Regression and Geostatistical Techniques: Considerations and Observations from Experiences in NE-FIA

Rachel Riemann[1] and Andrew Lister[2]

Abstract.—Maps of forest variables improve our understanding of the forest resource by allowing us to view and analyze it spatially. The USDA Forest Service's Northeastern Forest Inventory and Analysis unit (NE-FIA) has used geostatistical techniques, particularly stochastic simulation, to produce maps and spatial data sets of FIA variables. That work underscores the importance of generating uncertainty information along with the modeled estimates, the value of incorporating additional satellite and other data into the modeling, and the need to understand the characteristics of the output data set. In our study, we investigated three questions: Does spatial structure matter when satellite-derived and ancillary spatial data sets are incorporated into the modeling of forest attributes? If we use a modeling technique such as multiple linear regression, how do we calculate or estimate the uncertainty? And what are the characteristics of the output data set with respect to the original sample data and the ancillary data used?

Background

Spatial depictions of forest variables improve our understanding of the forest resource by allowing us to view and analyze it spatially and ask questions such as: How are things distributed spatially, and how are they related to other social, environmental, and historical patterns? Estimates in small areas may be improved because additional, relevant information is incorporated into the modeling and estimation. Two extremely valuable data sources for mapping forest attributes are the Forest Inventory and Analysis (FIA) plot data and satellite-derived imagery. FIA data contain an enormous amount of information on a large number of sample plots collected in an unbiased manner and spread relatively uniformly across nearly the entire United States. Many ecosystems, small and large, and much of the variation within them are captured. In addition, satellite sensors capture data at every location (at various resolutions), and the resulting imagery often is strongly related to many of the forest variables we are interested in. For example, in a New Jersey study, we found that the correlation between basal area and several imagery-based satellite layers is high (table 1). In this paper we review the lessons learned from using the geostatistical technique sequential Gaussian conditional simulation (SGCS) to model the relative basal area of individual tree species, and investigate the use of multiple linear regression to model similar variables using satellite-based data sets and other available spatial layers.

Geostatistical Techniques

Much of the spatial modeling work in NE-FIA has been conducted with plot data and geostatistical techniques. For exam-

Table 1.—*Pearson's Correlation Coefficients between three tassel cap bands for each of three seasons (2000 data) and normal–score transformed relative basal-area values for each of three species variables being modeled (1996 data; all values are significant ($p < 0.05$), and bold values are more highly significant ($p<0.005$)*

Item	Pitch pine	Red maple	Total coniferous basal area
Spring_bright	0.26	-0.17	**-0.43**
Spring_green	**0.49**	**-0.23**	**0.61**
Spring_wet	**0.39**	—	**0.60**
Summer_bright	**-0.57**	**0.30**	**-0.67**
Summer_green	**-0.63**	**0.34**	**-0.58**
Fall_bright	**-0.56**	0.19	**-0.71**
Fall_green	**-0.54**	0.21	**-0.53**

[1] Research Forester/Geographer, U.S. Department of Agriculture, Forest Service, Northeastern Research Station, c/o USGS, 425 Jordan Road, Troy, NY 12180.
[2] Forester, U.S. Department of Agriculture, Forest Service, Northeastern Research Station, 11 Campus Blvd., Newtown Square, PA 19073.

Figure 1.—*Like kriging, sequential Gaussian conditional simulation (SGCS) uses a model of the spatial structure present to estimate values at unknown locations: a) original plots (darker red are increasing values of %ba of hemlock); b) correlogram calculated from the sample plot data; c) model used to describe the structure depicted in the correlogram; d) a resulting output map of modeled values (darker green are increasing values of %ba of hemlock).*

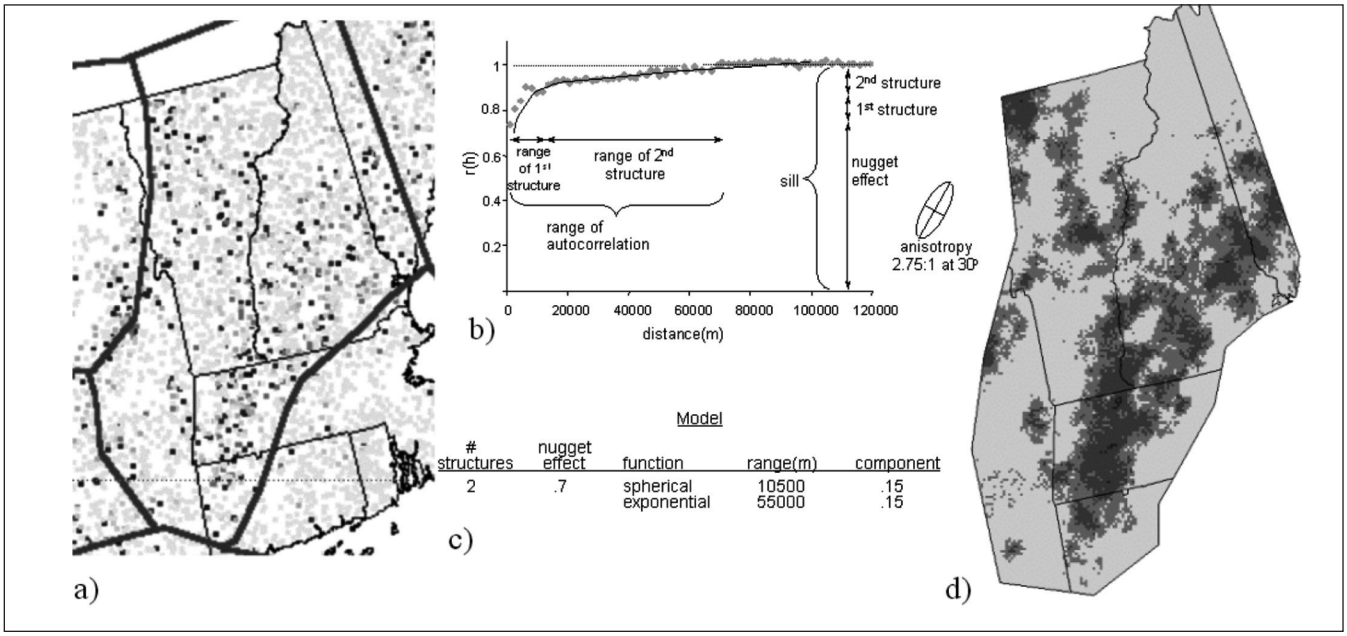
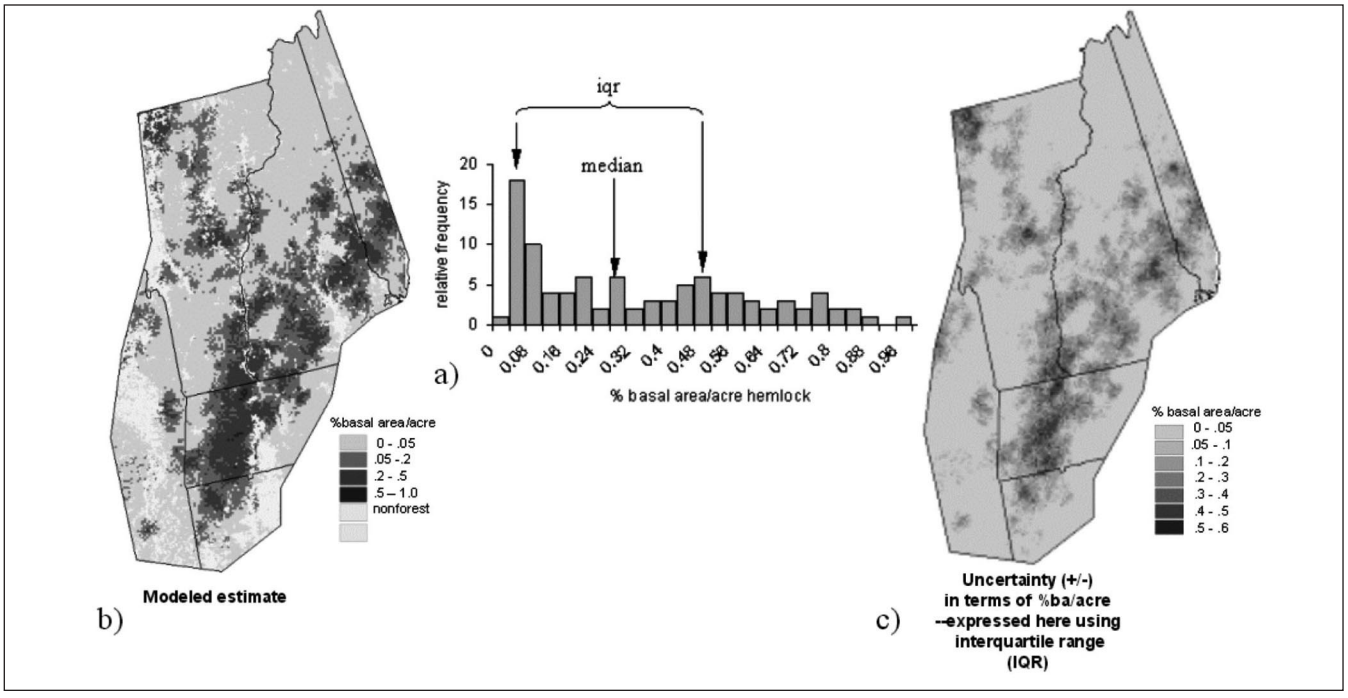


Figure 2.—*SGCS creates a distribution of possible values for each pixel (one generated with each simulation) from which the user can easily extract a clear measure of the uncertainty of each local estimate. a) the distribution of values at a single, randomly chosen cell; b) the modeled estimate where the value at the 65% percentile was chosen for each pixel (with nonforest areas masked out); and c) the value of the interquartile range (iqr) at each pixel, representing the range of uncertainty associated with each modeled value.*



ple, species distribution maps were created using SGCS (Riemann *et al.* 1997, Riemann and Lister,[3] http://www.fs.fed.us/ne/fia/spatial/specdist/species_dist.html). Like kriging, SGCS uses a model of the spatial structure present to estimate values at unknown locations (fig. 1). However SGCS differs in that it is a stochastic simulation technique that uses a random function to incorporate uncertainty into the model (Rossi *et al.* 1993). This creates a distribution of possible values for each pixel (one generated with each simulation) from which the user can easily extract a clear measure of the uncertainty of each local estimate (fig. 2). In addition, depending on the summary statistic chosen as the estimate (e.g., mean, median, or another percentile) this technique provides spatial output with a more realistic depiction of heterogeneity and is more effective than kriging in retaining original data characteristics.

Advantages and Limitations of SGCS

The use of SGCS offers several advantages. First is the way Monte Carlo techniques offer a clear depiction of the model uncertainty, which reflects both sample intensity and variation in the available data. This feature is valuable because some uncertainty is inevitable in all modeled output, and the magnitude and direction of uncertainty are important aspects of any modeled map created. In addition, since there is a probability distribution built for each location of the map, the user can depict the error in different ways based on her/his goals and choice of risk (e.g., of overestimating or underestimating). For example, if in a study of the hemlock woolly adelgid (an insect associated with hemlock mortality), the cost of sending crews to a site with an insufficient amount of hemlock is greater than missing a site that might have sufficient hemlock to complete the study, i.e., the cost of overestimation is greater than that of underestimation. The user would thus want to choose a depiction of error that reflects the wish to avoid wasting field crew resources—one that reduces the risk of overestimation.

Like all techniques, however, there are also limitations to SGCS that make us want to pursue additional approaches. First, not all variables of interest have a strong spatial structure that can be modeled and used in SGCS to improve estimates. Second, satellite data and other relevant spatial data layers (e.g., topographic information, climatic information, soils information), are becoming increasingly available and affordable and are of increasing quality. Also, our technical ability to display and manipulate these data is continually improving. Finally, because of the sampling intensity of FIA data and the level of unexplained variation typically remaining, using only FIA data in the modeling limits the spatial resolution and levels of uncertainty in the final output.

To address the shortcomings of univariate geostatistical methods, we are incorporating this increasing wealth of additional ancillary environmental information into the modeling process. Many multivariate modeling techniques are available for this, each of which utilizes and maintains different characteristics of the sample data, has different output characteristics, and makes different assumptions. Many of these methods are being investigated throughout FIA (e.g., Lister and Hoppus 2002; McRoberts *et al*. 2002; Moeur and Riemann 1999; Moisen and Frescino 2002; Ohmann and Gregory 2002). The goal of our study was to investigate the use of multiple linear regression to make predictions of FIA attributes by answering the following questions:

- Does spatial structure matter when satellite-derived and ancillary spatial data are incorporated into the modeling of forest attributes?
- If we use a modeling technique such as multiple linear regression, how do we calculate or estimate the uncertainty?
- Using this technique, what are the characteristics of the output data set based on the original sample data and the ancillary data used?

Methods

Data.—In all, 141 100-percent forested FIA plots from the 1998 inventory were used from the study area in central and southern New Jersey. FIA defines a forested plot as being at least 1 acre in area and 120 feet wide, having a minimum of 10-percent stocking, and an undisturbed understory.[4] Variables calculated from plot data and chosen for modeling were: the relative basal area of pitch pine (pp%ba), relative basal area of red maple (rm%ba), and total

[3] Riemann, R. and Lister, A.J. Stochastic simulation for mapping ground inventory variables: Creating and using the species distribution maps. Unpublished report on file at USDA FS NE-FIA, 11 Campus Blvd., Newtown Square, PA 19073.

coniferous basal area on a plot (conifba). Because of their skewed distributions, all three variables were normal–score transformed before all analyses—a 1-1 linear transform to a perfectly normal distribution (Deutsch and Journel 1998).

Only forested plots were used because we wanted to model the characteristics of forest land rather than the distinction between forest and nonforest land. Nonforest land was applied afterward as a mask on the modeled output, using data derived from a classification dedicated to accurately identifying those classes (Zhu and Evans 1994). Separating these two tasks and focusing on modeling only the forested population was considered a valuable part of the modeling exercise. Mixed plots, those partly on forest land and partly on nonforest land, add an additional complication when attempting to match plot data with the 30-m pixel of Landsat Thematic Mapper (TM)-derived layers. With mixed plots, being off by even a half pixel in co-registration of the data sets can place the plot in an entirely different land use class, which is well below the georeferencing accuracies currently achievable. This did reduce the usable number of plots available (from 206 forested and mixed plots to the 141 completely forested plots). Future studies should explore ways to address this uncertainty and use this plot information, but in this study mixed plots were simply removed.

The following predictor variables were used:

- Spectral information derived from Landsat ETM+ (USGS Eros Data Center 2002): three tassel cap bands each from three seasons (spring, summer-leaf on, fall-leaf off), image dates: approx. 2000 (range: 1-255).
- Topographic variables derived from the 30-m digital elevation model (DEM): elevation, slope, aspect, position indicator (range: 0-100, representing location between the valley (0) and the ridgetop (100)).
- Soils variables derived from STATSGO (USDA 1993): soil quality, soil carbon, available water content (soil_awc) (range: 1-255).
- Spatial information – X, Y (converted to a range of 1-255).

Modeling Approach.—Descriptive statistics were calculated and plotted to assess the characteristics of the dependent variables, and the data were normal-score transformed (a 1-1 linear transform of the data to a perfectly normal distribution) before further analysis. Variography was performed and variogram descriptive statistics were calculated to assess the degree of spatial continuity of each of the dependent variables. Scatterplots were constructed and correlation statistics were produced to assess the degree of correlation with predictor data and eliminate predictor data layers that were not linearly related to the dependent data layers. Data redundancy was reduced by removing one variable from pairs of predictor variables that were collinear. We performed stepwise linear regression to make maps of predicted pp%ba, rm%ba and conifba, both including and excluding X and Y as possible predictor variables. To assess the accuracy of the regression model, we analyzed the model fit and performed a tenfold cross-validation procedure in which successive sets of 10 percent of the data were withheld from the model and subsequently predicted. Scatterplots of observed vs. actual values and residual plots were produced and assessed. Finally, we compared the characteristics of the output data sets with those of the original sample data, looking for differences that might be effects of the modeling technique or data sets used.

Results and Discussion

Characteristics of the Data.—The skewed distributions of the dependent variables indicated a probable need for transformation before analysis, and indeed a clearer spatial structure and stronger correlations with the independent variables were observed with the normal-score transformed data.

The following are results of variography and correlation analyses:

Species	Variogram nugget (% of variation explained)	r value
pp%ba	0.45 (55%)	0.4–0.6
rm%ba	0.8 (20%)	0.2–0.3
conifba	0.55 (45%)	0.5–0.7

Both pitch pine relative basal area (pp%ba) and total coniferous basal area (conifba) had a strong spatial structure and fairly strong correlations with the independent/explanatory variables. Red maple had weak spatial structure and weaker correlations with the independent/explanatory variables.

[4] U.S. Department of Agriculture, Forest Service. 2000. Forest inventory and analysis national core field guide, volume 1: field data collection procedures for phase 2 plots, version 1.4. Unpublished report on file at USDA Forest Service, Forest Inventory and Analysis, Washington, DC.

Models.—The model developed by the linear regression using the full set of sample points and used for producing the final conifba map was:

conifba = -1.22186 + (0.03134*fall_wet) – (0.02951*fall_bright) + (0.02204* spr_green) - (0.01713 * sum_green) + (0.00265*y) - (0.00906*soil_awc)

$R^2 = 0.62$, p = 0.13, RMSE = 0.63.

The model used for the rm%ba final map was:

rm%ba = -1.1885 + (0.03983*sum_green) – (0.03552*fall_green) – (0.00774*position_ind) + (0.00825*soil_awc)

$R^2 = 0.25$, p = 0.046, RMSE = 0.88.

The model for predicting pp%ba was:

pp%ba = 1.63594 + (0.03333*spr_green) - (0.02823*sum_green) - (0.01128*soil_awc)

$R^2 = 0.54$, p = 0.0006, RMSE = 0.686.

The same larger set of potential predictor variables was provided for the development of the 10 validation models. Each of these models was similar to that developed from all the sample points.

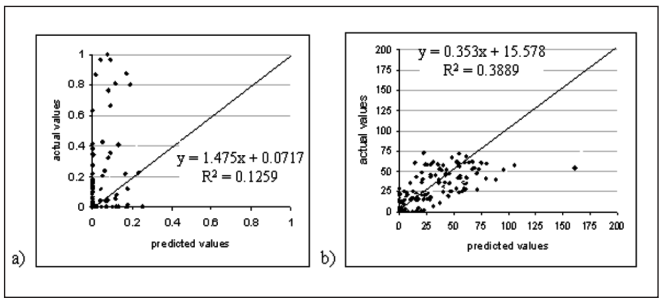
Calculating and Depicting Map Error/Uncertainty

How good is the model? How close to reality is the output map? How likely are we to find on the ground what is depicted in my map? A measure of uncertainty associated with each estimate can reveal some pertinent information, and it can capture one or more of the above elements of uncertainty since they are related. The most directly interpretable values for the user are measures of how the output map relates to reality, typically in terms of comparison to point locations on the ground. However, depending on the final use(s) of the output data set, other characteristics may also be important, such as the accuracy of area summaries or the spatial distribution of features.

We assessed the uncertainty of our regression models by examining the results of the tenfold cross validation. This involved dividing the data into 10 random parts, each containing 90 percent of the data, and running linear regression on each of the 10 sets. Each of the original values was then compared with that value predicted using the model created without it. Scatterplots of observed vs. predicted backtransformed values from this validation for each of the dependent variables are shown in figure 3. From these validation data we can calculate an RMSE value to describe the uncertainty associated with our output map. The validation RMSE with an average uncertainty of +/- 23 percent for estimates of red maple relative basal area, and 31 ft^2 for estimates of total coniferous basal area. However, RMSE is only a single average value for the entire map. How does the magnitude and direction of that difference vary with location and predicted value?

The previous analysis estimated the error associated with each known value. We also wanted to produce a spatial depiction of model uncertainty. Assuming that the validation we conducted using plots for which we had data provided a picture of the distribution of possible prediction errors, we grouped all the predicted values into classes, ensuring that enough plots fell in each, and calculated the average validation error associated with that class (table 2). We then reclassified our table of predictions to create an error map (fig. 4a) using table 2 as a lookup table. Figure 4b is the associated map of estimated values. Note that we are also assuming that we can associate a pixel (30- x 30-m grid cell) with a plot (a 4-point cluster of 1/24-acre circular plots spread over approximately 1 acre).

Figure 3.—*Actual vs. predicted values for a) proportion of red maple basal area and b) total coniferous basal area. Diagonal line is the line of 1:1 agreement.*



Checking Characteristics of the Output Data Set

What are the characteristics of the output maps in relation to the original sample data and the ancillary data used? When we checked the univariate statistics of the output data set, the predicted data duplicated the original data's sample histogram for all three variables fairly well (fig. 5). Looking at the correlogram/spatial structure of the output data set, the output is spatially more smoothed than the original data (table 3). This is partly due to the characteristics of the predictor data sets used.

Joint attribute structure—the relation between the forest variables predicted—is another important characteristic for those interested in querying the output data sets together. Although we did not model the three variables together, and thus did not explicitly attempt to retain this information in the modeling, it is nevertheless important to understand how this is expressed in the output data sets. Comparing the joint attribute structure in the output to the original data shows substantial similarities but also some truncation of the original ranges of values (fig. 6). Finally, looking at the output map and the input data sets that were used to create it, we observed that the characteristics of the source data sets were influencing the output in ways that may be undesirable. For example, in figure 7d, areas with the highest estimates (along the shorelines) are clearly influenced strongly by soil available water content (soil_awc) values (7a). In figure 7b, some relics of how the position_index value was calculated appear as "contour lines" in the rm%ba map. Such effects may be important clues to the driving factors associated with particular species, or, in this case, more likely suggest room for improvement in the model and/or in the

Table 2.—*Calculating average error for each class of predicted value—i.e., for each range of predicted values, the range of possible actual values*[a]

predicted classes	count	avg minus error (est too low)	avg plus error (est too high)	avg plus_minus error
0	25	3	0	2
10	17	6	3	5
20	19	4	9	6
30	24	12	15	13
40	11	13	15	14
50	16	18	13	16
60	17	16	19	17
80	12	21	23	22

[a] The average plus error can be sufficiently different from the average minus error such that one might want to depict them separately.

Figure 4.—*Map of the predicted values (a) and of the estimated uncertainty plotted from table 2 (b).*

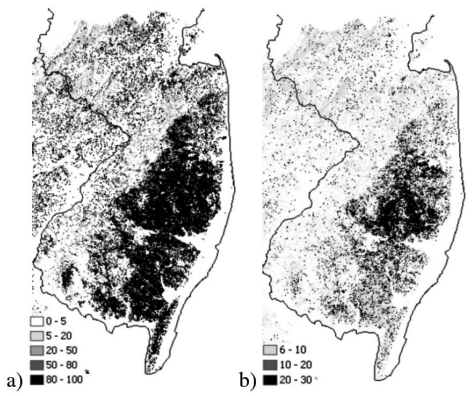


Figure 5.—*Univariate histograms of the sample data compared with predicted values.*

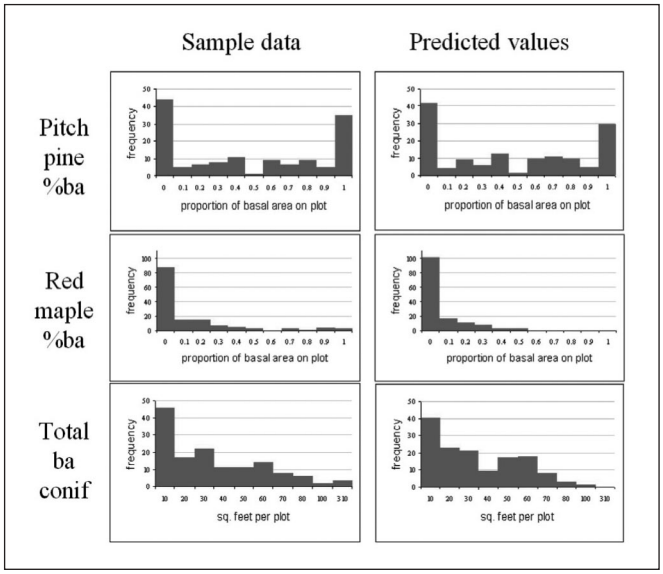
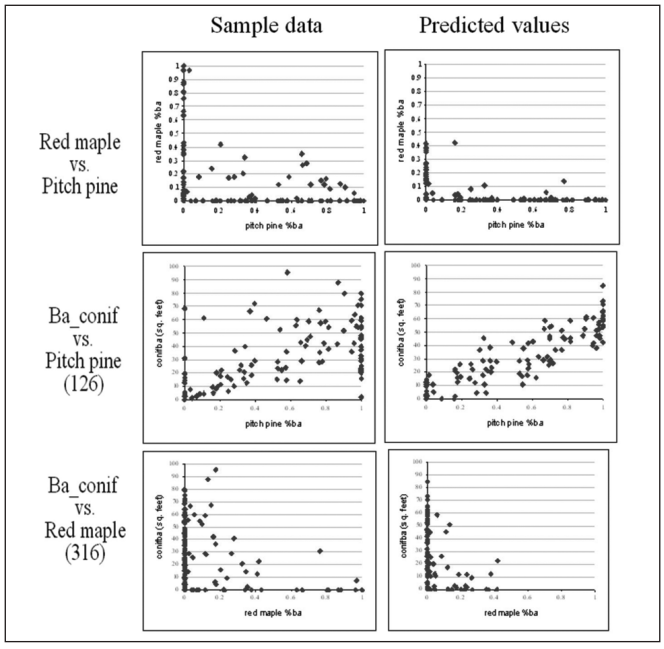


Figure 6.—*Scatterplots of the joint attribute structure of original and predicted data.*



independent/source data sets when they become available. The results also suggest that in future studies we may want to reevaluate the inclusion of ancillary spatial layers that adversely affect the final maps.

Incorporating Spatial Structure

On the basis of the nugget values calculated during the variography analysis, there is spatial structure in the sample data, particularly with relative pitch pine basal area and total coniferous basal area. In addition, the simple spatial location variables of standardized X and Y demonstrate some level of correlation (r values of 0.33 to 0.56 for X and 0.02 to 0.12 for Y)) with all three species or species-group variables. Whenever X or Y was one of the final independent variables in the model, it always smoothed the output data set (fig. 8), and did not noticeably change the amount of spatial structure remaining in the residuals (table 4). This limited effect of including or excluding X and/or Y as predictor variables in the model is probably because much spatial information is already implicitly contained in the satellite imagery and other spatial data sets when these are strongly correlated to the variable of interest. For variables that are both poorly correlated with the independent variables and contain a high level of spatial structure, incorporating spatial information could introduce important additional information (Goovaerts 1999, Moeur and Riemann 1999). None of the variables investigated contained this combination of characteristics. However, where including the spatial structure in the modeling is desirable, multivariate linear least

Table 3.—*Comparison of nugget values in the correlograms of the original values and the predicted values, expressed as % of the sill. A lower nugget indicates more spatial structure; there is a noticeable difference between the original and predicted values here.*

Item	Original	Predicted
rm%ba	80	12
conifba	55	8

Table 4.—*Nugget values in the residuals, expressed as % of the sill. A low nugget often indicates spatial structure not explained by the model. Including X in the model did not reduce the amount of spatial structure in the residuals.*

	Model	
Item	With X	Without X
rm%ba	52	68
conifba	100	100

Figure 7.—*Degree to which the spatial characteristics of the predictor data sets can contribute to the final map. This may or may not be realistic, but in the Northeast, where things are seldom driven so cleanly by a single variable, we would tend to suspect this as a characteristic of the data and model rather than of the phenomena: a) soil_awc; b) position_indicator; c) plus/minus error associated with predicted estimate; d) predicted %ba value for red maple.*

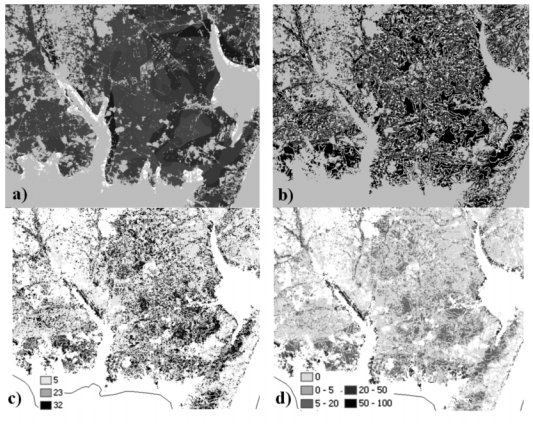
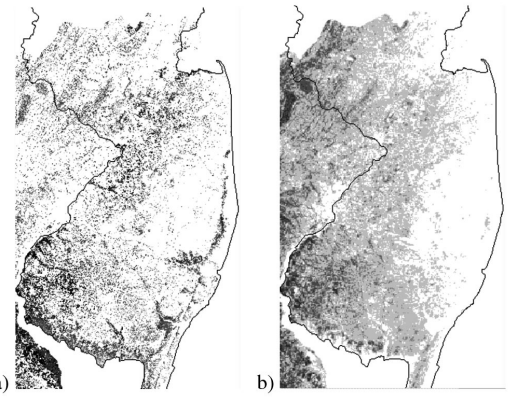


Figure 8.—*Effect of using X in the model for estimating red maple relative basal area (rm%ba): a) a map of the model developed excluding XY as optional predictor layers (-0.73814 + (0.01412 * sum_green) – (0.00759 * position_ind) – (0.00593 * X)), b) a map of the model developed including XY as options (-1.18850 + (0.03983 * sum_green) – (0.03552 * fall_green) + (0.00825 * soil_awc) – (0.00774 * position_ind)). Note the increased smoothing of the results in b) when the smoothly varying variable X is included.*



squares regression is probably not the best tool to take advantage of spatial information. Because it is mostly a per pixel modeling technique, linear regression does not easily incorporate information on the value, distance, or direction of neighboring data when producing an estimate. Other approaches may be much more effective in taking advantage of this spatial information.

Conclusions

Know the characteristics of your data and the phenomena going in—it will help greatly in the modeling, both in effectiveness and accuracy, for each has different characteristics that can affect the effectiveness of a particular modeling technique. Similarly, check the characteristics of the data coming out—it will help in understanding the characteristics of the output data set, which, in turn, may result in iterative improvement even if it does not contribute directly to a calculation of error/uncertainty. As stated earlier, the characteristics of the spatial data sets used in predicting the variable of interest affected the characteristics of the output. Such effects may be important clues to the driving factors associated with particular species, or, more likely in this case, signal possible improvements in the model and/or in the independent/source data sets.

Having some measure of uncertainty is SO important! Each of these maps is only a modeled estimate of what is occurring on the ground, so there is *always* some level of uncertainty as to the degree to which the modeled map reflects reality. Providing a measure of uncertainty with each estimate gives the user additional information to work with. With linear regression, a single validation RMSE value, such as a single %accuracy value for a classified map, doesn't tell the whole story with the mapped output of a regression model because we know there is spatial and class variability to that error and we want to know where that occurs. A single RMSE value calculated without the use of a validation data set is even less satisfactory because an uncertainty value calculated only from data that went into the modeling does not account for errors in our admittedly less-than-perfect input data sets. Error/uncertainty from comparisons between modeled estimates and plot data can be calculated when both are of a reasonably similar scale/resolution (e.g., a 30-m pixel with a 1/6-acre clustered plot).

Does spatial structure matter when we incorporate satellite-derived and ancillary spatial data sets into the modeling of forest attributes; that is, do we lose potential information by not including it? In this study, bringing the satellite and other layers into the modeling seemed to account for much of the variability that the spatial structure was describing. However, if we wish to include spatial structure in the modeling (as might be the case with a variable with weak correlations with the independent variables but a strong spatial structure), bringing X and Y into the regression as simple variables is only a partial solution with some consequences (i.e., smoothing), and is probably not the best technique for this task.

There are many characteristics of a spatial data set. Many maps are a compromise of some characteristics in favor of others, e.g., smoothing the map to discern patterns at the expense of local heterogeneity. Which aspects are most important in the output map will depend on how the data will be used. What the maps are being used for will direct/dictate how we look at them, what we consider to be accurate (or the most important aspect of accuracy), what we consider to be the dominant characteristic that makes us accept or reject a map, and what modeling techniques we choose because of the characteristics they preserve or the characteristics of their output.

The models produced from linear regression procedures in this study are by no means the best that could be obtained given the independent data currently available. However, they point out potential characteristics and tendencies that can result when input variables are used in a regression model to simply predict forest attributes by their relationship to other spatial variables. Characteristics of the input data sets, which may be derived from numerous sources and via other modeling techniques, can greatly influence the characteristics of the output.

Literature Cited

Deutsch, C.V.; Journel, A.G. 1998. GSLIB: geostatistical software library and user's guide. 2nd ed. New York: Oxford University Press.

Goovaerts, P. 1999. Performance comparison of geostatistical algorithms for incorporating elevation into the mapping of precipitation. In: Proceedings, IV international conference on geocomputation; 1999 July; Fredericksburg, VA. 17 p. www.geovista.psu.edu/sites/geocomp99/Gc99/023/gc_023.htm. [31 January 2003]

Lister, A.J.; Hoppus, M. 2003. Small area estimation using FIA plots and satellite imagery. In: Proceedings, 2002 IUFRO seminar on statistics and information technology in forestry; September 8–12; Blacksburg, VA. Blacksburg, VA: Department of Forestry, Virginia Polytechnical Institute and State University: 45–52.

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 technique. Remote Sensing of the Environment. 82(2/3): 457–468.

Moeur, M.; Riemann, R.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: 419–430.

Moisen, G.G.; Frescino, T.S. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling. 157: 209–225.

Ohmann, J.L.; Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, USA. Canadian Journal of Forest Research. 32: 725–741.

Riemann, R.H.; 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. *et al.*, eds. Modelling longitudinal and spatially correlated data: methods, applications, and future directions. Lecture notes in statistics, vol. 122. New York, NY: Springer-Verlag: 187–198.

Rossi, R.E.; Borth, P.W.; Tollefson, J.J. 1993. Stochastic simulation for characterizing ecological spatial patterns and appraising risk. Ecological Applications. 3(4): 719–735.

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

Zhu, Z.; Evans. D.L. 1994. U.S. forest types and predicted percent forest cover from AVHRR data. Photogrammetric Engineering & Remote Sensing. 60(5): 525–531.