A STATISTICALLY VALID METHOD FOR USING FIA PLOTS TO GUIDE SPECTRAL CLASS REJECTION IN PRODUCING STRATIFICATION MAPS

Michael L. Hoppus and Andrew J. Lister [1]

ABSRACT.—A Landsat TM classification method (iterative guided spectral class rejection) produced a forest cover map of southern West Virginia that provided the stratification layer for producing estimates of timberland area from Forest Service FIA ground plots using a stratified sampling technique. These same high quality and expensive FIA ground plots provided ground reference data for the classification method. Dividing the counties in the southern portion of the state into two groups, and using the FIA plots in each county group to make the strata for the other, avoids the potential bias incurred by having the plots stratify themselves. This procedure achieved the required precision of 3 percent sampling error per million acres of timberland.

BACKGROUND

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service is responsible for inventory and monitoring of the Nation's forests. 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, volume of timber, growth, and depletion of the Nation's forest land. In the East, statistical estimates derived from FIA inventories must meet specified sampling errors; for example, a 3-percent error per 1 million acres of timberland is the maximum allowable sampling error for area. Timberland is defined by FIA as forest land that is producing or capable of producing crops of 20 cubic feet of industrial wood per acre per year (Hansen and others 1992). Until now, FIA has reached this precision in part by stratifying the FIA ground plots using aerial photos and implementing a stratified sampling technique (Cochran 1977). 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 to 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 5 years.

The Northeastern FIA unit, responsible for surveying 13 Northeastern States, uses aerial photos from the National Aerial Photography Program (NAPP) for FIA ground plot stratification. NAPP currently is on a 7-year cycle. The high cost of additional qualified photointerpreters necessary to complete aerial photo stratification in all the states on a 5-year cycle plus the 7-year cycle of NAPP has led to investigations of the use of satellite imagery to create a stratification layer for stratifying the ground plots into homogeneous groups in order to reduce the variance of estimates (Cochran 1977). Furthermore, satellite image classification provides a forest cover map that can be used to evaluate forest distribution and change over space and time.

OBJECTIVES

The primary purpose of this project was to estimate the timberland area contained within the 14 southern counties of West Virginia. Additionally, we wanted to make and evaluate FIA's first practical application of satellite image-based stratification. Preliminary results of this analysis were requested by the State forester of West Virginia. An

[1] Research Forester and Forester, respectively, U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA 19073. Phone (610) 557-4039; e-mail: mhoppus@fs.fed.us

important goal of the image classification design was to use the FIA ground plots as reference data in a statistically valid way. The plots provide abundant high quality ground truth that is well distributed, recent, and accurately located with GPS. The risk, of course, is that use of the plots to make stratification maps may produce statistically invalid estimates from a post-stratified random sampling method, i.e., the plots might erroneously be used to derive the stratification layer that will then in turn be used to stratify them. We hypothesized that timberland area estimates produced from stratification of the plots based on satellite imagery would be more precise than estimates provided by simple random sampling of ground plots.

METHODS

Ground Plots

FIA ground plots were distributed randomly over the landscape in the 14 counties of the FIA's Southern unit of West Virginia with an intensity of approximately one plot per 6,000 acres or one plot about every 3 miles. Data were collected in 2000 and 2001. Many data were obtained on each plot in accordance with FIA protocol, including current land use class. Timberland is one land use class whose area can be estimated from the survey (USDA Forest Service 2001).

The layout and geometry of the ground plots have implications for use as image classification reference data. Two issues must be considered when comparing the pixel brightness values of Landsat TM images and their corresponding FIA plot data. First, the FIA plot consists of four 48-foot-diameter circular subplots: one in the center, and three symmetrically separated from the central subplot by 120 feet (fig. 1). This means that the plot places a subplot on a minimum of 4 pixels. Each subplot covers only 19 percent of the area of a pixel. Second, the location accuracy of the pixel is plus-or-minus 30 m on average and the GPS location accuracy of the plot center is about plus-or-minus 10 m. Therefore, there is not a clear one-to-one relationship between plot data and a single pixel. We have found from previous work that plots closer than about 90 m (3 pixel widths) to a different land cover type should not be used for reference data to classify pixels. For the same reason, only plots that have a single land use class over the four subplots should be used to help classify the satellite image.

The plot coordinates were converted to a vector GIS representation of the information. ERDAS Imagine (ERDAS, Atlanta, Georgia 30329) image processing software was used for image classification.

Satellite Image Classification

An innovative classification technique called Iterative Guided Spectral Class Rejection (IGSCR) was used to produce a forest/nonforest map of the area from the six non-thermal bands of Landsat TM. This technique was developed with the specific objective of producing FIA phase 1 inventory: FIA plot stratification to improve the precision of estimates (Wayman and others 2001). At its core, the method uses a large number of reference pixels, whose spectral signatures and forest or nonforest information class are known, to label the spectral classes created from an unsupervised classification. An unsupervised classification is a common technique used to group pixels that have similar spectral characteristics. Acquiring sufficient amounts of accurate forest reference data that can be used for creating the pool of reference pixels is often very difficult and expensive (Congalton and Biging 1992). This project made use of the large number of existing FIA ground plots as its source of high quality, accurately located reference data to be used as training sites for image classification.

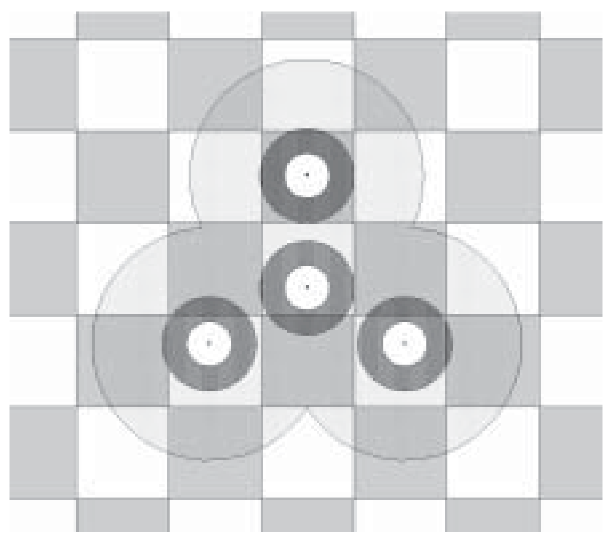

Figure 1.—*The FIA ground plot geometry versus 30-m TM pixels. The plot consists of a cluster of four 0.017-ha subplots. The dark gray circles represent the area of locational error due to GPS errors. The larger gray circles represent the potential locational error due to image registration.*

For details of the image classification algorithms and the ERDAS Imagine Software usage, see ERDAS (1997). Reference pixels for forested areas were identified by overlaying the location of homogeneous timberland plots onto the satellite image using ERDAS Imagine. By using a region growing tool in this software, a cluster of spectrally similar pixels was created for each timberland plot. Reference pixels for nonforest areas were created by interpreting digital ortho-maps and growing a region of spectrally similar pixels at the corresponding location on the satellite imagery. We did not use nonforest FIA plots for locating nonforest training sites because FIA plots do not characterize nonforest areas well from a spectral point of view. Areas covered with trees may have a nonforest definition, such as city parks. Typical nonforest areas, such as roads, buildings, quarries and pastures, are easy to identify on digital ortho-maps.

Four raw Landsat TM images that cover the area were mosaicked together. Documentation on how the software performs this task is weak, and there was some concern that differences in pixel brightness values between scenes at the overlap area would create spectral confusion. A careful examination of the finished forest/nonforest map showed no evidence of mapping inaccuracies along the seams of the mosaicked images. However, it is important to note that when maps were produced from the same mosaic, using more traditional supervised and unsupervised classification methods, definite inconsistencies occurred at the seams. The classification method we used did not appear to produce any noticeable inconsistencies at the seam lines of the scenes. The satellite scenes were all acquired during late May and early June 2000.

An unsupervised classification (ISODATA) method clustered the mosaicked image into 100 spectrally similar classes. The IGSCR algorithm accepts and labels a spectral class when 90 percent of the reference pixels contained within a given class have the same information class: forest or nonforest. At least 10 reference pixels have to be located within the class. All classes not labeled in the first iteration are “rejected” into a single group of pixels that are again subjected to an unsupervised classification. A portion of these second iteration classes is labeled, and a third iteration of the process is performed. The signature files for all the classes labeled as pure forest or nonforest are then merged into one signature file and used in a supervised classification using the maximum-likelihood decision rule. This resulted in a forest/ nonforest raster map that was used as a basis for all stratification methods.

STRATIFICATION METHOD

A simple method was used to keep the plots from being stratified by the same map that they produced. The 14 counties were divided into two sets. FIA plots from each of the two county sets were used to create reference pixels that were used to produce two independent forest/nonforest maps using the IGSCR technique. The forest/nonforest maps of county set A were produced with the help of plots located only in county set B and vice versa. The result was that all FIA plots located in each of the counties were labeled, and their estimates weighted, by strata created from a map produced by plots located outside the county.

Stratified random sampling procedures were applied to the FIA plots located within the 14 counties of FIA unit 3 of West Virginia to estimate timberland area. For details of the stratified random sampling protocol, see Cochran (1977). Three stratification designs were compared in this project:

1. Two strata—Forest and Nonforest;
2. Four strata—Forest, Nonforest, Forest within 2 pixels of Nonforest, and Nonforest within 2 pixels of Forest;
3. Four strata (based on the number of forested pixels within a 5 X 5 pixel window)—0-6 forested pixels, 7-17 forested pixels, 18-22 forested pixels, and 23-25 forested pixels.

A previous study found that nearly all of the significant Landsat TM predictor variables for timberland area on an FIA plot were based on images created by 3 X 3 or 5 X 5 filters. These filters were designed to calculate the average or standard deviation of pixel brightness values (raw bands and transformed layers) within the window or, when applied to classified images, they summed the count of forested pixels within the window (Hoppus and others 2001). This indicates that the geometry match between the plots and the TM pixels requires a measurement of each pixel’s neighborhood for best results when using Landsat TM derived maps to stratify FIA ground plots. The numbers of forested pixels in each of the four strata based on the 5 X 5 pixel neighborhood were chosen based on the results of this study (fig. 2).

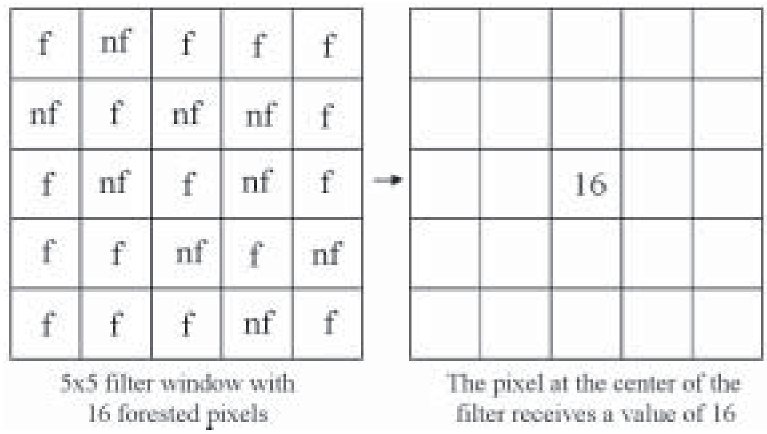

Figure 2.—*Filtering the forest/nonforest map into strata. A 5 X 5 pixel filter window is passed over the forest/nonforest map. All forested pixels within the window are counted and the sum placed in the center of the window. This results in a new raster map of forest pixel counts. Pixels with values equal to forested neighborhood counts are grouped into strata.*

Placement of an FIA plot into a given stratum was based on the value of the pixel coinciding with the center of the central subplot. It was predicted that the four strata designs would reduce the variance of the estimate of timberland area more than the two strata forest/nonforest design because they would be more likely to group plots based on their proximity to forest edges and/or areas of differing heterogeneity.

Stratified Estimation Formulae Using Satellite Derived Strata

The data collected were calculated using the FINSYS computer system developed at the Northeastern Research Station. The inventory estimates are based on stratified random sampling methodology.

Assumptions for the Image Strata:

1. Each image pixel in the entire strata is assigned to one of the classes, e.g., forest or nonforest.
2. Each plot is also assigned to one of the same set of classes.

Notation:

Each class is called a "strata" and for strata "h" (Cochran 1977)

N = total number of pixels

N_h = number of pixels in class h

n_h = number of plots in class h

$W_h = N_h / N$ proportion of total pixels in stratum "h"

y_{hi} = proportion of plot "i" in stratum "h" that is timberland

$\bar{y}_{hi}$ = mean proportion of timberland for stratum "h", which

$= \sum_{i=1}^{n_h} y_{hi} / n_h$ = estimate of the proportion of land area that is timberland for stratum "h"

L = total number of strata

1. To obtain an estimate of the proportion of timberland for the entire area, calculate a weighted mean of the individual stratum means. Weights should be proportional to the stratum size because it is improper to weight the mean representing a very small area the same as one representing a very large area. The overall stratified estimate is

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

2. The estimated variance of $\bar{y}_h$ is obtained with the sums of squares about the mean

$$s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$$

and the variance of the mean is

$$\text{var}(\bar{y}_h) = s_h^2 / n_h$$

3. The estimated variance of $\bar{y}_{st}$, with a correction factor for a "finite population," is

$$\text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \; \text{var}(\bar{y}_h)(1 - f_h) \text{, where } f_h = n_h / N_h$$

4. To calculate an estimate of total timberland area, multiply the estimate of the mean proportion by the known total land area "A"

$$\hat{y}_{st} = \text{A} \times (\bar{y}_{st})$$

The variance of $\hat{y}_{st}$ is

$$\text{var}(\hat{y}_{st}) = A^2 \; \text{var}(\bar{y}_{st})$$

The standard error is the square root of $A^2 \operatorname{var}(\bar{y}_{st})$

$$SE(\hat{y}_{st}) = A \times \operatorname{var}(\bar{y}_{st})^{1/2}$$

5. The sampling error is equal to the standard error divided by the estimate

$$SE = A \times SE(\bar{y}_{st}) / \hat{y}_{st}$$

RESULTS

The estimates, sampling errors, and relative sampling efficiencies for total timberland in FIA unit 3 are presented in table 1. Simple random sampling of the plots, without stratification, provided an estimate of 4,005,380 acres of timberland with a sampling error of 1.6 percent. The required precision for this area is 1.5 percent in order not to exceed an FIA limit of 3 percent sampling error per million acres of timberland. All three of the stratification designs, based on forest/nonforest maps produced by applying the IGSCR technique to four mosaicked Landsat TM scenes, provided estimates with precisions less than 1.5 percent. It is clear that the stratification method that grouped and weighted the plots based on the number of forested pixels in the surrounding 25-pixel neighborhood reduced the variance of the estimate the most (table 1).

Another way to compare the efficiency of a stratification method is to calculate how many extra plots would be required to reduce the variance to the same level. The percent increase in precision is equal to

$$1 - (\text{original number of plots / original number plus extra plots})^{0.5}.$$

A total of 1,722 plots are required to reduce the variance from 1.6 percent to 1.1 percent, the sampling error for the stratified sampling estimate provided by the 5 X 5 forested pixel sum. That's an increase of 908 plots at an additional cost of $700,000.

DISCUSSION

Our results were consistent with those of McRoberts and others (in press), where forest land estimates of portions of four states were produced with more precision using a satellite-based stratification method with four strata with a 2-pixel edge. We had expected that stratified random sampling would lead to an increase in the precision of estimates of timberland area over those from simple random sampling. Our main concern, however, was that the satellite stratification technique would not reduce the variance below the acceptable limit. Furthermore, we wanted to devise a method that did not use the FIA plots to make a map that would be used to stratify the same plots. Our method of splitting the counties into two groups and using the opposite counties' plots for stratification does not violate the assumptions of independence of strata and the corresponding plots.

Table 1.—*Estimates, sampling errors, and efficiencies for total timberland in FIA unit 3 of West Virginia provided by sampling FIA plots stratified three different ways using a forest/nonforest map. Four mosaicked Landsat TM scenes were classified using the IGSCR technique to produce the map*

Landsat TM map strata	Timberland	Sampling error	Efficiency
	Acres	*Percent*[a]	*Percent*[b]
Simple Random Sampling	4,005,380	1.60 (3.20)	n/a
IGSCR (F/NF Plot-Pixel)	4,038,696	1.23 (2.47)	69
IGSCR (F/NF/2-Pixel Edge)	4,038,155	1.16 (2.33)	90
IGSCR (5X5 Pixel F Sum)	4,056,300	1.10 (2.22)	112

[a] Sampling error for entire study area (per million acres of timberland).

[b] Sample sizes would have to be increased by this percent to achieve the same precision without stratification.

Our results indicate that even without stratification, we would almost have met our requirements. It should be noted, however, that this region of West Virginia is approximately 80 percent forested. When sampling for proportions (forest and nonforest), the variance is generally greater when the two classes are nearly equal. In this study, the classes are not equal, and thus the variance is expected to be lower. Our results also indicate that stratification using the classification derived from the iterative guided spectral class rejection method detailed in Wayman and others (2001) can reduce variance below that from simple random sampling. We thus conclude that there is a clear benefit to using classified satellite imagery, not only from a variance reduction standpoint, but also for its possible use in other applications such as fragmentation analysis, change detection, and graphical display of forest land distribution.

LITERATURE CITED

Cochran, W.G. 1977. Sampling techniques. 3d ed. New York, NY: John Wiley and Sons. 428 p.

Congalton, R.G.; Biging, G. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. Photogrammetric Engineering and Remote Sensing. 58(12): 1669-1671.

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

Gillespie, A. 1999. Rationale 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 database: users manual. Gen. Tech. Rep. NC-151. St Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 48 p.

Hoppus, M.; Arner, S.; Lister, A.J. 2000. Stratifying FIA ground plots using a 3-year old MRLC cover map and current TM derived variables selected by “decision tree” classification. (on file at the Northeastern Research Station, USDA Forest Service, Newtown Square, PA 19073). In: Proceedings of the 2000 FIA Symposium; 2000 October 17-18; Salt Lake City, UT: 19-24.

McRoberts, R.E.; Wendt, D.G.; Nelson, M.D.; Hansen, M.H. In press. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Submitted to Remote Sensing of Environment. September 2001. (On file at the North Central Research Station, USDA Forest Service, St. Paul, MN.)

USDA Forest Service. 2001. Forest Inventory and Analysis Northeast Field Guide. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Forest Inventory and Analysis Unit. 298 p.

Wayman, J.P.; Wynne, R.H.; Scrivani, J.A.; Reams, G.A. 2001. Landsat TM-based forest area estimation using iterative guided spectral class rejection. Photo-grammetric Engineering and Remote Sensing. 67(10): 1155-1166.