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.

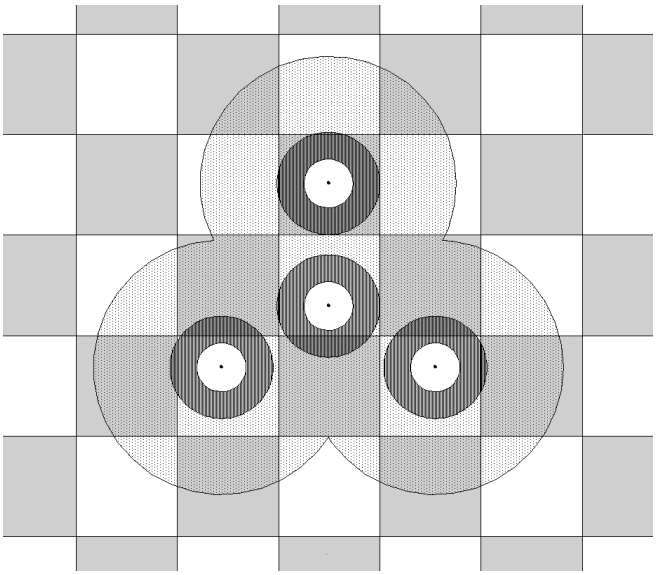

Figure 1—The FIA ground plot geometry versus 30m TM pixels. The plot consists of a cluster of four 0.017 ha subplots. The dark grey circles represent the area of locational error due to GPS errors. The larger grey circles represent the potential locational error due to image registration.

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

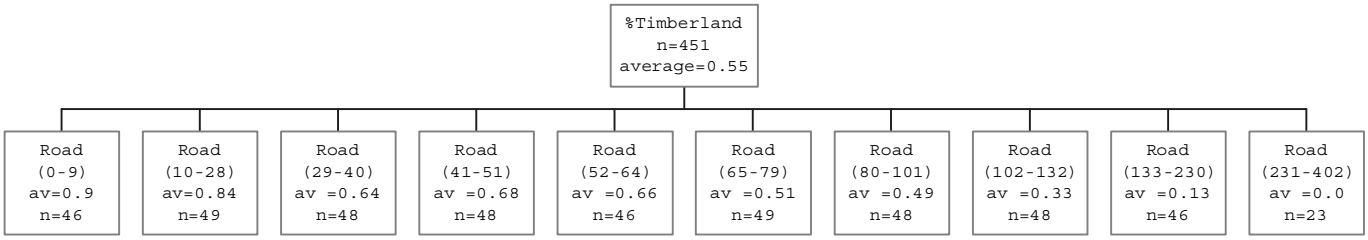

Figure 2—The decision tree algorithm first splits the continuous response variable into 10 classes of approximately equal size by consecutive groups of values of each of one of the predictor variables. Here the predictor variable, roads within 1 km, splits the response variable, percent timberland per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

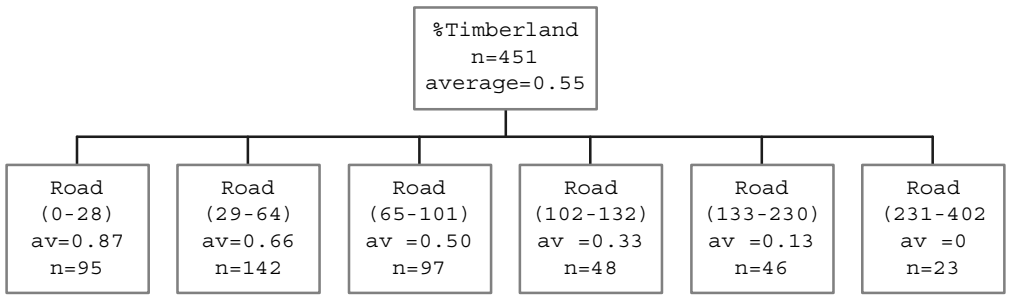

Figure 3—Groups of the response variable are merged when they are found to be statistically similar by an *F*-test. Here seven of the original 10 classes of the response variable, percent timberland per plot, have been merged into three, while only three of the original classes remain unchanged.

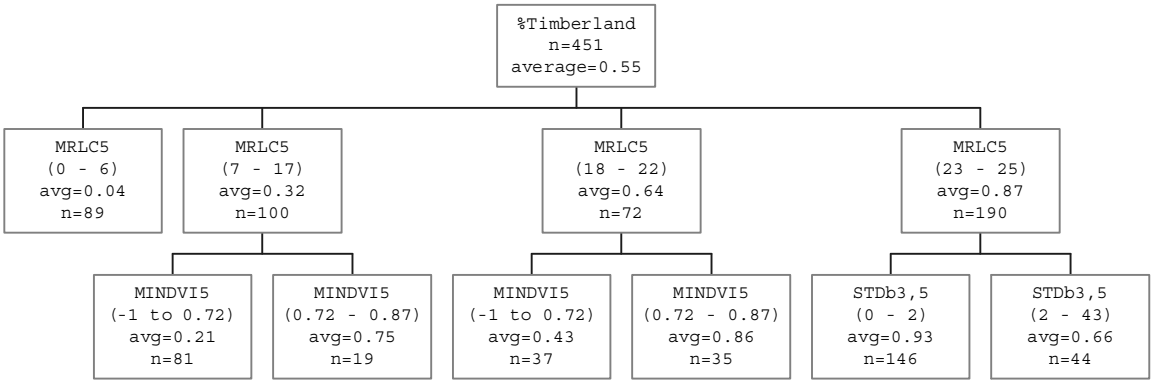

Figure 4—Final decision tree model for classes of percent timberland for all Connecticut ground plots. Each box shows the predictor variable, predictor variable class boundaries, average percent timberland, and the number of plots.

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.