter measurements taken at heights corresponding with the truth data were compared.

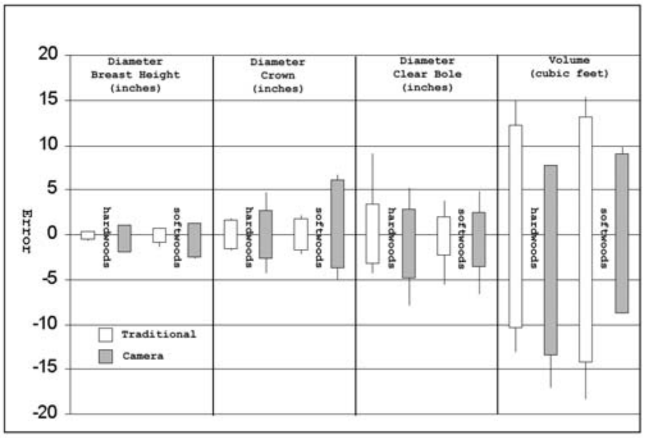


Figure 2—Errors for camera and standard methods for varying stem characteristics. The box indicates a 95 percent confidence interval about the mean (assuming normal distributions), and tails indicate the maximum and minimum observations.

**Table 1—Difference statistics (method – felled tree) by category for hardwoods and softwoods**

| Division | Method | n | Mean | Stnd.Dev. | Minimum | Maximum |
|---|---|---|---|---|---|---|
| *Diameter at breast height (inches)* | | | | | | |
| Hardwood | Standard | 20 | -0.08 | 0.19 | -0.80 | 0.10 |
| | Camera | 20 | -.46 | .77 | -1.90 | 1.00 |
| Softwood | Standard | 20 | -.12 | .44 | -1.40 | .80 |
| | Camera | 16 | -.71 | .97 | -2.70 | 1.00 |
| *Clear bole diameters (inches)* | | | | | | |
| Hardwood | Standard | 159 | .10 | 1.66 | -4.30 | 9.00 |
| | Camera | 166 | -.96 | 1.98 | -7.90 | 5.30 |
| Softwood | Standard | 173 | -.15 | 1.11 | -5.60 | 3.70 |
| | Camera | 145 | -.54 | 1.60 | -6.70 | 4.80 |
| *Crown diameters (inches)* | | | | | | |
| Hardwood | Standard | 69 | -.03 | .84 | -1.80 | 1.80 |
| | Camera | 151 | .03 | 1.40 | -4.40 | 4.60 |
| Softwood | Standard | 81 | .01 | .95 | -2.10 | 2.10 |
| | Camera | 93 | 1.15 | 2.52 | -5.00 | 6.60 |
| *Height (feet)* | | | | | | |
| Hardwood | Standard | 20 | .94 | 3.22 | -5 | 9 |
| | Camera | 20 | -2.32 | 7.26 | -15 | 9 |
| Softwood | Standard | 20 | .11 | 5.79 | -7 | 15 |
| | Camera | 16 | .04 | 9.41 | -14 | 16 |
| *Volume (cubic feet)* | | | | | | |
| Hardwood | Standard | 20 | .88 | 5.78 | -13.16 | 14.94 |
| | Camera | 20 | -2.90 | 5.39 | -17.04 | 3.20 |
| Softwood | Standard | 20 | -.50 | 6.98 | -18.39 | 15.31 |
| | Camera | 16 | .12 | 4.57 | -8.65 | 9.79 |

## Diameters at Breast Height

For diameters at breast height the standard method is more accurate for all taxonomic divisions (fig. 3). Within this category there is a concern that contact dendrometer (mechanical caliper or diameter tape) measurements, acquired to determine tree size class, were substituted as the optical dendrometer measurements for the standard method. The existence and severity of this operator bias could not be documented, but is mentioned as a possibility.

Discrepancies greater than 1 in. using the camera method were all negative and occurred primarily on larger stems (fig. 3). A number of probable causes exist for the variability of the d.b.h. errors. A high frequency of understory vegetation can cause improper range data collection. The range filter used in the diameter processing software is not sensitive to occlusions within a certain threshold (5 ft) of the estimated stem face range. Correction for measuring range to the stem face instead of the stem axis may have a

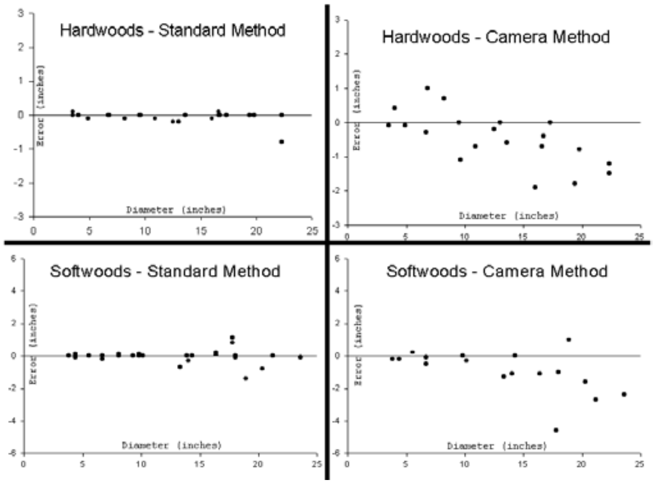


Figure 3—Diameter at breast height errors.

slight effect. Difficulty in locating stem edges due to foreground occlusion or background obfuscation may have been caused by the understory vegetation. Observations made on unimpeded open field targets indicate an instrument bias correlated with the inclination angle.

## Clear Bole Diameters

Multiple comparison tests (table 2) demonstrated no significant differences among the softwood means for diameters on the clear bole, but the camera was significantly lower than the two other methods for hardwoods. Extreme negative values are present predominantly for large diameters (fig. 4) and low heights (fig. 5). All values greater than 24 inches were situated below breast height and are subject to the same error explanations as d.b.h. These lower height measurements had small effects on the volume determination (especially if useable volume was to be considered), due to the short lengths with which these measurements were associated. Some extreme

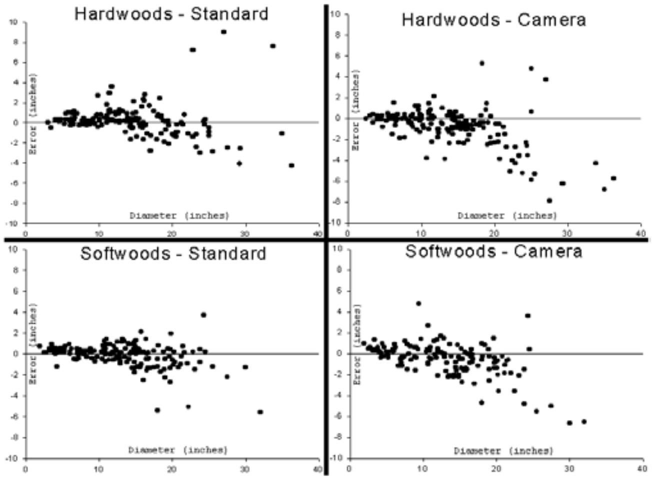


Figure 4—Clear bole diameter errors.

**Table 2—Bonferroni Multiple Comparision test results from a randomized block design analysis on the inventory attributes**

| | | Least squares means | | |
|---|---|---|---|---|
| Division | P-value | Camera | Standard | Fell |
| | - - - - - - - - *Diameter at breast height (inches)* - - - - - - - | | | |
| Hardwoods | 0.0045 | 12.22 a | 12.60 b | 12.68 b |
| Softwoods | .0103 | 12.23 a | 12.74 ab | 12.94 b |
| | - - - - - - - - - - - - *17.3 diameter (inches)* - - - - - - - - - - - - | | | |
| Hardwoods | .0009 | 10.68 a | 11.53 b | 11.32 b |
| Softwoods | .1852 | 10.44 a | 10.86 a | 10.75 a |
| | - - - - - - - - - *Clear bole diameters (inches)* - - - - - - - - - | | | |
| Hardwoods | .0026 | 12.34 a | 13.27 b | 13.19 b |
| Softwoods | .0668 | 11.67 a | 12.11 a | 12.25 a |
| | - - - - - - - - - - - Crown diameters (*inches*) - - - - - - - - - - - | | | |
| Hardwoods | .8162 | 6.83 a | 6.68 a | 6.58 a |
| Softwoods | .4271 | 7.27 a | 6.58 a | 6.73 a |
| | - - - - - - - - - - - - - - - - *Height (feet)* - - - - - - - - - - - - - - - | | | |
| Hardwoods | .1102 | 65.60 a | 68.85 a | 67.92 a |
| Softwoods | .981 | 67.31 a | 66.94 a | 67.28 a |
| | - - - - - - - - - - - - - *Volume (cubic feet)* - - - - - - - - - - - - - | | | |
| Hardwoods | .0185 | 36.38 a | 40.16 b | 39.28 |
| Softwoods | .3663 | 41.52 a | 39.70 a | 41.40 a |

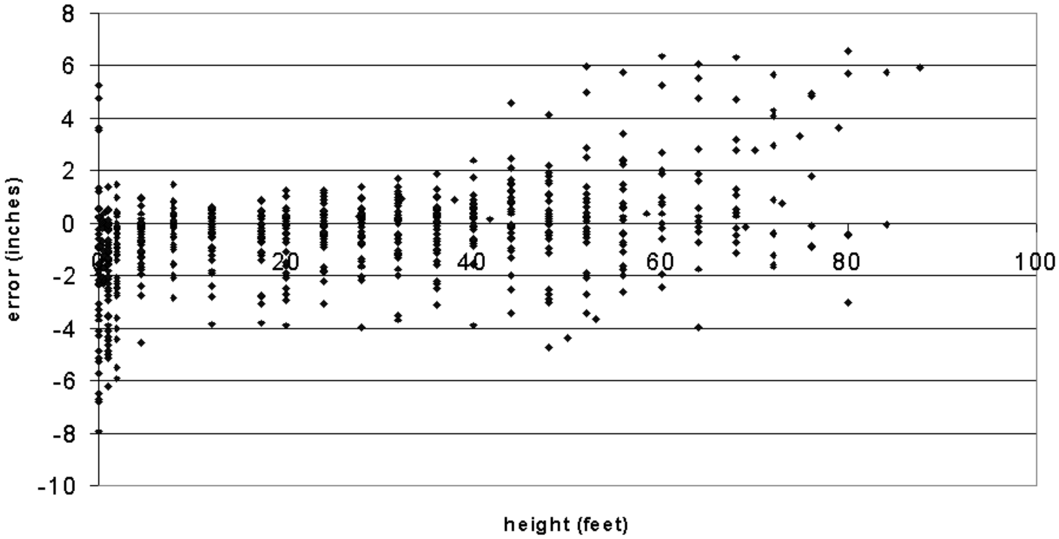


Figure 5—Camera diameter errors by height.

positive values in the hardwoods using the standard methods may have been the result of estimation due to exceeding the range of the pentaprism.

### Crown Diameters

Using the standard method, variance (fig. 6) seemed to be independent of dimension in the crown measurements. The multiple comparison tests (table 2) did not indicate significant differences among the means, though standard errors (fig. 6) were less for the standard method. Camera mean errors were positive for both hardwoods and softwoods, and extreme positive errors can be found in the softwoods. Twenty of the 22 camera errors >4 inches (fig. 6) were from three stems, so there is a potential for the camera estimates to have a correlated error. Especially among the hardwood crown diameters, the differing data collection protocol of the standard method greatly influenced the number of samples that could be compared.

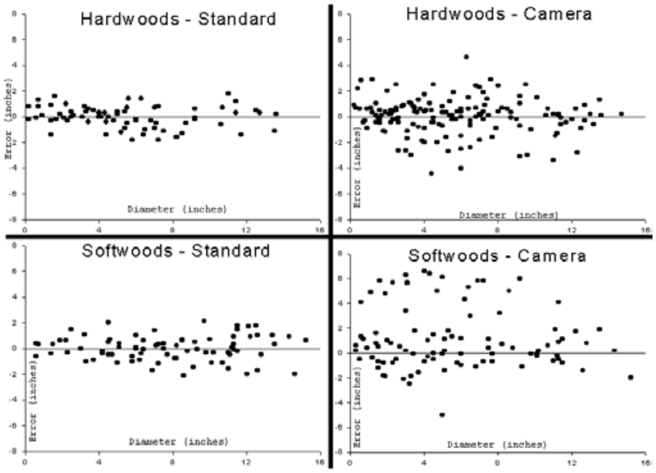


Figure 6—Crown diameter errors.

### Height

Total tree heights (actually stem length) ranged from 23 to 93 ft. Table 1 reveals that the camera variance is greater; however, the multiple comparison tests (table 2) do not reveal any significant differences among the means. It is thought that the camera method may improve for the hardwoods in a leaf-off condition if a better range to the top can be achieved. Then, by taking this length from ground to highest tip and projecting it back over to the plumb stump axis, a more accurate total tree height can be calculated.

### Volume

The hardwood least squares means between the camera and standard methods are significantly different from each other based on the multiple comparison tests (table 2), though neither method is significantly different from the true mean. Although the results using the standard method were better for each separate category previously mentioned, the volume results were slightly less satisfactory than the camera measurements. The camera and felled-tree methods used the 4 ft incremented diameters in order to calculate volume, whereas the standard method used shorter segments in places where forks or significant taper changes were exhibited. This may provide a partial explanation of this unexpected outcome. On the camera side, the negative trend still exists in the hardwoods. However, on some softwood stems the negative lower diameter errors are offset by positive upper-stem errors.

### CONCLUSIONS

For heights or diameters alone, the camera method was not quite as accurate as the standard method. For volume, the camera method was marginally better than the standard method. A summary of the multiple comparison tests (table 2) shows that the standard method least squares means do not differ significantly from the felled-tree method least squares means in any category. Camera method diameters on the lower portion of the stem are generally lower than the true diameters. It is evident that extreme diameter measurement errors are related to their position on the stem. Negative errors occur predominantly below breast height, and most of the extreme positive errors occur in the crown. Sources of instrument and methodological bias are being investigated. If the biases present at the two extremities of the stem can be removed, this instrument can produce some results comparable to the standard methods with considerable timesavings in the field.

## FUTURE WORK

A few hardware improvements are needed, including the integration of the ranging and video data streams and incorporation of the video tape recorder into the ruggedized unit, before the camera can be considered a reliable production instrument. The information extraction suite of algorithms needs to be expanded to provide increased automation. Image matching can be implemented to photogrammetrically determine heights, providing greater accuracy for spatial measurements, and edge detection can be used to eliminate manual coordinate capture. Algorithms can also be developed to incorporate more variables of interest, such as crown characteristics, biomass models, and stem quality.

## ACKNOWLEDGMENTS

The authors thank the reviewers, Susan Schmidt, Richard Oderwald, and Randolph Wynne, for their attention and constructive comments.

## REFERENCES

**Bradshaw, F.J.** 1972. Upper stem diameter measurements with the aid of 35mm photographs. Australian Forest Research. 6(1): 17–20.

**Clark, N.A.** 1998. An assessment of the utility of a non-metric digital camera for measuring standing trees. Blacksburg, VA: Virginia Polytechnic Institute and State University. 119 p. M.S. thesis. [Available on the Web at: http://scholar.lib.vt.edu/theses/available/etd-110298-111109/]. [Date assessed: unknown].

**Clark, N.A.** 2000. Initial results from a video-laser rangefinder device. In: proceedings for the 3rd southern forestry GIS conference; October 10–12, 2000. Athens, GA. 6 p.

**Crosby, P.; Barrett, J.P.; Bocko, R.** 1983. Photo estimates of upper stem diameters. Journal of Forestry. 81(12): 795–797.

**Juujärvi, J., Heikkonen, J.; Brandt, S.; Lampinen, J.** 1998. Digital-image-based tree measurement for forest inventory. SPIE conference on intelligent robots and computer vision XVII: algorithms, techniques, and active vision. Bellingham, WA: The International Society for Optical Engineering: 114–123. Vol. 3522.

# LAND USE, RECREATION, AND WILDLIFE HABITATS: GIS APPLICATIONS USING FIA PLOT DATA[1]

**Victor A. Rudis**[2]

**Abstract**—Spatial contexts govern whether and how land is used. Forest surveys inventory land uses from sampled plots and provide common forest resource summaries with limited information about associated nearby uses, or the landscape context. I used the USDA Forest Service's Forest Inventory and Analysis program of the South-Central States survey region (Alabama, Arkansas, Louisiana, Mississippi, east Oklahoma, Tennessee, and east Texas) to derive landscape context information. Methods employed moving averages (statistical combinations of sample plot observations with those from adjacent sample plots) to portray the spatial context, or "neighborhood" for forest resource appraisals. The survey region had 32,000 plots with land use information, and half of the plots classed as forest land provided more detailed information. Results yielded regional maps with displays of high and low probability of common land uses. For forest land, attributes shown include roads, forest fragment size, and hunting signs. Models of land use "hot spots" of competing and complementary uses are provided, forest land attributes important to selected recreational opportunity and wildlife habitat appraisals are discussed.

## INTRODUCTION

The clearing of extensive forested areas for agricultural use was once a common practice in the United States (Williams 1989). Deforestation of this magnitude is not as widely practiced today, but anthropogenic influences continue to affect remnant, as well as regenerated forests. Livestock grazing intrudes upon otherwise exclusive forest land use in pasture-dominated regions (Rudis 1998, 2000). Major roads, urban and built-up land, and associated higher population densities encroach on an otherwise rural forested landscape, thereby reducing timber harvests (Barlow and others 1998). This urban sprawl frequently clashes with other rural land needs (Befort and others 1988).

Silvicultural management regimes, as well as complementary and competitive income sources, may also differ among regions dominated by neighborhoods with important recreational, urban, or agricultural pursuits, or with habitats of critically endangered wildlife populations. Depending on user demand and quality of neighborhood resources (e.g., scenery, game), income from nontimber forest enterprises can vary widely. In the southern United States, for example, lease fees in 1989 for hunting alone averaged between $1 and $15 per acre per year (Thomas and Shumann 1993). Forests near areas with high population densities are unlikely commercial wood sources as nontimber uses (e.g., aesthetics, real estate) may outweigh their use for timber production (Wear and others 1999).

The USDA Forest Service's Forest Inventory and Analysis (FIA) program monitors the status and change in forest land and provides sample-based information about forest resources. Commonly, FIA data users analyze and summarize FIA data from sample plots but often ignore the context, that is, the "neighborhood" of the samples. In addition, attributes that index nontimber forest products and uses are not widely known, such as those associated with recreation opportunities and wildlife habitats.

The main objective of this study was to illustrate the use of landscape context attributes for forest resource appraisals. A second objective was to consider the importance of selected attributes for recreation and wildlife habitat appraisals at the landscape level of analysis. Forest attributes included in this paper include roads, forest fragment size, and hunting signs.

Roads provide access to forests for passive uses like sightseeing and for extractive uses like timber harvesting. Roads and allied roadside vegetation management also alter the wildlife habitat value of forests (Forman and Deblinger 1998). For example, forests with extensive roads are less likely to support viable populations of black bear (Rudis and Tansey 1995) and snakes (Rudolph and others 1998). Roaded forests, by definition, are also less likely to support primitive recreation opportunities, such as hunting and backpacking (USDA Forest Service 1982).

Forest fragment size is inversely related to population density (Rudis 1998). Among bottomland hardwood forests, large fragments (>1,000 ha) are comparatively wetter and older. Large bottomland hardwood fragments contain fewer human intrusions than small (<100 ha) fragments (Rudis 1995). Large forest fragments are in short supply. They are valued for primitive recreation opportunities (Rudis 1987, 1995) and are key habitats for wildlife in need of seclusion from humans (e.g., black bear, venomous snakes) or requiring large expanses of forest land (e.g., Cerulean warbler). Simply because of their size and their scarcity, large forest fragments offer economic opportunities like tourism as well as reserves of future timber supplies. Small fragments have lower potential for a variety of resources but may be suited to other uses, such as picnicking, thermal cover for livestock, and windbreaks in agriculture-dominated neighborhoods.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, USDA Forest Service, Southern Research Station, 201 Lincoln Green, Starkville, MS 39759.

Hunting signs observed and associated with forested areas index a number of phenomena. At the very least, such an index shows the prevailing cultural practice and landowner concern about hunting activities. The presence of more abundant signs in one region or time period than in others also suggests (1) a concentration of landowners with hunting interests or leased land by hunt clubs, (2) greater apprehension over landowner liability, (3) concern for trespass by hunters, (4) conflict between landowners and sportsmen regarding hunting activities, and (5) a shortage in the supply of hunting areas relative to demand. An increase in sign density between surveys suggests a change in landowner attitudes toward hunting activities and a decline in subsistence hunting opportunities by low-income residents.

## METHODS

Data used were from the U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis (FIA) surveys conducted between 1988 and 1995 for the South-Central States (Alabama, Arkansas, Louisiana, Mississippi, east Oklahoma, Tennessee, and east Texas). FIA sampled land use systematically in a three-phase design involving forest and nonforest determinations from aerial photos, a check of photointerpretation for a portion of these, and a ground sample of a still smaller subsample (Miller and Hartsell 1992.) Although sampling with aerial photographs was 25 times more intensive, geographically referenced observations were available only for ground-sampled plots. Because of this lack of geographic referencing, other sample information was not included.

Each of the 32,000, 0.4-ha ground-sampled plots had an approximate latitude and longitude from reference maps. Samples were spaced at 4.8 km intervals. FIA crews obtained more detailed attribute information on about 17,000 plots classified as forest land. The definition of forest land included areas 0.4 ha and larger, >37 m in width, and not developed for nonforest uses.

To generate land use and forest attribute maps from ground-referenced information, I created a digital map of forest inventory plots (positional accuracy better than 0.8 km) to a geographic surface with the aid of ArcView geographic information science (GIS) software and maps (ESRI, Inc. 1996a, 1999.) Further details and additional examples of early results appear elsewhere (Rudis 1991, 2000, and in press.)

I transferred plot attribute information to 2.4-km grid cells oriented in cardinal directions to increase computation efficiency. The dimension of the grid cell was "small enough to define the most detailed geographic feature" (ESRI, Inc. 1996b), yet large enough to minimize computer memory storage space and software calculations. With a 2.4-km grid, I nominally assigned every plot to a unique grid cell.

For indicator attributes, I recoded observations as 0=absent, or 1=present, averaged the observations for a given range of samples, and obtained average probability of occurrence, in percent. For interval attributes, I used attribute values themselves to calculate averages and compared average values using standard deviations above and below the mean. When needed, I transformed the values to obtain a normalized frequency distribution.

### Moving Averages

Averages per grid cell were estimates from plots within a circle of a given radius. The term is referenced as a "spatially moving average." Grain size was defined as the radius of the circle used to calculate the spatially moving average. Only grid cells associated with sample plots contributed to the averages. A grain size of 4.8-km radius encompassed about 13 of the 2.4-km grid cells (7,240 ha), up to five of which contained FIA sample plots. For land use occurrence probability estimation, results yield an occurrence probability of 0, 1, 2, 3, 4, or 5, out of five samples.

The spatially moving average yields an isotropic probability for forest land. Small-scale aerial photographs, however, show forest land as more frequent along the direction of steep terrain and adjacent to water courses, and as associated with particular soils and climates. The simplified averaging procedures used in this report ignored them. Furthermore, because FIA locations of sample plots were on a regular grid, I made no extensive examination of alternative grain sizes.

I employed the circular neighborhood mean statistics function within ArcView with the Spatial Analyst extension (ESRI, Inc. 1999). Calculation of means for forest-collected attributes provided averages for adjacent nonforest land, a scenario in classical statistics comparable to drawing a regression line beyond the range of the sampled region. To mask these areas, I created a layer of grid cells with 20 percent or less forest land probability, based on a 4.8-km radius grain size and averaged from all sampled plots. For the seven-State FIA survey region, this nonforest mask included extensive areas of nonforest land in the Mississippi Alluvial Plain. I also masked out areas with no FIA plot samples (nonsurveyed locations and largely nonforested counties in western Oklahoma, western Texas, and extreme southern Louisiana).

For forested plot attributes, I used a radius of 24 km—a size with resolution suitable for multicounty decisions, e.g., multiagency, Federal, and regional planning. The 24-km radius grain size approximated the size of a county planning area, or portion of a large city, which a larger grain size could obscure. This grain size is likely coarse for local management purposes, but provides broad contextual information. The 24-km grain size yielded averages nominally represented by 25 forested plots. Exceptions were in sparsely sampled, sparsely forested regions, where averages were based on fewer samples.

### Kriging

Unlike the more straightforward "averaging" technique listed above, kriging is memory intensive for large data sets. Kriging yields grid-cell averages based on a distance-weighting scheme, with the nearest sample plots, typically 16, contributing the most information. The radius specified is large to ensure that averages are based on 16 plots, even in a sparsely populated sample region. Contrary to moving average interpolation, changing the radius will usually yield only small differences in resulting patterns.

The analyst must choose among several weighting schemes, based on an examination of the geographic relationship of sampled values and sometimes by knowledge of the spatial association of phenomena under study. In a linear weighting scheme, the value at one grid cell location corresponds directly (1 to 1) with the value at an adjacent grid cell location.

I converted indicator plot values to an indicator probability surface interpolated with kriging, using ArcView with the Spatial Analyst extension and GS+ software (Gamma Design Software 1998). GS+ interpolated the surface with a 2.4 km grid.

### Field Attributes

FIA field crews made a general determination of land use on each sample plot. On forest land, they collected traditional timber variables, such as stand diameter class (i.e., stand size in timber reports), forest type, harvest activity, owner class, site productivity, and stand origin. Crews also collected nontraditional variables like livestock grazing, presence of trash, and proximity to urban and built-up land. Maps of other results based on moving averages appear elsewhere (Rudis 2000, and in press).

Selected results in this paper include the use of moving averages to assess forest land, forest fragment (patch) size, road proximity, and hunting signs. More recent results include averages using kriging, with overlays of county-based ecological subregion boundaries (Rudis 1999) to highlight regional, within-State differences. The example includes a spatial prediction of predominant land use.

**Land use**—The classification of land at 0.4-ha sample plots by use classes. Categories were forest, cropland, pastureland, urban and other land uses, marsh, and noncensus water. Definitions follow Anderson and others (1976) land use classifications. FIA survey manuals describe additional details (FIA Staff 1994).

**Forest fragments**—Contiguous forest cover unbroken by nonforest cover. A “contiguous” forest meant a patch of forest unbroken by water or nonforest land cover >37 m wide, as determined by field visits and the aid of 1:58,000 scale high-altitude color-infrared aerial photographs. Forest fragment size classes (and midpoints used in averages) were 0.4 to 4.0 ha (midpoint 2 km), 5 to 20 (12), 21 to 40 (30), 41 to 202 (121), 203 to 1,012 (607), 1,013 to 2,023 (1,518), and >2,023 (set at 3,323 ha). Because the frequency distribution of forest fragment size class was lognormal, I calculated averages using logarithm-transformed midpoint values. Though one fragment could be large enough to be associated with more than one sample plot, I assumed every plot was a different fragment.

**Roads**—Travel corridors associated with vehicular transportation. From the sample plot to the nearest road, FIA field crews measured proximity in 30 m intervals to 1600 m (100 ft intervals, to 5300 ft). FIA field crews judged roads as capable of travel by four-wheel drive vehicles, termed “truck-operable or better” roads.

**Hunting signs**—Signs encountered by field crews within 400 m (1,320 ft) of a sample plot. These signs listed “no hunting,” “hunting restricted,” or “posted” and were commonly associated with the sample plot.

## RESULTS AND DISCUSSION

Figure 1 illustrates the distribution of forest land by sample plot location and by interpolated forest land probability. Forest land probability was low in the Mississippi Delta (western Mississippi, east Louisiana, and eastern Arkansas) and other predominantly agricultural areas along major rivers. Forest land probability was also low in the Blackland Prairie crescent spanning the States of Mississippi and Alabama. Forest land probabilities were higher in other areas.

### Forest Fragments

I used forest fragment size class to illustrate an example of an interval attribute. Figure 2 depicts the spatial distribution of mean fragment size. Most of the large fragments were either in mountainous areas, such as the Boston Mountains of the Ozark National Forest, or in low-lying areas, such as the Atchafalaya Basin of Louisiana. Black bear occupy many of these same sites (Maehr 1984). The most fragmented forests (Memphis, Central Tennessee, Longview [Texas], and agriculture-dominated areas) do not contain black bear.

Large fragments that occur on Federal and State land serve as habitats for wildlife in need of seclusion and provide primitive recreational opportunities. Results indicate that the public agencies associated with these areas have been successful in conserving these uses.

Figure 3 (see page 134 ) illustrates that roaded forests near roads were abundant throughout the South and more abundant in selected regions. An extensive road network appears near forests throughout south Mississippi and parts of other States, particularly within the Southern Coastal Plain. By contrast, roadless forested areas were rare. The only extensive roadless forested area was in Louisiana’s Atchafalaya Basin, which suggests that its scientific and ecological value may surpass its value for forest production or development.

### Land Use

The last two examples employ kriging to depict land use in east Texas. Figure 4 (see page 135) shows land use in east Texas for the 1992 FIA survey. A linear weighting scheme yielded the highest r-square autocorrelation (> 0.65) and lowest residual sums-of-squares.

Pastureland dominated in the western part of east Texas, urban areas to the southwest (Houston area), and forests to the east. Given these patterns, one might logically expect livestock to use forests in pasture-dominated areas. In fact, livestock grazing occurred on a third of the forests in the western ecological province, compared with 10 percent throughout the south central region (Rudis 1998). The urban-dominated areas encompassed the outskirts of major cities—the most prominent of which was in the southwest corner (Houston metro area, Harris County).

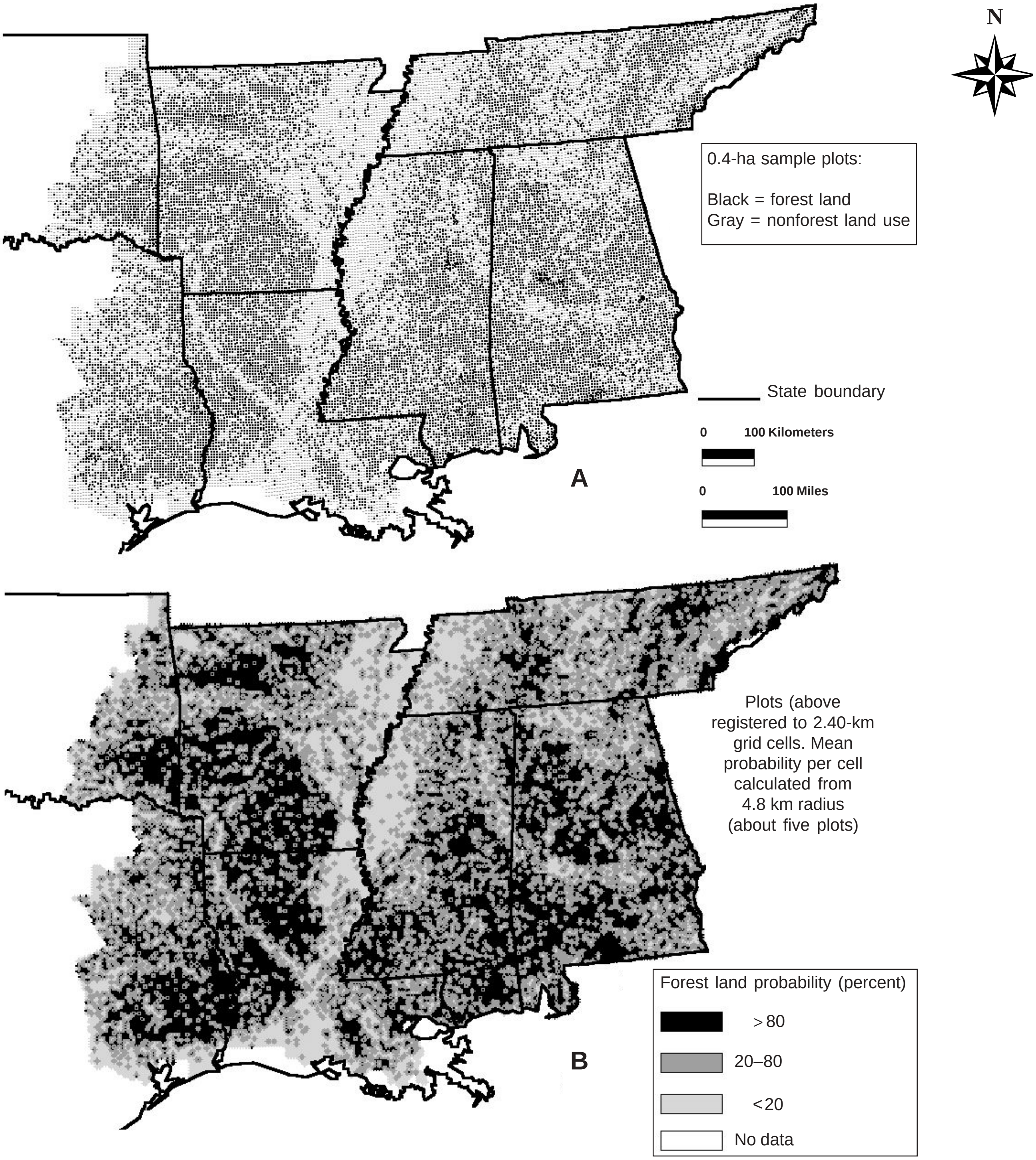

N
0.4-ha sample plots:
Black = forest land
Gray = nonforest land use
State boundary
0 100 Kilometers
0 100 Miles
A
Plots (above registered to 2.40-km grid cells. Mean probability per cell calculated from 4.8 km radius (about five plots)
Forest land probability (percent)
■ >80
▩ 20–80
□ <20
□ No data
B

Figure 1—Forest and nonforest land use, U.S. Forest Inventory and Analysis surveys. South-Central United States, 1988–1995: (A) sample plot locations, (b) forest area probability. Forest land probability was generated using a 4.8 km radius moving average and mapped to a 2.4 km grid.

Each of these patterns suggest that forests in nonforest-dominated areas serve more as shade for livestock or as landholdings for urban uses than as forests with continuing timber production potential. Hence, forest resource appraisals stratified by predominating land-use class—whether it is pasture, urban, or forest—will likely improve estimates of forest resource supply.

## Hunting Signs

Forest land with hunting signs represents 11 percent of the resource in the South-Central States FIA survey region (Rudis 1998). The spatial distribution of forests with hunting signs appears in figure 5 (see page 136) for two sample periods. In both surveys, hunting restrictions were more frequent in the northern half of east Texas. One

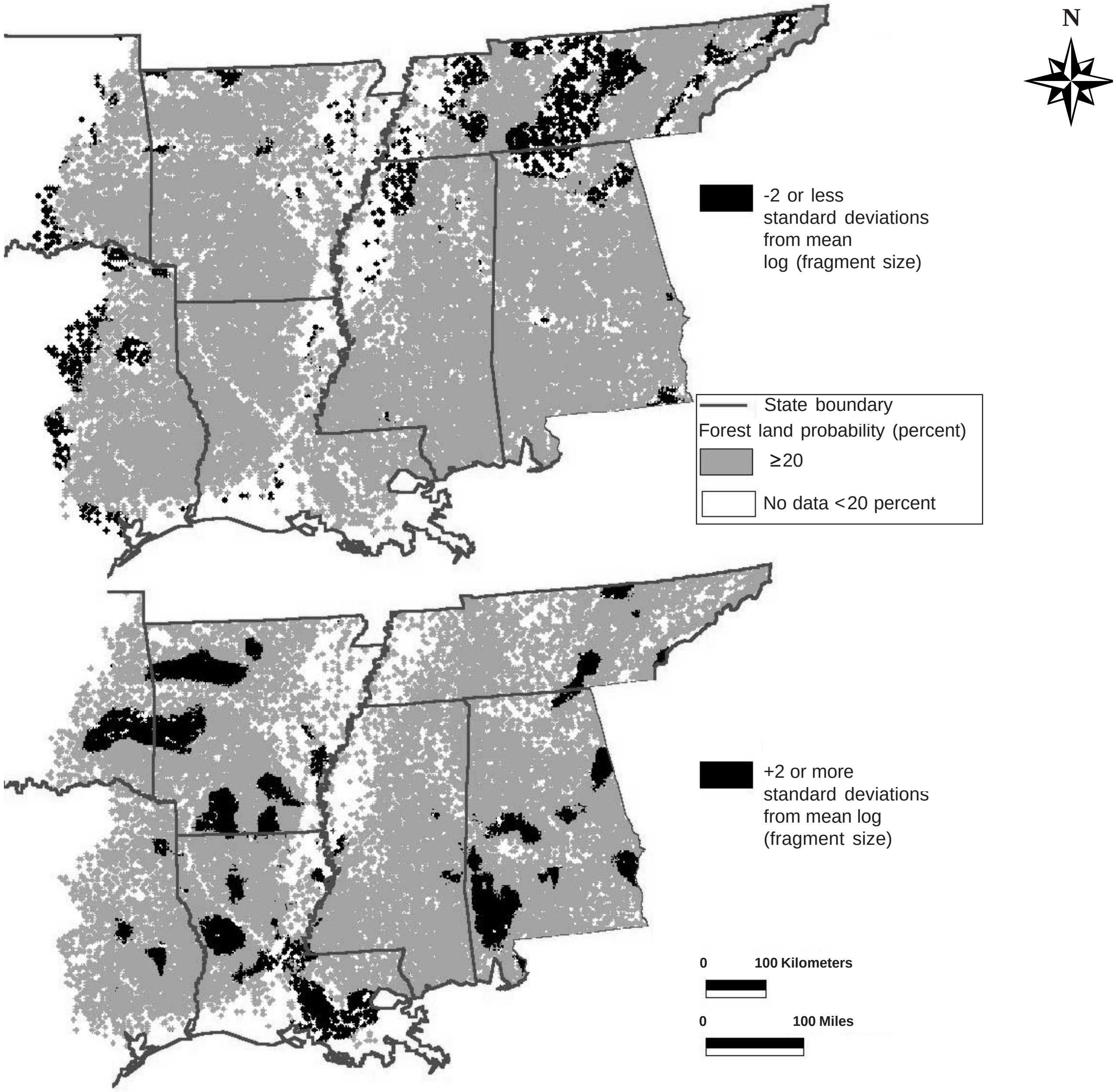


Figure 2—Deviation from average forest fragment class, forest land probability 20 percent or more, South-Central United States, 1988 to 1995 surveys. Averages were generated using a 24 km radius moving average and mapped to a 2.4 km grid.

hypothesis is that because both forests and public land areas are relatively limited to the north, landowners may be (1) selling more private land leases for hunting on a per-acre basis, and (2) resisting public use of forest land for hunting. Analysts need additional evidence, such as deer kill surveys, sportsman license sales, or landowner studies, to draw definitive inferences. The suggested increase in restricted forests for Wood County, TX, bears further investigation.

## FUTURE PROSPECTS

Critical to any mapping scheme is having geographically referenced observations from which to draw inferences. From a mapping standpoint, the more information received from all sampled plots, not just those visited on the ground and not just those having detailed attributes only for forested land, the better will be the resulting estimation of the “neighborhood.” Studies have already implicated road density, road proximity, and fragment size class in the distribution of wildlife populations (Rudolph and others 1998, Rudis and Tansey 1995) and recreation uses (Rudis 1987). The next steps in the analyses are to improve the reliability of these indices as surrogates for the number of recreation users, the percentage of landowners with specific intentions, and the number of hunters. Such improvements could take the form of user and landowner surveys.

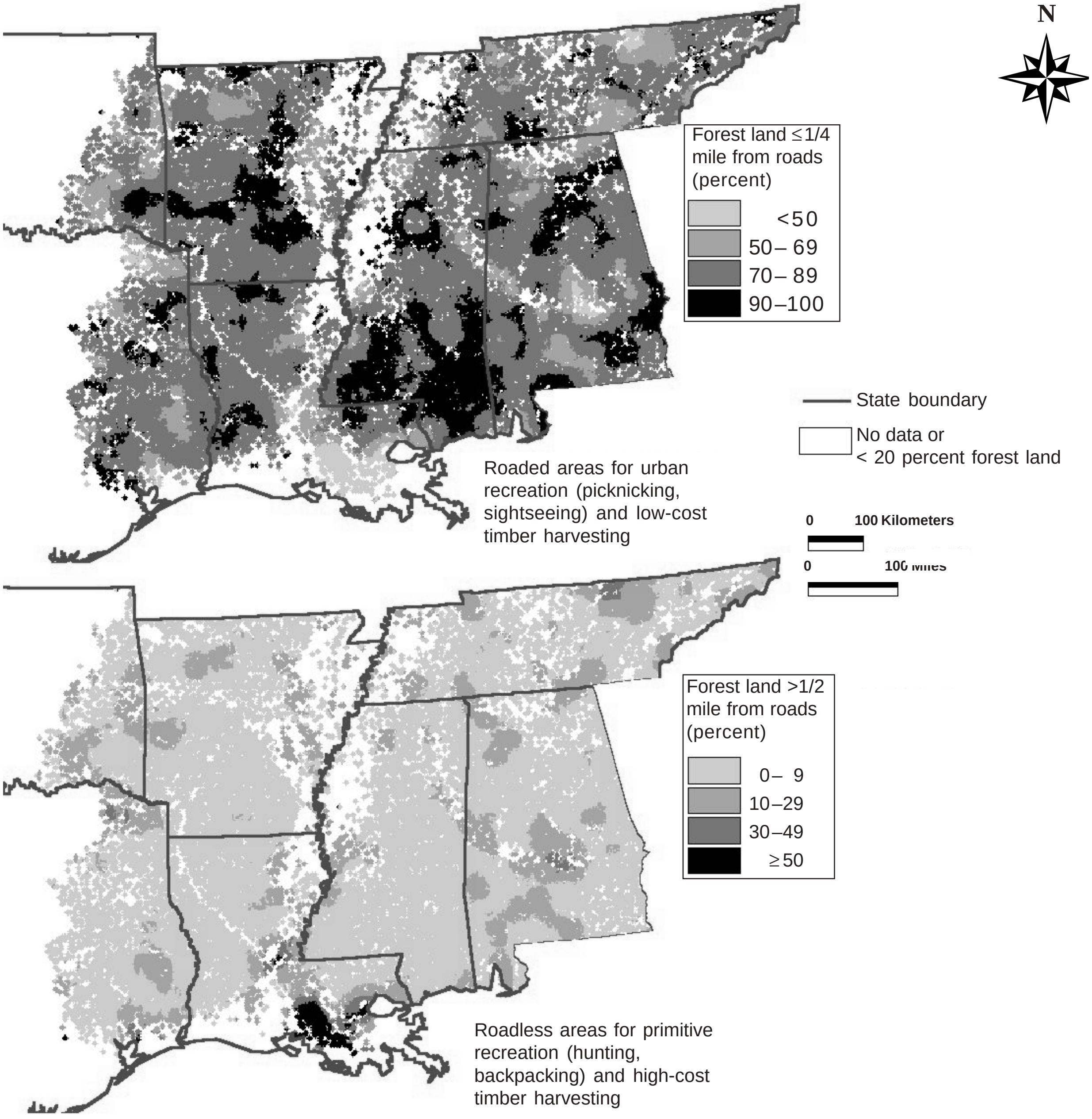

N
Forest land ≤1/4 mile from roads (percent)
<50
50– 69
70– 89
90–100
State boundary
No data or < 20 percent forest land
Roaded areas for urban recreation (picknicking, sightseeing) and low-cost timber harvesting
0 100 Kilometers
0 100 Miles
Forest land >1/2 mile from roads (percent)
0– 9
10–29
30–49
≥50
Roadless areas for primitive recreation (hunting, backpacking) and high-cost timber harvesting

Figure 3—Roadless and roaded forest probability, forest land probability 20 percent or more. South-Central United States, 1988-1995 surveys. Averages were generated using a 24 km radius and a 2.4 km grid interpolation.

Map certainty is the confidence of an attribute's value at a given location and attribute variation near the location. A straightforward appraisal of map certainty is to list the number of samples used to estimate the value of each grid cell. Among land-use attributes, map certainty is greater in regions with more samples. Among forest land attributes, map certainty is greater in regions with abundant forests. Map certainty in land use estimates is relatively uniform because FIA sampling is regularly spaced throughout the survey region. Certainty in mapped forest attributes information is lower in sparsely forested regions, such as the Mississippi Delta, and higher in densely forested regions, such as the Boston Mountains of Arkansas.

Relaxation of the isotropic forest probability assumption and incorporation of information from other data sources are other ways to improve resulting maps. Incorporation of classified digital imagery from satellite sensors permits a reduction in the grain size (and an increase in the resolution) of sensor-detected earth cover classes, while still providing thematic information from ground-sampled observations. Concurrence of prediction in attributes mapped from other geographically referenced, correlated data, such as soils, climate, and geology, boosts confidence in attribute variation for a given location.

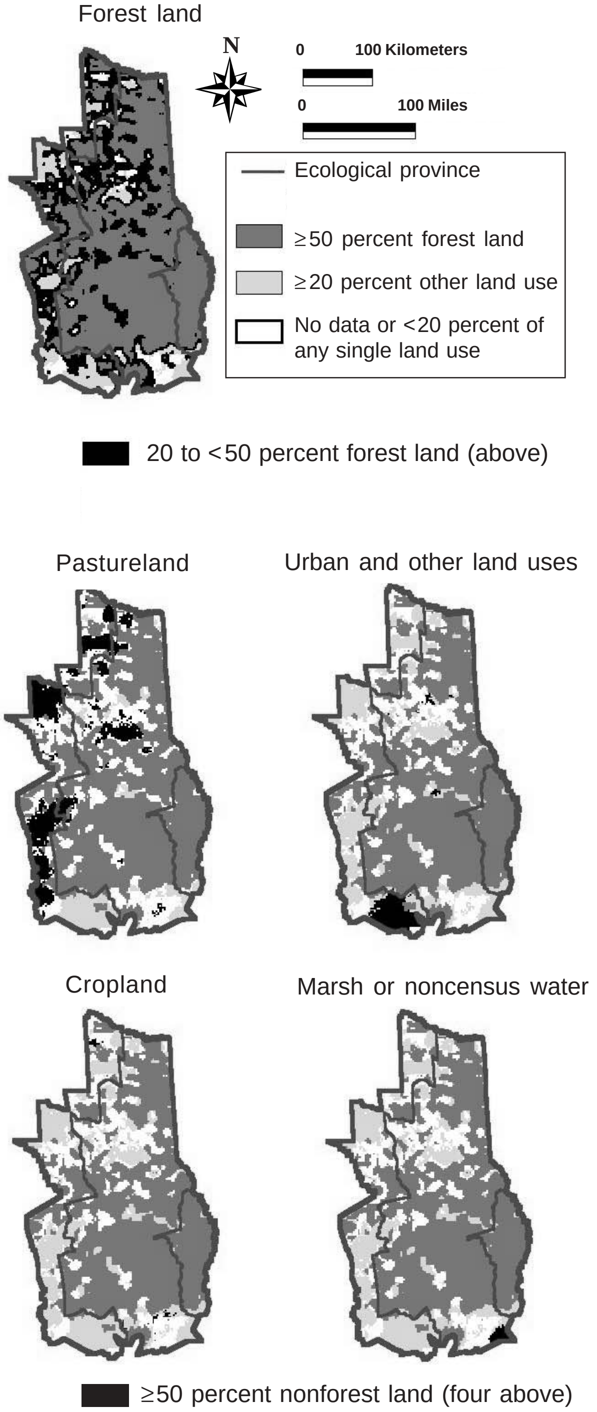

Forest land
N
0 100 Kilometers
0 100 Miles
Ecological province
≥50 percent forest land
≥20 percent other land use
No data or <20 percent of any single land use
20 to <50 percent forest land (above)
Pastureland
Urban and other land uses
Cropland
Marsh or noncensus water
≥50 percent nonforest land (four above)

Figure 4—Land use probability, east Texas Forest Inventory and Analysis survey. Averages were generated using kriging with a linear weighting scheme for the 16 nearest samples and a 2.4 km grid interpolation.

In conclusion, GIS provides additional tools for evaluating the spatial context of FIA plots for forest resource appraisals, particularly the context of associated nontimber values. Examples in this paper portray land use, wildlife habitat, and recreational opportunities. Knowing where the phenomena occur, even in general terms, provides the analyst with added information about likely timber supplies, occurrence “hot spots” of predominant and potentially competing resource uses, and change over time.

## ACKNOWLEDGMENTS

I thank Greg Reams for formally inviting me to make the presentation; and Bill Cooke and Andy Lister for constructive remarks on an earlier manuscript draft.

## LITERATURE CITED

**Anderson, J.R.; Hardy, E.E.; Roach, J.T.; Witmer, R.E.** 1976. A land use and land cover classification system for use with remote sensor data. U.S. Geological Survey Professional Paper 964. Washington, DC. 28 p.

**Barlow, S.A.; Munn, I.A.; Cleaves, D.A.; Evans, D.L.** 1998. The effect of urban sprawl on timber harvesting: a look at two southern States. Journal of Forestry. 96(12): 10–14.

**Befort, W.A.; Luloff, A.E.; Morrone, M.** 1988. Rural land use and demographic change in a rapidly urbanizing environment. Landscape and Urban Planning. 16: 345–356.

**ESRI, Inc.** 1996a. ESRI data and maps [Compact Disc]. Redlands, CA: Environmental Systems Research Institute, Inc. Vol. 1.

**ESRI, Inc.** 1996b. Working with the ArcView spatial analyst. Redlands, CA: Environmental Systems Research Institute, Inc.: 2–11.

**ESRI, Inc.** 1999. ArcView GIS: Vers. 3.2. Redlands CA: Environmental Systems Research Institute, Inc. [CD-ROM].

**FIA Staff.** 1994. Forest survey inventory work plan: Arkansas 1994-1995. Starkville, MS: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station, Forest Inventory and Analysis Unit. 61 p. + appendices.

**Forman, R.T.T.; Deblinger, R.D.** 1998. The ecological road-effect zone for transportation planning and Massachusetts highway example. In: Evink, G.L.; Garrett, P. [and others]. 1998. Proceedings of the international conference on wildlife ecology and transportation; 1998 February 10–12; Fort Meyers, FL. FL–ER–69–98. Tallahassee, FL: Florida Department of Transportation: 78–96.

**Gamma Design Software.** 1998. GS+: geostatistics for the environmental sciences. Professional ed., Vers. 3.1.7. Plainwell, MI: Gamma Design Software.

**Maehr, D.S.** 1984. Distribution of black bears in eastern North America. Black Bear Research and Management. 7: 74–75.

**Miller, P.E.; Hartsell, A.J.** 1992. Forest statistics for east Texas counties. Resour. Bull. SO–173. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station. 55 p.

**Rudis, V.A.** 1987. Recreational use of forested areas by Alabama residents. Res. Pap. SO–237. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station. 37 p.

**Rudis, V.A.** 1991. A recreation and landscape perspective of Alabama’s changing forest environment: the human character of forests. In: Jones, R.H., ed. Proceedings: Alabama’s forest resources: past, present and future; 1991 May 29–30; Auburn, AL. Auburn, AL: Auburn University, School of Forestry, Alabama Cooperative Extension Service: 93–114.

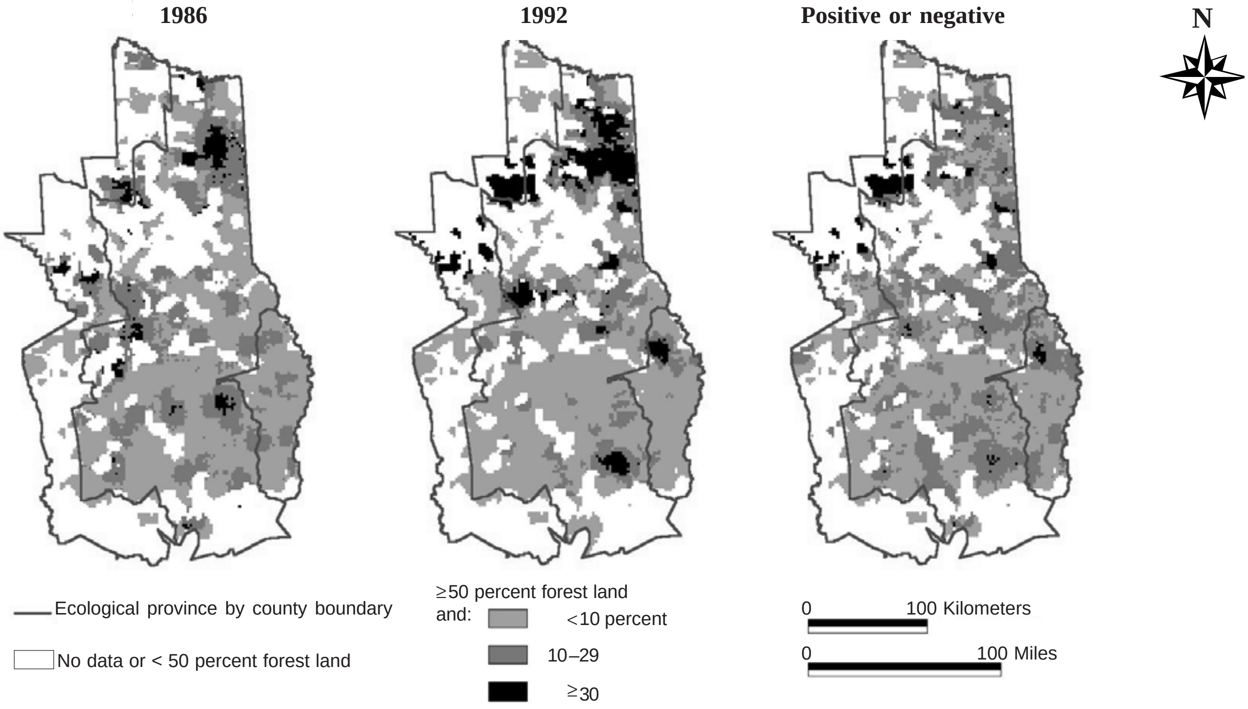

1986
1992
Positive or negative
N
Ecological province by county boundary
No data or < 50 percent forest land
≥50 percent forest land and:
<10 percent
10–29
≥30
0 100 Kilometers
0 100 Miles

Figure 5—Forests with restrictive hunting signs, forest land probability 50 percent or more, east Texas, 1986 and 1992 surveys, and change since 1986. Average probability of forests with restrictive hunting signs was generated using a 24 km radius and a 2.4 km grid interpolation. Forest land probability was generated with kriging, as in figure 4.

**Rudis, V.A.** 1995. Regional forest fragmentation effects on bottomland hardwood community types and resource values. Landscape Ecology. 10(5): 291–307.

**Rudis, V.A.** 1998. Regional forest resource assessment in an ecological framework: the Southern United States. Natural Areas Journal. 18(4): 321–334.

**Rudis, V.A.** 1999. Ecological subregion codes by county, coterminous United States. Gen. Tech. Rep. SRS–36. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 95 p.

**Rudis, V.A.** 2000. Using widely-spaced observations of land use, forest attributes, and intrusions to map resource potential and human impact probability. In: Hansen, Mark; Burk, Tom, eds. Integrated tools for natural resources inventories in the 21st century: an international conference on the inventory and monitoring of forested ecosystems; 1998 August 16–19; Boise, ID. Gen. Tech. Rep. NCRS–212. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 721–733.

**Rudis, V.A.** 2001. Landscape context and regional patterns in Arkansas' forests. In: Guldin, James M., tech. compiler. Symposium on Arkansas forests; 1997 May 30–31; Little Rock, AR. Gen. Tech. Rep. SRS–41. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station: 24–45.

**Rudis, V.A.; Tansey, J.B.** 1995. Regional assessment of remote forests and black bear habitat from forest resource surveys. Journal of Wildlife Management. 59(1): 170–180.

**Rudolph, C.D.; Burgdorf, S.J. [and others].** 1998. The impact of roads on the timber rattlesnake (*Crotalus horridus*) in eastern Texas. In: Evink, G.L. [and others]. Proceedings of the international conference on wildlife ecology and transportation; 1998 February 10–12; Fort Meyers, FL. FL–ER–69–98. Tallahassee, FL: Florida Department of Transportation: 236–240.

**Thomas, M.G.; Schumann, D.R.** 1993. Income opportunities in special forest products: self-help suggestions for rural entrepreneurs. Agric. Info. Bull. 666. Washington, DC: U.S. Department of Agriculture, Forest Service. 206 p.

**USDA Forest Service.** 1982. ROS users guide. Washington, DC: U.S. Department of Agriculture, Forest Service. 38 p.

**Wear, D.N.; Liu, R. [and others].** 1999. The effects of population growth on timber management and inventories in Virginia. Forest Ecology and Management. 118: 107–115.

**Williams, M.** 1989. Americans and their forests: a historical geography. New York: Cambridge University Press. 599 p.

# INTEGRATING P3 DATA INTO P2 ANALYSES: WHAT IS THE ADDED VALUE[1]

James R. Steinman[2]

**Abstract**—The Forest Inventory and Analysis and Forest Health Monitoring Programs of the USDA Forest Service are integrating field procedures for measuring their networks of plots throughout the United States. These plots are now referred to as Phase 2 (P2) and Phase 3 (P3) plots, respectively, and 1 out of every 16 P2 plots will also be a P3 plot. Mensurational methods will be identical on both types of plots, as will the procedures used for coding tree damages. Measurements of crown dieback, crown density, and foliage transparency; and measurements related to soils, lichens, and ozone indicators will distinguish P3 tree data from P2 tree data. Questions arise as to what value the unique P3 data add to reporting forest health conditions, and whether the P3 attributes can be extended to the greater number of P2 plots and forest landscape. This paper explores the latter question by showing how representative the P3 plots are of the forest as depicted by P2 plots. In empirical analyses of P2 and P3 data recently collected in Georgia, the P3 data were treated as a one-sixteenth subset of the P2 data. Stratifications of the data by forest-type group demonstrated that P3 plots were representative of the predominant forest-type groups and spatial distributions showed how the two types of plot data were comparable at different levels of resolution.

## INTRODUCTION

The Forest Inventory and Analysis (FIA) and Forest Health Monitoring (FHM) Programs of the USDA Forest Service are integrating field procedures for measuring plots throughout the United States. All FHM plots [herein referred to as Phase 3 (P3)] will be collocated on a systematic grid with 1 of every 16 FIA plots [herein referred to as Phase 2 (P2)]. Mensurational methods on both types of plots will be identical, and FHM procedures for recording tree damages will be applied to the P2 plots. Crown measurements of dieback, density, and foliage transparency in 5-percent classes will remain as the only attributes that distinguish P3 tree data from P2 tree data. Measurements related to soils, lichens, and ozone bio-indicators will also be unique to P3 plots, and other data related to woody debris, herbs, and shrubs will most likely be collected in the near future on these plots.

Integration of the P2 and P3 field procedures will also result in combined use of P2 and P3 attributes for the reporting of forest health conditions. In these analyses, the P3 data will be regarded as a one-sixteenth subset of the P2 plots with the additional attributes described above. This proposed use of the data gives rise to several related questions:

1. Do P3 plots represent the forest landscape as depicted by P2 plots?
2. What is the appropriate spatial scale of use for the P3 data?
3. Can attributes unique to the P3 plots be extended to the P2 plots?

Past analyses have addressed only the first two questions by showing that estimates of some attributes from P2 and P3 data are comparable at a regional scale of resolution (Brooks and others 1992). However, recent unpublished applications of the P3 data have demonstrated their use for smaller geographic areas. Given this interest, the objective of this paper is to explore the spatial relationships between P2 and P3 data at different scales.

## METHODS

Empirical data from Georgia were used in an analytical approach to compare spatial distributions of various attributes common to the P2 and P3 data. Georgia was selected as a case study because (1) recent years of measurement for P2 and P3 data closely coincide (1997 and 1995, respectively), and (2) sampled data distributions for the State are similar among forest-type groups and stand sizes (table 1). Analytical methods focused on whether spatial distributions of the P2 and P3 data were

**Table 1—Percentages of forest-type groups and stand size in Georgia as estimated by Phase 2 (P2) and Phase 3 (P3) data sources**

| | Data source and year | | |
|---|---|---|---|
| Forest-type groups and stand size | P2 1997 | P3 1995 | Difference P2 – P3 |
| | - - - - - - - - *Percent* - - - - - - - - | | |
| Forest-type groups | | | |
| Oak-hickory | 23 | 17 | 6 |
| Oak-gum-cypress | 16 | 13 | 3 |
| Oak-pine | 15 | 21 | -6 |
| Loblolly-shortleaf pine | 30 | 28 | 2 |
| Longleaf-slash pine | 14 | 19 | -5 |
| Stand size | | | |
| Sawtimber | 40 | 41 | -1 |
| Poletimber | 24 | 24 | 0 |
| Seedling-sapling | 36 | 35 | 1 |

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Research Forester, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804.

still comparable at higher levels of resolution within the State.

In 1995, P3 data were collected in Georgia on a systematic grid of hexagons wherein one P3 plot was located within each hexagon cell. In 1997, P2 data were collected from a nonsystematic plot network dissimilar to the new systematic grid of the Southern Annual Forest Inventory System (SAFIS) (Roesch and Reams 1999). Therefore, to make realistic comparisons between data from P2 and P3 plots, it was necessary to simulate the SAFIS grid using the 1997 data.

### Simulating the New P2 Grid

The new SAFIS P2 grid was simulated by choosing plots measured in 1997 that were nearest to the center of each grid cell. This technique populated about 4,500 of the 6,413 SAFIS grid cells with data. Empty cells occurred mostly in areas that were purposely undersampled in 1997, such as nonforested landscapes (e.g., the Atlanta area) and the Okefenokee Swamp. Aside from these areas, the simulation of the SAFIS grid produced a uniform and representative sampling intensity of 16 P2 plots for each P3 hexagon cell (fig. 1).

### Quantifying Spatial Associations between P2 and P3 Plots

The simulated grid was used to depict spatial distributions of P2 plots located within different forest-type groups throughout Georgia. Likewise, spatial distributions of P3 plots within corresponding forest-type groups were then overlaid for comparison. Considering that individual plots can occupy multiple forest conditions, plots were considered representative of a given forest stratum only if at least 50 percent of their sampled area was within that stratum. Plots sampling the oak-hickory and loblolly-shortleaf pine forest-type groups were used in example analyses because these groups are abundant yet unevenly distributed in Georgia.

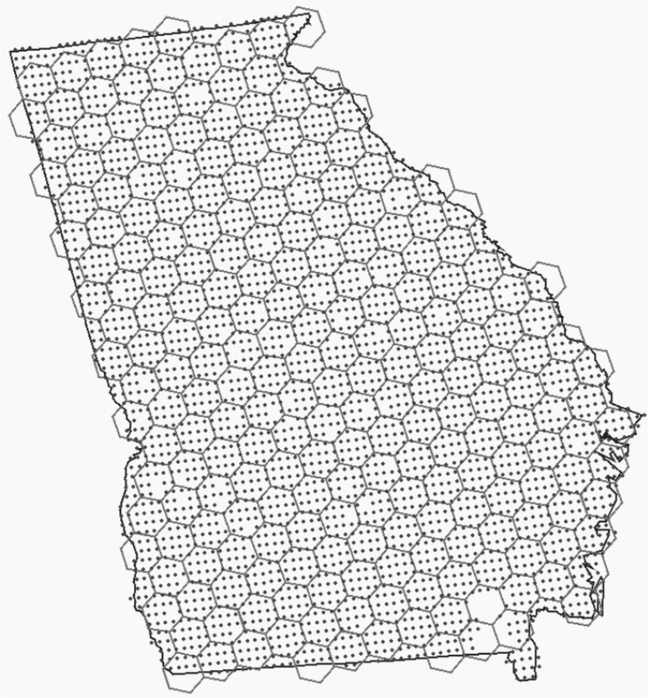


Figure 1—P3 hexagons in Georgia overlaid with a simulated grid of P2 plots using data from 1997 to simulate the one-sixteenth P3 sampling intensity of P2 plots.

The ratio of the number of P2 to P3 plots was calculated for each forest-type group at different spatial scales and compared to the base-grid ratio of 16:1. The smallest unit of area (highest resolution) used for analysis was a P3 hexagon grid cell, where the number of P2 plots representing a given forest stratum was compared with that depicted by the individual P3 plot for the cell. A tally of all counts was then used to examine how the classification of a P3 cell compared to that of each P2 plot contained within the cell. This Geographic Information System technique was easy to implement and provided an unbiased match between P2 plots and P3 hexagons.

In a similar manner, a coarser resolution was analyzed using a cluster of seven P3 hexagon cells, with one cell surrounded by six others. This technique involved classifying each cluster of seven cells, or "hepta-hexagon," according to the P3 attributes of the center cell, and then determining how many P2 plots within the cluster had matching attributes. Each P3 hexagon cell was evaluated, which resulted in a sequence of overlapping clusters equal in number to the number of individual P3 hexagons. However, only hepta-hexagon clusters located completely within Georgia were retained for analysis.

## RESULTS

### Ratio of P2 Plots per P3 Hexagon

A tally of all P2 and P3 plots that sampled oak-hickory forest conditions showed a total of 670 P2 and 29 P3 plots, equivalent to a ratio of 23:1. This deviation from the base-grid ratio of 16:1 corresponds to a slightly greater estimate of oak-hickory forest abundance obtained from the P2 data (table 1) and suggests that the P3 plots under-sampled the resource. For plots that sampled loblolly-shortleaf pine forest conditions, the ratio of P2 to P3 plots was 18:1, which was expected, considering that both types of plots provide similar estimates of the loblolly-shortleaf pine abundance in the State.

A visual display of locations of P2 and P3 oak-hickory plots illustrates their respective spatial distributions within Georgia (fig. 2). From these data it is evident that the number of oak-hickory P2 plots within each oak-hickory P3 hexagon is much less than 16. Conversely, a large number of oak-hickory P2 plots are located in areas not represented by oak-hickory P3 hexagons.

A cross-classification of the plot distributions quantifies the disparity in plot locations (table 2). All 29 oak-hickory P3 hexagons contain 10 or fewer oak-hickory P2 plots. Furthermore, about two-thirds of the hexagons contain fewer than six P2 plots. Conversely, a large number of oak-hickory P2 plots are located in areas where the nearest P3 plots are in other forest-type groups or are nonforested.

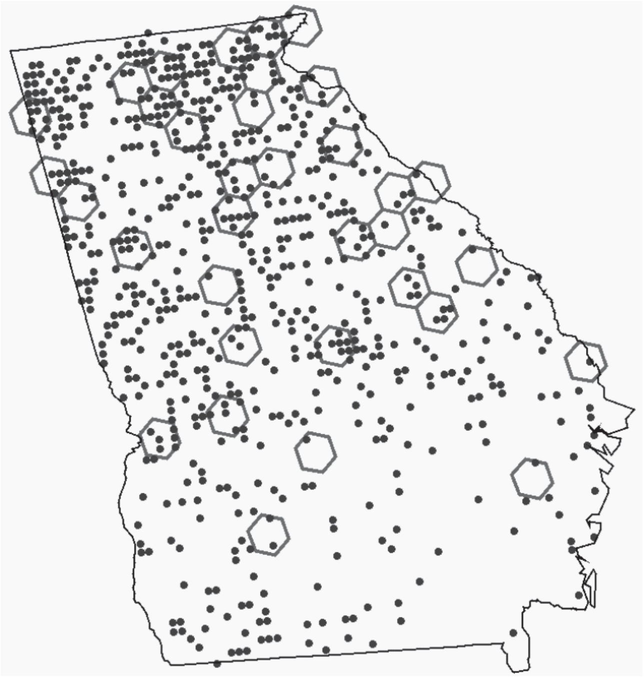


Figure 2—P3 hexagons and P2 plots, each representing sampled oak-hickory forest-type groups.

**Table 2—Spatial associations between Phase 2 (P2) and Phase 3 (P3) plot locations that sample oak-hickory forest-type groups**

| Number of oak-hickory P2 plots per P3 hexagon | Number of P3 hexagons | |
|---|---|---|
| | Oak-hickory | Other |
| 0 | 0 | 29 |
| 1–5 | 19 | 135 |
| 6–10 | 10 | 25 |
| 11–15 | 0 | 2 |
| >15 | 0 | 0 |

Similar results were found for the loblolly-shortleaf forest-type group. The arbitrary difference was that loblolly-shortleaf pine types are more abundant in the southern part of Georgia.

## Ratio of P2 Plots per Cluster of Seven P3 Plots

As expected, use of the clusters of seven P3 hexagons resulted in greater numbers of oak-hickory P2 plots within each hepta-hexagon cluster (table 3). All but three of the clusters with an oak-hickory P3 plot in the center hexagon cell also contained at least 16 oak-hickory P2 plots somewhere within the cluster. Visual inspection of the plot and cluster distributions also confirmed that the area defined by the hepta-hexagon clusters captured most of the P2 plots (fig. 3).

Results at this spatial scale were also similar in analyses of the distributions of the loblolly-shortleaf pine types.

**Table 3—Spatial associations between Phase 2 (P2) and Phase 3 (P3) plot locations that sample oak-hickory forest-type groups**

| Number of oak-hickory P2 plots per P3 hexagon | Number of P3 hexagons | |
|---|---|---|
| | Oak-hickory | Other |
| 0 | 0 | 0 |
| 1–5 | 1 | 22 |
| 6–10 | 1 | 24 |
| 11–15 | 1 | 23 |
| >15 | 16 | 81 |

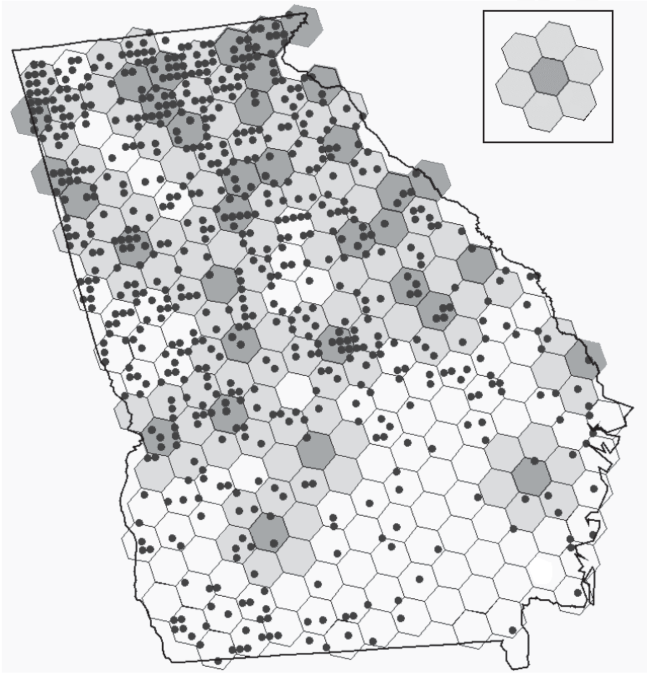


Figure 3—Clusters of seven P3 hexagons used to search a more extensive area for neighboring P2 plots that represent oak-hickory forest-type groups.

Locations of P2 plots in this stratum were strongly associated with seven-hexagon clusters that had a P3 loblolly-pine plot in the central cell.

## DISCUSSION

Although this study used only a subset of empirical data from Georgia, some general conclusions can be inferred from its analyses. Findings help quantify which spatial resolutions are suitable for use with the P3 data.

For a given P3 plot, the distance to neighboring P2 plots with similar attributes can be great. The Georgia data show that a low number of oak-hickory P2 plots are usually found within the hexagon cell of an oak-hickory P3 plot. This finding confirms that an individual P3 plot is not necessarily representative of its surrounding hexagon and that the P3 hexagons are thus not an appropriate level of resolution for interpretation. However, results showed that 16 P2 plots that match one attribute were found by expanding the search area to a cluster of seven hexagons. For some individual forest strata, this spatial resolution can therefore be achieved.

However, results from this study also imply that analyses of more detailed forest strata would involve searching larger areas to find 16 P2 plots for every matching P3 plot. For example, searches for P2 and P3 plots that sample pole-sized, oak-hickory forests on well-drained sites would obviously result in a smaller number of P3 plots and require going a greater distance to find 16 neighboring P2 plots (fig. 4). In other words, forest strata by several attributes will have lower spatial resolutions of interpretation than those strata defined by just one attribute. In some instances, a forest stratum of interest may be too detailed to obtain much spatial resolution within a State, and the default approach to regional analyses would be necessary.

This paper did not directly demonstrate how data from the P3 plots can be extended to P2 plots. However, one plausible method is to assign P3 values to neighboring P2 plots within the same strata. The distance at which P2 plots were to be considered neighbors would depend on the results stated in this paper. In addition, it would be of interest to examine the variability of P3 data attributes themselves.

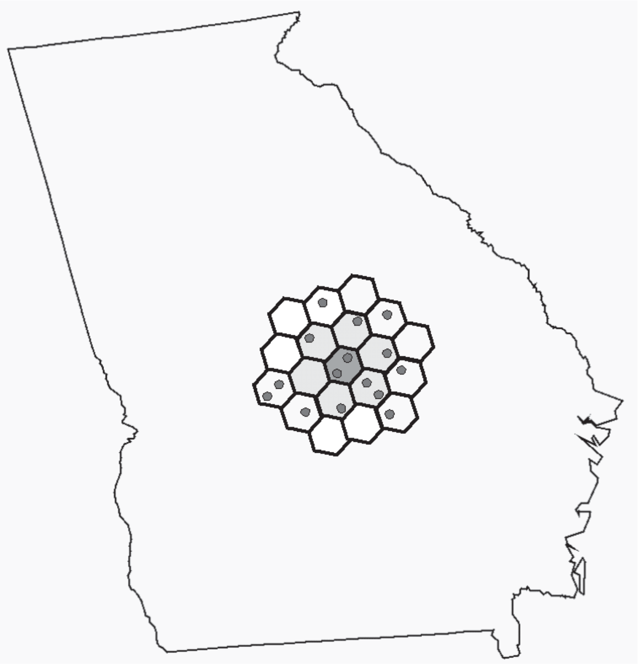


Figure 4—A cluster of 19 P3 hexagons to illustrate a more extensive search area for P2 plots with matching attributes.

## REFERENCES

**Brooks, R.T.; Dickson, D.R.; Burkman, W.B. [and others].** 1992. Forest health monitoring in New England: 1990 annual report. Resour. Bull. NE–125. Radnor, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 111 p.

**Roesch, F.A.; Reams, G.A.** 1999. Analytical alternatives for an annual inventory system. Journal of Forestry. 97(12): 33–37.

# PLOTVIEW SOFTWARE FOR RETRIEVING PLOT-LEVEL IMAGERY AND GIS DATA OVER THE WEB[1]

**Ken Boss**[2]

**Abstract**—-The Minnesota Department of Natural Resources Division of Forestry Resource Assessment office has been cooperating with both the Forest Service's FIA and Natural Resource Conservation Services's NRI inventory programs in researching methods to more tightly integrate the two programs. One aspect of these ongoing efforts has been to develop a prototype intranet application to provide interactive, map-based access to "raw" FIA and NRI plot data. The application runs on a password-protected web site at http://www.ra.dnr.state.mn.us/plotview. Products retrievable for any given plot on the interactive map include scanned aerial photographs, scanned field plot sheets, scanned USGS quad maps, digital ortho quad photography, and, in the case of the NRI program, database outputs of the raw data collected for the plot. All software driving the interface is freely distributable under open source licenses.

## INTRODUCTION

PlotView is a web-based intranet application designed to provide online access to centrally maintained stores of plot data from anywhere that an internet connection is available. The design of the system obviates any need for duplication of large image datasets, while simultaneously providing instantaneous access to those datasets from virtually any location. Field data gatherers, crew managers, quality control personnel and data analysts are all enabled to access the data they need when they want it, thereby avoiding the lengthy delays encountered in the past when trying to access plot data.

## MAP INTERFACE

One of the great strengths of the PlotView application is that it enables users to view plot locations in a landscape context via online maps. The maps are easily navigated through a variety of means, including place name searches, public land survey specifications, map coordinate input, or point-and-click pan and zooming on the map itself. The map view may be toggled to a satellite image view, allowing users to readily identify both cultural and natural features in proximity to the plot(s) of interest.

## PLOT DATA ACCESS

Data pertaining to a given plot can be accessed either by entering the plot ID into a text box in the interface, or by clicking on the plot centroid symbol on the map. The data returned for a given plot may include any or all of the following:

Scanned aerial photographs—these can include any number of inventory-specific photos taken over the plot through the years, and/or more generic photo resources such as NHAP, NAPP or state and local photographic holdings.

Scanned plotsheets—available only for the FIA program, included specifically for access to the hand-drawn maps indicating plot access routes.

Plot-specific database contents—available currently only for NRI plots. Taps into the "dataview" web application at Iowa State for database outputs specific to a given plot. Integration with FIA plot databases is also possible.

DRGs (scanned USGS quad maps)—presented in a "seamless" fashion, so that users can view any area without regard to the boundaries of the original paper maps. Presented with GIS overlays, including plot centroid and label.

DOQs (quad-based orthorectified photography from USGS) —as with the DRGs, above, DOQs are presented in a "seamless" mode, so that any area of interest may be viewed without regard to original photo boundaries. Also presented with GIS overlays, including plot centroid and label.

## SUPPORTING SOFTWARE

PlotView is supported by a number of freely-distributable open source software packages, listed below.

MapServer (http://mapserver.gis.umn.edu)—provides the interactive web-mapping capabilities.

Perl (http://www.perl.com)—the scripting "glue" that holds the various pieces together.

MySQL (http://www.mysql.com)—the database system that holds the metadata for the variety of data elements associated with the plots. Note that MySQL is not required for the PlotView application to work, and may be supplanted with nearly any web-accessible database system (Oracle, Sybase, Informix, etc.)

Apache Web Server (http://www.apache.org)—Web serving software. May be supplanted by a variety of other web serving software packages.

PDFlib (http://www.pdflib.com)—PDF file format generator. Enables the PDF encapsulation of images and maps for printing.

[1] Paper presented at the Second Annual Forest and Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Web Developer, Minnesota Department of Natural Resources, Division of Forestry—Resource Assessment, 413 SE 13th Street, Grand Rapids, MN 55744.

There are a couple of proprietary software products that, while not required to run PlotView, are indispensable in preparing GIS and image data for web distribution.

ArcView and/or ArcInfo (http://www.esri.com)—GIS software used to prepare the map datasets served with the MapServer. Note that the MapServer is presently poised to incorporate the OGR and GDAL open source software libraries, which will greatly expand the sources of both vector and raster GIS data that the MapServer can work with.

MrSID Encoder (http://www.lizardtech.com)—Software for the wavelet compression of image data; enables speedy delivery of large image datasets over the internet.

## FURTHER INFORMATION

PlotView is best understood by visiting the web site itself. Visit http://www.ra.dnr.state.mn.us. Note that you will need a user ID and password in order to access the site, though for security reasons these cannot be printed here. Contact your FIA program manager or NRI ICCS leader for this information.

Feel free to direct any technical or administrative questions concerning PlotView to the author of this paper.

# Index of Authors

**Reams, Gregory A.; McRoberts, Ronald E.; Van Deusen, Paul C., eds.** 2001. Proceedings of the second 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. 143 p.

Documents progress in developing techniques in remote sensing, statistics, information management, and analysis required for full implementation of the national Forest Inventory and Analysis program's annual forest inventory system.

**Keywords:** Annual forest inventory, estimation, imputation, inventory design, modeling, remote sensing.