Variable Selection Strategies for Small-area Estimation Using FIA Plots and Remotely Sensed Data

Andrew Lister, Rachel Riemann, Jim Westfall, and Mike Hoppus[1]

Abstract.—The USDA Forest Service's Forest Inventory and Analysis (FIA) unit maintains a network of tens of thousands of georeferenced forest inventory plots distributed across the United States. Data collected on these plots include direct measurements of tree diameter and height and other variables. We present a technique by which FIA plot data and coregistered remotely sensed raster data were used to predict the basal area of deciduous trees at a spatial resolution of 30 m. Results varied, generally indicating that culling putatively unrelated variables did not improve estimates over those obtained using all the potential variables in the model.

The USDA Forest Service's Northeastern Forest Inventory and Analysis unit (NE-FIA) is charged with conducting a portion of a national forest inventory. NE-FIA uses data collected on a network of ground plots to produce reports on the status of the region's forests.

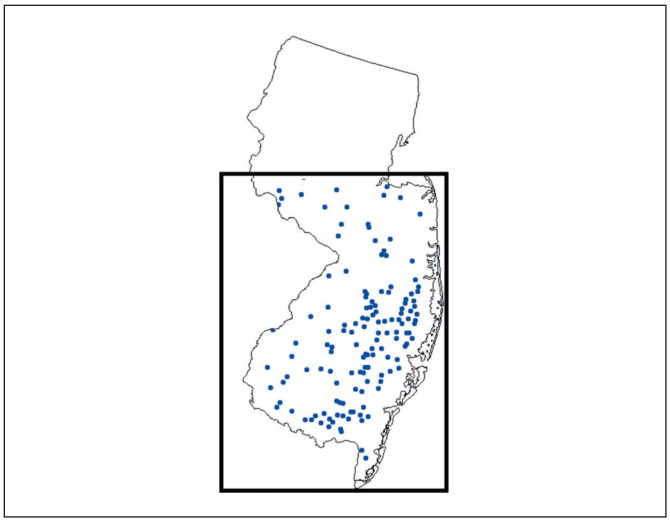
In addition to tabular reports, analysts and data consumers frequently request spatially explicit, highly resolute maps of forest variables. To produce these maps, data from geographic information systems (GIS) and satellites are often used to build models that predict attributes such as volume, biomass, and basal area.

There are many choices of GIS data and satellite layers for a given region. The National Land Cover Dataset (NLCD) project, a USGS-led, collaborative effort among several governmental and nongovernmental groups, is producing national land cover maps using GIS and satellite data. The USGS Eros Data Center compiled 18 GIS and satellite imagery layers for a mapping area covering several Mid-Atlantic States (NLCD mapping zone 60).[2] These data layers are coregistered, so they can be easily combined with NE-FIA plot data to produce a data set that can be used for predictive modeling. The goal of the current study was to assess the effects of subsetting these 18 layers to arrive at a model training set that would lead to more accurate predictions of the basal area of deciduous trees.

Methods

The study area included that portion of New Jersey covered by the NLCD imagery data (fig. 1). Data were collected on NE-FIA plots in New Jersey between 1998 and 1999.[3] The total amount of deciduous tree basal area measured on each plot was used as the dependent variable in the predictive modeling. Only

Figure 1.—*The study area in central and southern New Jersey; 141 homogeneous, forested plots were used for the analysis.*



[1] Forester, Research Forester, Research Forester, and Supervisory Forester, respectively. U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA 19073. Phone: 610–557–4038; e-mail: alister@fs.fed.us.

[2] Homer, C.; Gallant, A. 2001. Partitioning the conterminous United States in mapping zones for Landsat TM land cover mapping. USGS Draft White Paper, on file at USGS Eros Data Center, 47914 252nd Street, Sioux Falls, SD 57198.

[3] 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. USDA Forest Service, internal report. On file at USDA Forest Service, Washington Office, Forest Inventory and Analysis, Washington, DC.

completely forested plots were used in the analysis.

The portion of the NLCD data in New Jersey was used as a source of potential predictor variables. The NLCD data set was assembled by mosaicking, georeferencing, and radiometrically correcting three-season satellite imagery collected by the Landsat 7 satellite between 1999 and 2001 (USGS Eros Data Center 2002). These assembled raw images were transformed using the Tassled Cap (TC) transformation, a procedure that produces new images consisting of three layers per original seasonal six-band image (USGS Eros Data Center 2002). The TC transformation typically is used because the composite layers have a higher correlation with some features of vegetation than do the constituent layers. In addition to these nine TC layers, elevation, slope percentage, aspect, and slope position index were derived from digital elevation models (DEMS), which are raster GIS layers with a value for elevation at each pixel location. Slope, aspect, and slope position (ranging from 0 in the valley to 100 on the ridgetop) also were calculated for each pixel using a GIS. Soil quality, available water content (awc), and soil carbon percentage (variables that often are considered when measuring site quality) were derived from the STATSGO soils data set produced by NRCS (USDA Soil Conservation Service 1993). Layers consisting of geographic Easting and Northing also were created. All NLCD data layers were coregistered, standardized to be within the range of 0-255, and resampled to a 30-m pixel size.

Values of predictor variables at plot locations were obtained with Erdas Imagine software. Scatterplots of deciduous basal area vs. each of the predictor variables were generated and correlation matrices were created with SAS software. To create a subset of predictors for modeling, variables that were not significantly correlated with basal area were excluded from modeling, as were plots that were significantly correlated but subjectively considered weakly related after assessing the scatterplots.

The two modeling data sets (the full set and the subset) were used to produce maps of deciduous basal area for each 30-m pixel defined by the predictor layers. The technique used was a minimum-distance supervised classification, which is in effect a *k*-nearest neighbor imputation with a *k* of 1 (McRoberts *et al.* 2002, Franco-Lopez *et al.* 2001). This procedure is based on the multidimensional Euclidean distance between pixels where basal area is unknown but the predictor variables have known values, and a pixel with known values for both basal area and predictor variables. The basal area of the plot whose associated pixel has the smallest multidimensional Euclidean distance from the unknown pixel is assigned to the pixel being evaluated. Each pixel is treated in this way until a continuous map of basal area is produced.

The modeling procedure is such that a given plot's value never influences the prediction at its own location; it always is a different plot whose value is assigned to the pixel on which a plot sits, making it possible to use the modeling data for validation. To assess the accuracy of the resulting maps, scatterplots of observed vs. predicted basal area were generated from the original data, and simple linear regression models describing the relationship between observed and predicted values were created. Histograms of absolute error were generated for both maps.

Table 1.—*Correlation coefficients (r values) and p values from correlation analyses of the relationship between deciduous basal area and several predictor variables (N=141)*

Predictor variable	r value	p value
Position index	– 0.09	0.31
Slope	0.29	<0.001
Aspect	0.26	<0.01
Elevation	0.25	<0.01
Easting	– 0.54	<0.0001
Northing	– 0.24	<0.01
Soil water content	0.44	<0.0001
Soil carbon	0.08	0.33
Soil quality	0.33	<0.0001
Summer brightness	0.60	<0.0001
Summer greenness	0.65	<0.0001
Summer wetness	– 0.10	0.24
Fall brightness	0.58	<0.0001
Fall greenness	0.59	<0.0001
Fall wetness	– 0.06	0.45
Spring brightness	0.30	<0.01
Spring greenness	– 0.54	<0.0001
Spring wetness	– 0.41	<0.0001

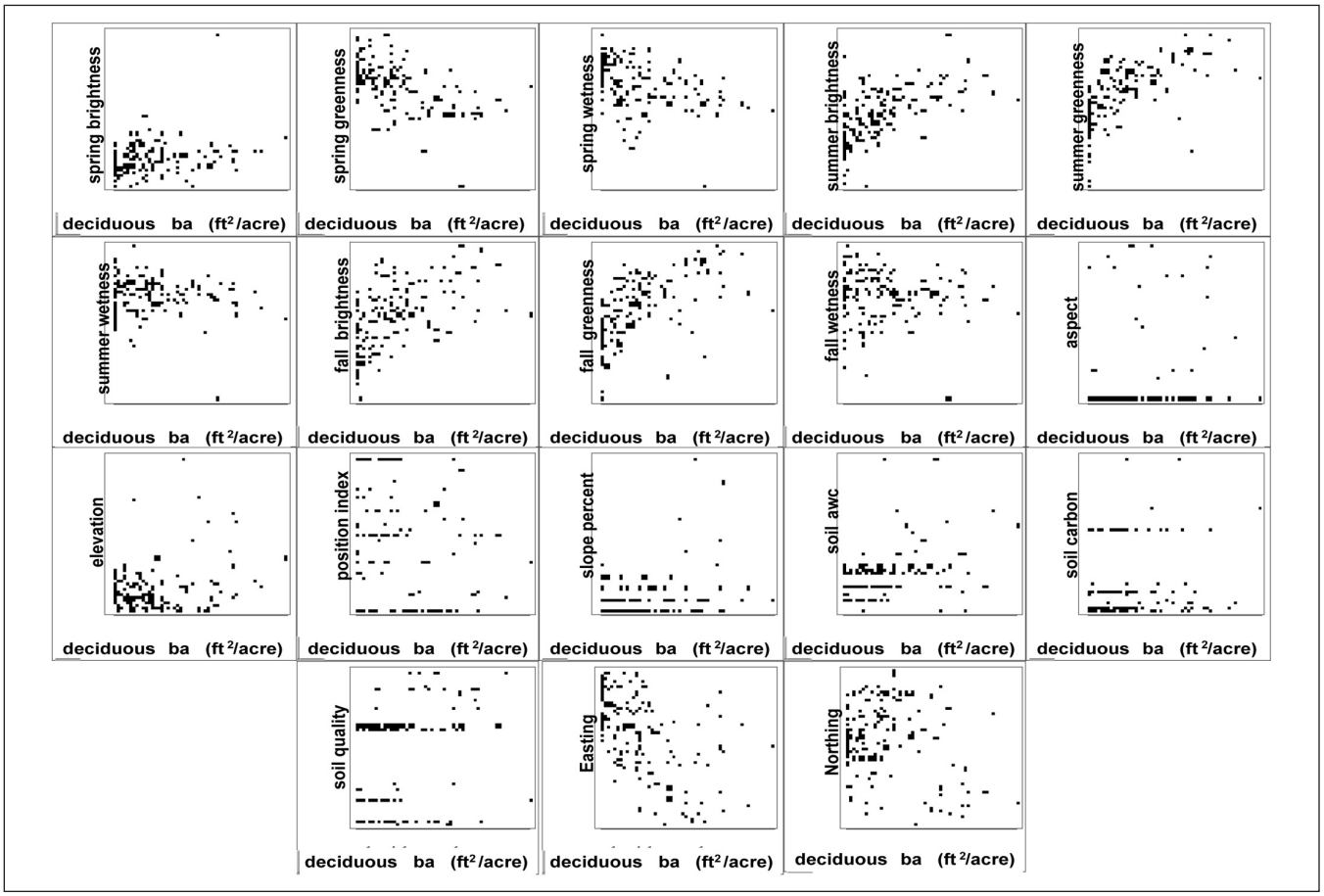
Figure 2.—*Final map of predictions of deciduous basal area for central and southern New Jersey. The map was produced using all 18 GIS and imagery layers. Lighter values indicate higher levels.*

Results and Discussion

The final map is shown in figure 2. The correlation statistics and resulting p values are shown in table 1. Fourteen of the 18 original predictor variables had significant correlation coefficients. Position index, soil carbon, and summer and fall wetness (two of the TC layers) were not significant ($p>0.05$). After subjectively assessing the scatterplot matrix (fig. 3), we decided to eliminate aspect, elevation, slope percentage, and Northing.

We had hypothesized that the DEM-based layers would be related to site quality and probability of development; low slope sites closer to a valley floor should have deeper, moister soil and be more prone to human development than sites on steep hillsides or ridgetops. There were no or only weak relationships between topographic site factors and basal area of

Figure 3.—*Scatterplot matrix showing relationship between the basal area of deciduous trees on FIA plots and several GIS and imagery-based values (see Methods for information on predictor variables). Y axis values ranged from 0 to 255; x axis values ranged from 0 to 192 ft²/acre (N=141).*



deciduous trees measured on NE-FIA plots in our study area, possibly because there are higher order interactions among topographic variables or between them and other unmeasured variables. Similarly, Xu and Prisley (2000) hypothesized that the local variation in carbon distribution could be due to factors such as previous land use, forest growth phase, forest type differences, geomorphology, natural disasters, or other unmeasured factors.

The lack of relationship between basal area and soil carbon might be caused by the same phenomenon or by the nature of the soil carbon data. Soil carbon may not be measurably biologically related to basal area production. The relationship between Northing and basal area did not appear to be linear. Perhaps the relationships could have been improved via transformation or by creating composite variables to test the effects of interactions, but we chose not to transform the data to preserve biological interpretability of our model outputs.

The TC wetness layer historically has been used to represent different levels of soil moisture. Perhaps during summer and fall, little bare soil was exposed on the FIA plots, making the wetness layer less useful during these seasons. However, before the deciduous trees produce leaves in spring, the satellite acquires reflected light from bare soil beneath the trees and thus might be measuring an ecological factor that affects deciduous basal area.

The scatterplots and diagnostic statistics of the regressions of observed vs. predicted for the full model and for the subset model are shown in figures 4a and 4b, respectively. Considering the shape of the scatterplot and the regression outputs, the full model performed better than the subset model. The R^2 value was higher, the slope was closer to 1, and the y intercept was closer to zero (figs. 4a - 4b). The histogram of absolute errors (fig. 5) indicates that the subset model performed slightly better in the second and third lowest basal area categories, but the subset model's errors generally had higher variances than those from the full model.

These results were unexpected. We had hypothesized that several of the potential predictor layers would be extraneous, that is, the effects of strong predictors on the estimate would be diluted. But we found that the full model performed better in a validation. The scatterplots indicate that the full model's accuracy was consistent throughout the distribution of observed val-

Figure 4.—*Observed vs. predicted scatterplots from validation of the full model (A) and the subset model (B). The full model used 18 GIS and imagery layers; the subset model used only 10 (N=141).*

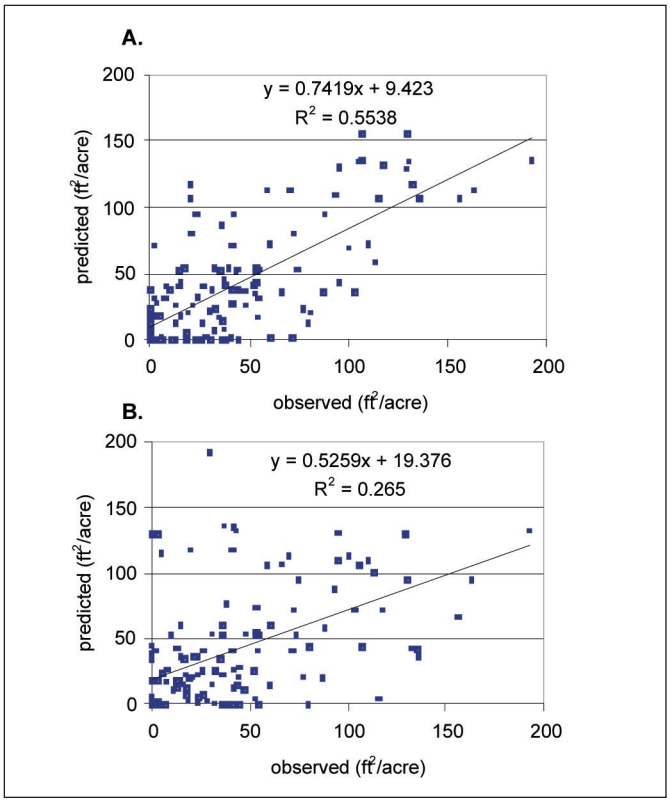
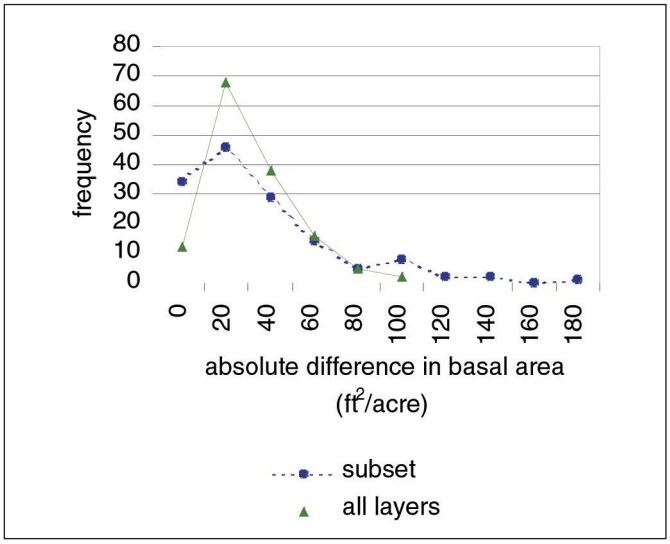


Figure 5.—*Absolute error histograms of the full model and the subset model. Absolute errors were calculated for each model by calculating the absolute difference between observed and expected (N=141).*



ues, whereas for the subset model, the accuracy was much worse in the lower tail of the observed data's distribution. A possible explanation is that, when extraneous data layers are used as predictors, the potential negative impact of anomalous model training data is mitigated. An unknown pixel that is incorrectly classified by a "bad" training site using the subset model might not be classified incorrectly if the distances are perturbed slightly by the addition of extraneous data layers. That extra distance raises the probability that a "better" training site might be assigned to that unknown location.

If this is the case, there must be a tradeoff between diluting the strength of mechanistic relationships between predictors and dependent data and susceptibility to poor training data. Our results suggest that the extraneous data layers served as a safety net, removing some of the effects of outlaying data points but not increasing the overall variance of our residual error to an unacceptable level.

In future studies, we plan to analyze the effects of individual training data sites on the accuracy of our modeling. Much of the variance in our absolute errors may be due to rogue training data. We also plan to test the effects of transforming the variables and creating composite layers consisting of interactions of the GIS and imagery layers. And we will investigate additional variable reduction methods, including multiple linear regression, principal components analysis, and other univariate and multivariate techniques. We also will assess the relationship between the amount of training data used and the number of spectral bands used. A phenomenon called the "Hughes Phenomenon" (Hughes 1968) occurs when accuracy is degraded when one increases the number of predictor variables but does not change the number of training sites.

Literature Cited

Franco-Lopez, H.; Ek, A.R.; Bauer, M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of the Environment. 77(3): 251–274.

Hughes, G.F. 1968. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. 14: 55–63.

McRoberts, R.E.; Nelson, M.D.; Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors. Remote Sensing of the Environment. 82: 457–468.

U.S. Department of Agriculture, Soil Conservation Service. 1993. State soil geographic data base (Statsgo) data user's guide. Misc. Pub. 1492. Washington, DC: U.S. Department of Agriculture, Natural Resources Conservation Service.

U.S. Geological Survey, Eros Data Center. 2002. MRLC 2000 image preprocessing procedure. http://landcover.usgs.gov/pdf/image_preprocessing.pdf. and on file at USGS Eros Data Center, 47914 252nd Street, Sioux Falls, SD 57198.

Xu, Y.J.; Prisley, S.P. 2000. Linking STATSGO and FIA data for spatial analyses of land carbon densities. In: Hubbard, W.G; Jordin, J.B., eds. Proceedings of the 3rd USDA Forest Service Southern GIS Conference; 2000 October 10–12; Athens, GA.