A Model-Based Approach to Inventory Stratification

Ronald E. McRoberts[1]

Abstract.—Forest inventory programs report estimates of forest variables for areas of interest ranging in size from municipalities to counties to States and Provinces. Classified satellite imagery has been shown to be an effective source of ancillary data that, when used with stratified estimation techniques, contributes to increased precision with little corresponding increase in costs. A new approach to stratification based on using satellite imagery and a logistic regression model to predict proportion forest area is proposed. The results suggest that precision may be substantially increased for estimates of proportion forest area and volume per unit area.

Introduction

The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture (USDA) Forest Service reports estimates of forest variables for medium to large geographic areas of interest (AOI) such as counties, national forests, and States based on data collected from arrays of field plots. Due to budgetary constraints and natural variability among plots, sufficient numbers of plots frequently cannot be measured to satisfy precision guidelines for the estimates of many variables unless the estimation process is enhanced using ancillary data. Classified satellite imagery has been accepted as a source of ancillary data that can be used with stratified estimation techniques to increase the precision of estimates with little corresponding increase in costs (Hansen and Wendt 2000, McRoberts *et al*. 2002, Hoppus and Lister 2003). The objective of the study was to evaluate the utility of satellite image-based stratifications derived from logistic regression predictions of proportion forest area (P) for increasing the precision of estimates of volume per unit area (V) and P.

Data

The FIA program has established field plot locations using a sampling design that is assumed to produce a random, equal probability sample (McRoberts and Hansen 1999). The sampling design is based on a tessellation of the United States into approximate 2,400-ha hexagons derived using the Ecological Mapping and Assessment Program methodology (White *et al*. 1992). At least one permanent plot has been established in each hexagon. The hexagonal array has been divided into five nonoverlapping, interpenetrating panels, and measurement of plots in one panel is completed before measurement of plots in the next panel is initiated. Panels are targeted for selection on a 5-, 7-, or 10-year rotating basis, depending on the region of the country.

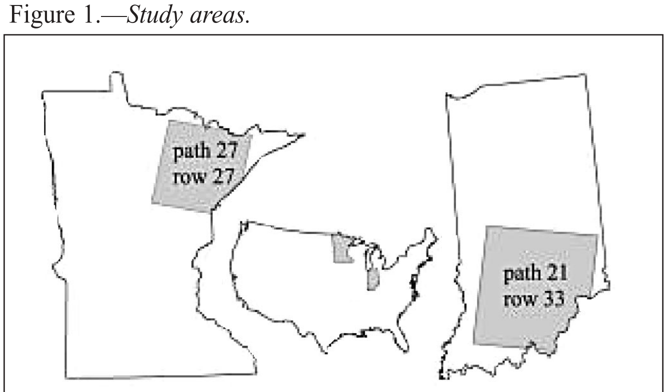
In general, locations of forested or previously forested plots are determined using Global Positioning System receivers, while locations of nonforested plots are determined using aerial imagery and digitization methods. Each field plot consists of four 7.31-m (24-ft) radius circular subplots. The subplots are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0°, 120°, and 240° from the center of the central subplot. Among the observations field crews obtain are the proportions of subplot areas that satisfy specific ground land use conditions. Subplot estimates of P are obtained by collapsing ground land use conditions into forest and nonforest classes consistent with the FIA definition of forest land. Field crews also measure the diameter at breast height (d.b.h.), 1.37 m (4.5 ft) and the height of each tree with d.b.h. ≥ 12.5 cm (5 in). Statistical models are used to predict the volume of each tree from the d.b.h. and height measurements, and volumes of all trees with d.b.h. ≥ 12.5 cm on each subplot are added to obtain subplot estimates of V. The national FIA program uses an infinite sampling framework and attributes aggregations of data for the four subplots to the point

[1] Mathematical Statistician, U.S. Department of Agriculture, Forest Service, 1992 Folwell Avenue, St. Paul, MN 55108. Phone: 651–649–5174; e-mail: rmcroberts@fs.fed.us.

corresponding to the center of the central subplot. For two study areas (fig. 1), one in southern Indiana and one in northern Minnesota, FIA plot data and three dates of Landsat Thematic Mapper (TM) or Enhanced TM+ imagery were used. Observations of P and V obtained between 1999 and 2003 were available for 1,211 FIA plots in the Indiana study area and for 2,114 FIA plots in the Minnesota study area.

Landsat imagery for one Indiana scene (path 21, row 33) and one Minnesota scene (path 27, row 27) was obtained from the Multi-Resolution Land Characterization 2001 land cover mapping project (Homer *et al*. 2004) of the U.S. Geological Survey. Imagery for three dates corresponding to early, peak, and late vegetation green-up (Yang *et al*. 2001) were obtained for each scene: April 2001, July 2000, and October 2001 for the Indiana scene and March 2000, July 1999, and October 1999 for the Minnesota scene. Preliminary analyses indicated that Normalized Difference Vegetation Index (NDVI) (Rouse *et al*. 1973) and the tasseled cap (TC) transformations (brightness, greenness, and wetness) (Kauth and Thomas 1976, Crist and Cicone 1984) were superior to both the spectral band data and principal component transformations in predicting P. Thus, 12 satellite image-based predictor variables were used: NDVI and the three TC transformations for each of the three image dates. Because plots would eventually be assigned to strata derived from pixel classifications or predictions, the constraint that a plot could not sample multiple strata had to be accommodated, and the FIA plot configuration requires a 3 x 3 block of pixels for geospatial coverage, the mean of each transformation of the spectral values was calculated for each 3 x 3 block of pixels and attributed to the center pixel of each block. Similarly, the FIA plot observations of P and V were attributed to the pixel containing the plot center.

Figure 1.—*Study areas.*



Methods

Stratified Estimation

Stratified estimation requires accomplishment of two tasks: (1) calculation of the relative proportion of the land area corresponding to each stratum, and (2) assignment of each plot to a single stratum. After the classifications or predictions for the satellite imagery have been obtained and aggregated into useful strata, the two required tasks are relatively easy to accomplish. The first task is accomplished by counting the number of pixels in each stratum and then calculating the relative proportions of pixels in strata. The second task is accomplished by assigning plots to strata on the basis of the stratum assignments of their associated pixels.

Stratified estimates for FIA variables are calculated using standard methods (Cochran 1977):

$$\overline{Y}_{Str} = \sum_{h=1}^{H} w_h \overline{Y}_h, \tag{1}$$

and

$$Var\left(\overline{Y}_{Str}\right) = \sum_{h=1}^{H} w_j^2 \frac{\hat{\sigma}_j^2}{n_j}, \tag{2}$$

where

$$\overline{Y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}, \tag{3}$$

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(Y_{hi} - \overline{Y}_h\right)^2, \tag{4}$$

and where Y_{hi} is the i^{th} observation in the h^{th} stratum of the variable of interest; h=1,…H denotes strata; w_h is the weight for the h^{th} stratum, calculated as the proportion of pixels in the AOI assigned to the stratum; n_h is the number of plots assigned to the h^{th} stratum; $\overline{Y}_h$ is the sample mean for the h^{th} stratum; and $\hat{\sigma}_h^2$ is the sample estimate for the stratum variance.

The FIA program uses stratified estimation but not stratified sampling. For estimation purposes, at least five plots per stratum are considered necessary to obtain reliable stratified estimates. If fewer than five plots are assigned to a stratum, the user has four options: (1) combine similar strata, (2) increase the size of the AOI so that the stratum includes a sufficient number of plots, (3) combine strata and increase the size of the AOI, or (4) do not use stratified estimation.

The effectiveness of a stratification is often evaluated using relative efficiency (RE) calculated as follows:

$$RE = \frac{\hat{Var}\left(\overline{Y}_{SRS}\right)}{\hat{Var}\left(\overline{Y}_{Str}\right)}, \quad (5)$$

where $\hat{Var}(.)$ is estimated variance, $\overline{Y}_{SRS}$ is the estimate of the mean obtained under the assumption of simple random sampling (SRS), and $\overline{Y}_{Str}$ is the estimate of the mean obtained using stratified estimation. RE > 1.0 indicates that the strata and stratified estimation have the desired effect of reducing variance and increasing precision, while $RE \approx 1.0$ indicates the strata are having little benefit.

Model Prediction

Predictions of P for individual pixels in each study area were obtained using a logistic regression model (LOG):

$$E(P) = \frac{1}{1+\exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_{12} X_{12}\right)} \quad (6)$$

where E(.) is statistical expectation, exp(.) is the exponential function, the βs are parameters to be estimated, and the Xs are the 12 transformations of the satellite image spectral values. Separate sets of parameter estimates were obtained for each study area.

Because the estimates of all parameters of model (6) are obtained from observations for all plots in the study area, the model prediction for each image pixel will also be based on the observations for all plots in the study area. Thus, because the same plots are assigned to strata as were used to calibrate the model from whose predictions the strata were derived, concern that the plots do not constitute random samples of strata may exist. Breidt and Opsomer (2002) showed that for samples smaller than those used for model calibration for this study, this concern may be dismissed.

Analyses

For each study area, the pixel predictions of P were grouped into 0.01-wide classes beginning with $\hat{P}$ = 0.01 and ending with $\hat{P}$ = 1.00. Plots were assigned to the resulting 101 classes on the basis of the class assignments of the pixels containing the plot centers. The optimal grouping of the 101 classes into four strata was determined, subject to the constraint that no stratum with fewer than five plots was permitted. Within each study area, three optimality criteria were considered: RE_P, RE_V, and $RE_P + RE_V$.

As a basis for comparison, means and standard errors for P and V were obtained for each study area under the SRS assumption. In addition, means, standard errors, RE_P, and RE_V were obtained for the approach to stratification used by the regional FIA program of the North Central Research Station (NC), USDA Forest Service. With the NC approach, four strata are derived from the 21 classes of the National Land Cover Data (NLCD) (Vogelmann *et al*. 2001, Homer *et al*. 2004). First, the NLCD classes are aggregated into forest and nonforest classes, and second, 2-pixel-wide forest edge and nonforest edge classes are constructed along the forest/nonforest boundary (Hansen and Wendt 2000, McRoberts *et al*. 2002). These four classes—Forest, Forest Edge, Nonforest Edge, and Nonforest—are then used as strata. For the Indiana study area, the four strata were derived from the 1992 version of the NLCD, while for the Minnesota study area, the four strata were derived from the 2001 version of the NLCD.

Results

Estimates of mean P and V were nearly indistinguishable for the different approaches to stratification (table 1). In all cases, $RE_P > RE_V$, which is consistent with previous findings and can be attributed to the closer relationship between P and a forest/nonforest classification than between V and the classification. For both study areas, the LOG approach was superior to the

Table 1.—*Comparisons of results of stratifications.*

Approach	Strata optimization criterion	Indiana study area Mean	SE	RE	Minnesota study area Mean	SE	RE
Proportion forest area							
SRS		0.3383	0.0127	1.00	0.7279	0.0091	1.00
NC		0.3545	0.0067	3.60	0.7267	0.0073	1.53
LOG	RE_A	0.3373	0.0052	5.87	0.7298	0.0060	2.33
LOG	RE_V	0.3349	0.0060	4.50	0.7241	0.0064	1.74
LOG	$RE_A + RE_V$	0.3393	0.0053	5.72	0.7298	0.0060	2.33
Volume per unit area (m^3/ha)							
SRS		47.44	2.15	1.00	48.91	1.25	1.00
NC		49.75	1.53	1.99	48.78	1.17	1.13
LOG	RE_A	47.32	1.37	2.47	49.12	1.09	1.32
LOG	RE_V	46.97	1.31	2.71	48.62	1.06	1.37
LOG	$RE_A + RE_V$	47.57	1.35	2.54	49.12	1.09	1.32

NC approach with respect to increasing the precision of estimate of mean P and V. The largest RE_P for the LOG approach was 63 percent greater than RE_P for the NC approach for the Indiana study area and 51 percent greater for the Minnesota study area. The largest RE_V for the LOG approach was 36 percent greater than RE_V for the NC approach for the Indiana study area and 21 percent greater for the Minnesota study area.

As expected, for each study area the largest RE_P was obtained for the LOG approach when the strata boundaries were selected to maximize RE_P, and the largest RE_V was obtained when the strata boundaries were selected to maximize RE_V. The decrease in RE_P when the strata boundaries were selected to maximize RE_V was approximately 25 percent for both study areas, although the decrease in RE_V when the strata boundaries were selected to maximize RE_P was less than 10 percent for both study areas. Thus, RE_P is apparently more sensitive to changes in strata boundaries than is RE_V. When the strata boundaries were selected to maximize $RE_P + RE_V$, the relative decreases in both RE_P and RE_V for both study areas were less than 5 percent. Finally, very little advantage was realized for either study area when selecting strata boundaries to maximize $RE_P + RE_V$ as compared to selecting them to maximize RE_P.

Conclusions

The LOG approach was superior to the NC approach for estimating both mean P and V for both study areas. Increases in RE_P were 63 and 52 percent for Indiana and Minnesota, respectively, and increases in RE_V were 36 and 21 percent for Indiana and Minnesota, respectively. Although selection of strata boundaries to maximize $RE_P + RE_V$ rather than to maximize either RE_P or RE_V individually had a slight advantage, selection of boundaries to maximize RE_P was nearly as effective.

Greater REs could possibly have been achieved with a few more strata for these data sets. However, the bimodal distributions of the plots (figs. 2a and 2b) with respect to $\hat{P}$, with most plots either completely forested or completely non-forested, suggest that the minimum of five plots per stratum would be difficult to achieve with larger numbers of strata, particularly for the smaller geographic areas for which the FIA program reports estimates.

Figure 2a.—*Frequency distribution of Indiana study area plots.*

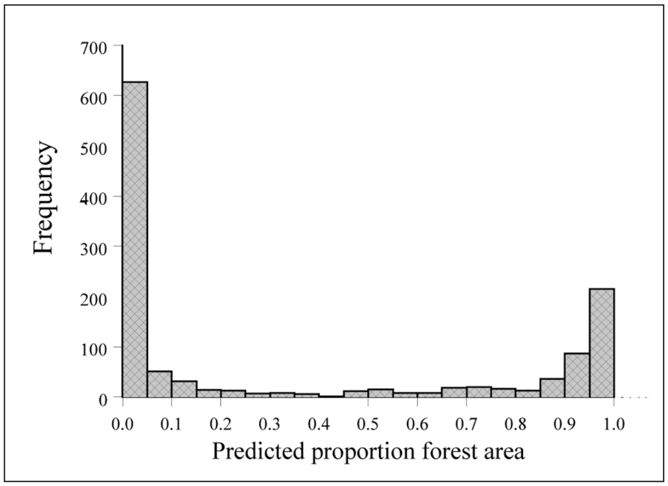
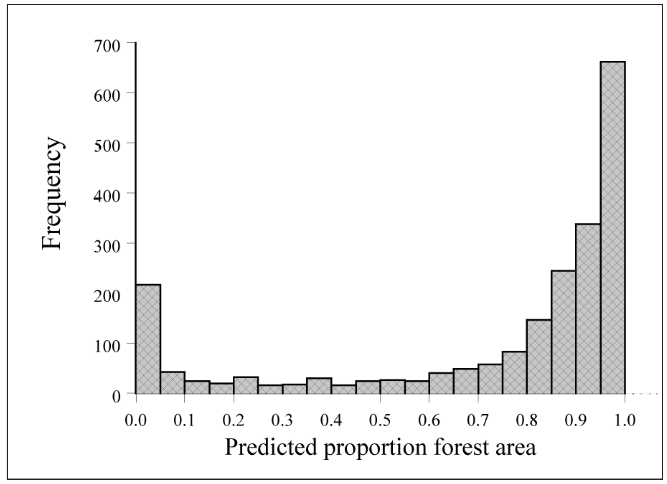


Figure 2b.—*Frequency distribution of Minnesota study area plots.*



Literature Cited

Breidt, F.J.; Opsomer, J.D. 2002. Endogenous post-stratification in surveys: classifying with a sample fitted model. Preprint Series #02-23, Department of Statistics, Iowa State University. www.stat.colostate.edu/~jaybreidt/lectures/epse.pdf.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. New York: John Wiley & Sons.

Crist, E.P.; Cicone, R.C. 1984. Application of the tasseled cap concept to simulated Thematic Mapper data. Photogrammetric Engineering & Remote Sensing. 50: 343–352.

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, 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.

Homer, C.; Huang, C.; Yang, L; Wylie, B.; Coan, M. 2004. Development of a 2001 national landcover database for the United States. Photogrammetric Engineering & Remote Sensing. 70(7): 829–840.

Hoppus, M.L.; Lister, A.J. 2003. A statistically valid method for using FIA plots to guide spectral class rejection in producing stratification maps. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; Moser, J.W., eds. Proceedings, 3rd annual forest inventory and analysis symposium; 2001 October 17–19; Traverse City, MI. Gen. Tech. Rep. NC-230. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 44–49.

Kauth, R.J.; Thomas, G.S. 1976. The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. In: Proceedings, the symposium on machine processing of remotely sensed data; 1976 June 29–July 1; West Lafayette, IN: Purdue University: 41–51.

McRoberts, R.E; Hansen, M.H. 1999. Annual forest inventories for the North Central Region of the United States. Journal of Agricultural, Biological, and Environmental Statistics. 4: 361–371.

McRoberts, R.E.; Wendt, D.G.; Nelson, M.D.; Hansen, M.H. 2002. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sensing of Environment. 81: 36–44.

Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. 1973. Monitoring the vernal advancement and retrogradation (green wave effect) of natural resources. Prog. Rep. RSC 1978-1. College Station, TX: Remote Sensing Center, Texas A&M University.

Vogelmann, J.E.; Howard, S.M.; Yang, L.; *et al*. 2001. Completion of the 1990s National Land Cover Data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering & Remote Sensing. 67: 650–662.

White, D.; Kimerling, J.; Overton, S.W. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. Cartographic and Geographic Information Sciences. 19: 5–22.

Yang, L.; Homer, C.G.; Hegge, K.; *et al*. 2001. A Landsat 7 scene selection strategy for a National Land Cover Database. In: Proceedings, IEEE 2001 international geoscience and remote sensing symposium; 2001 July 9–13; Sydney, Australia [CD-ROM]: 1123–1125.