

# **Geostatistical Prediction and Mapping for Large Area Forest Inventory Using Remote Sensing Data**

Qingmin Meng

Warnell School of Forestry and Natural Resources,  
University of Georgia, Athens, GA 30602 USA

E-mail: qmeng@uga.edu

Tel: 706-542-1180 (O)

**Abstract:** Large area forest inventory is important for understanding and managing forest resources and ecosystems. The purpose of traditional large area forest inventory is to provide unbiased and reliable forest resource information, though typically these inventories lack fine spatial resolution. Remote sensing, the Global Positioning System (GPS), and geographic information systems (GIS) provide new opportunities for forest inventory. By integrating remote sensing, GPS, and GIS, it is possible to predict forest parameters at fine spatial resolutions. The research described here develops a new systematic geostatistical approach for large area forest inventories, where one type of forest parameter, such as basal area, height, health conditions, biomass, or carbon can be incorporated as a response variable and the geostatistical approach can be used to predict un-inventoried points. Using basal area as an illustration, this approach includes univariate kriging (ordinary kriging and universal kriging) and multivariable kriging (co-kriging and regression kriging). The combination of bands 4, 3, and 2, as well as the combination of bands 5, 4, and 3, along with normalized difference vegetation index (NDVI) and principal components (PCs) are used in co-kriging and regression kriging. Cross validation using the training dataset and validation based on 200 random sampling points indicate that the regression kriging is the best geostatistical method for spatial predictions of pine basal area. Finally, pine basal area is mapped using regression kriging, and standard errors also are mapped to assess the dispersions of the spatial prediction.

**Key words:** Geostatistical approach, kriging, regression kriging, remote sensing, Mapping

## **Introduction**

Large area forest inventories generally are based on plot sampling, and small area forest inventories usually are processed forest stand units. These two traditional inventories can be integrated by combining ground inventory and remote sensing data and processing them in geographical information systems (GIS).

Remote sensing, the Global Positioning System (GPS), and GIS provide new opportunities for forest inventory. It is now easy to measure the locations of survey plots, forest stands, and stand boundaries in the field with an accuracy rate of  $\pm 5\text{m}$  using differential GPS. Developments

in sensor technology have also enabled acquisition of remotely sensed data at a range of scales. Remote sensing data are available from satellite sensors providing images with medium spatial resolution of 20~30 m (Landsat TM, Landsat ETM+, SPOT HRVIR) as well as high spatial resolution of less than 5 m (Ikonos, QuickBird, LIDAR, and others). Integration of these technologies allows achievements in forest metrics using raster data with cell sizes of 30 m, 20 m, 10 m, 5 m, or 1 m. These raster data can be estimated from remote sensing data by modeling the relationships between the image's digital numbers (DN) and the forest variables inventoried with GPS. Geographic information systems and spatial modeling are efficient tools to model, estimate, map, and predict spatial characteristics of stands or trees. Generally, the two ways to obtain the fine spatial forest information are spatial modeling and nonspatial modeling.

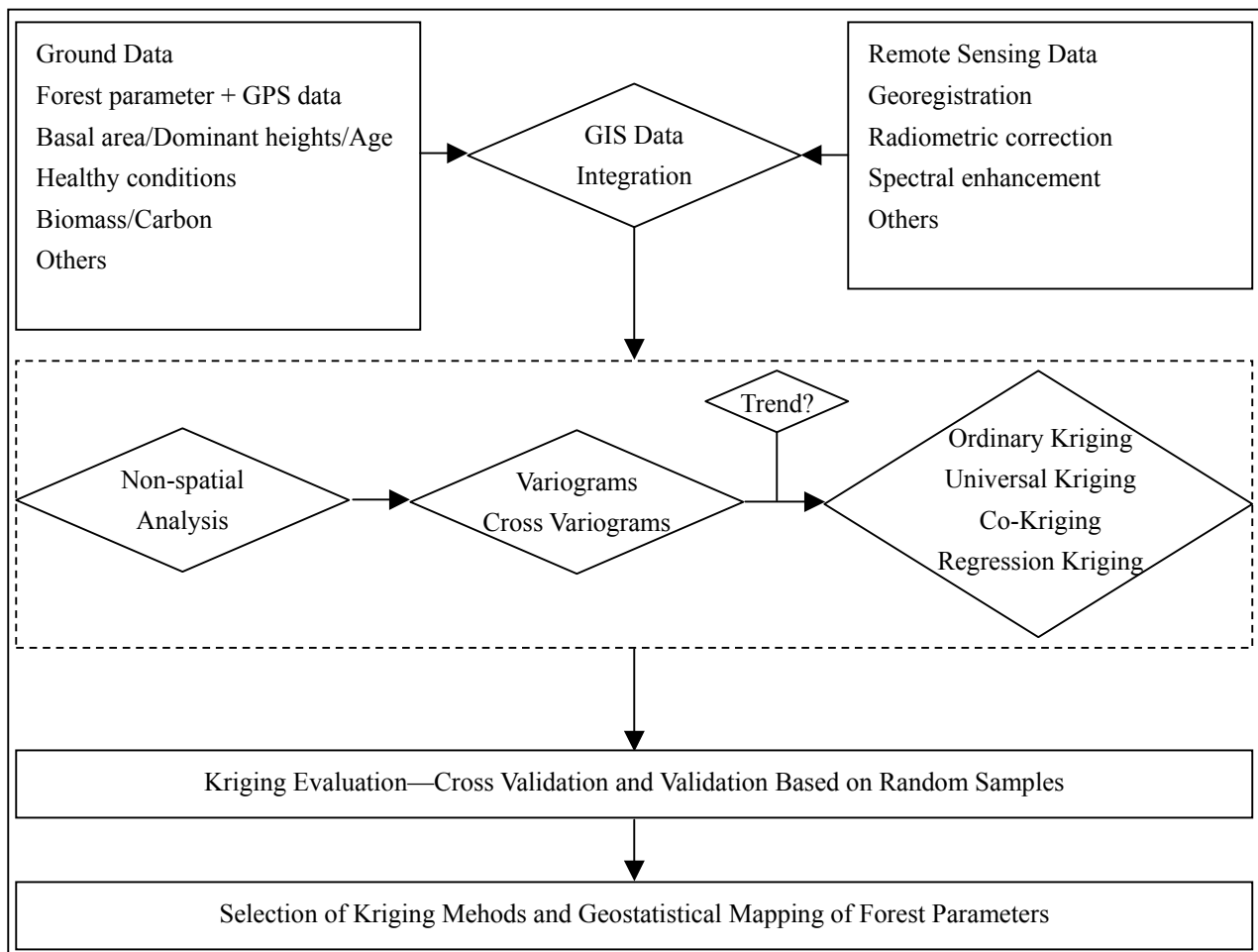
Nonspatial modeling methods have been widely applied in forest research. Ordinary least-squares (OLS) regressions are the common models applied for estimations of forest variables (Ardö, 1992; Dungan, 1998; Trotter, Dymond & Goulding, 1997). K nearest neighbor (KNN) methods for achieving forest metrics using remote sensing data have been applied in Finland and America for forest inventories (Franco-Lopez, et al. 2001; Holmström and Fransson, 2003; Moeur and Stage, 1995; Tomppo, 1991). Artificial neural networks also are used for estimating forest variables using remote sensing data (Foody & Boyd, 1999; Foody, 2000; Tatem *et al.*, 2001).

Tokola *et al.* (1996) applied both linear regression and the KNN method on forests in the southern boreal vegetation zone in Finland using data from Landsat TM and SPOT. The authors reported standard errors of stem volume prediction from 70 to 80 m<sup>3</sup>/ha (more than 60% of the mean) at the plot level. Trotter *et al.* (1997) used Ordinary Least Squares to predict stem volume of mature plantations in New Zealand and reported a root mean square error (RMSE) greater than 100 m<sup>3</sup>/ha (with a mean stem volume of 413 m<sup>3</sup>/ha) for pixel predictions. The K nearest neighbor method was applied by Holmström & Fransson (2003) to predict forest variables using a combination of SPOT-4 and low-frequency radar data from the airborne CARABAS system. The study by Holmström & Fransson (2003) used data from a coniferous forest in southwest Sweden and reported RMSEs of 64% (of the mean) of stem volume using optical data and 53% using the combination of optical and radar data. The stem volume of the sample plots (10 m radius) was in the range of 0-750 m<sup>3</sup>/ha with a mean value of 171 m<sup>3</sup>/ha.

Many studies have conducted spatial predictions using remotely sensed data (Atkinson *et al.* 1994; Atkinson and Lewis, 2000; Chica-Olmo and Abarca-Hernandez, 2000; Curran, 1988; Curran and Atkinson 1998; Dungan 1998; Dungan, *et al.* 1994; Lark, 1996). Few studies have been conducted on estimations of forestry relevant variables using spatial models, though, a large number of spatial-statistical and prediction models are available in the literature (e.g. Cressie, 1993; Goovaerts, 1997; Odeh, *et al.* 1995; Odeh and McBratney, 2000; Wackernagel, 1994). Berterretche *et al.* (2005), Tuominen *et al.* (2003), and Zhang *et al.* (2004) applied geostatistical models to estimate forest variables, leaf area index, and classify forest lands based on remote sensing data. Gilbert & Lowell (1997) used kriging to predict stem volume in a 1500 ha balsam fir (*Abies balsamea*) dominated forest. Prediction based on 5.6 m and 11.3 m radius plots resulted in a prediction RMSE of 54% (of the mean) and 39-46%, respectively. Similar accuracy was

obtained by prediction using the sample average only. Methodologically, the accuracy rate of the predicted variable could be improved by incorporating close field observations as predictors in spatial modeling.

Rarely has research explored the integration of remote sensing data, GPS, ground data, GIS, and geostatistics to estimate forest parameters at a fine spatial resolution for large areas. One systematic geostatistical approach for spatial forest inventory is developed and explored in this research. Compared to the typical ordinary kriging (OK) and universal kriging (UK) using only one variable, this research develops a systematic geostatistical approach—co-kriging (CoK) and regression kriging (RK) using remotely sensed data as predictors—to improve spatial predictions of forest variables by integrating GPS, ground inventory data, remote sensing, and GIS. This systematic geostatistical approach is summarized in a flow chart (Figure 1), and provides new insights for forest parameter estimation, and not only considers the associations between one forest parameter and DN but also incorporates the spatial dependence of the forest parameter into the process of spatial prediction. In this study, basal area is used as the response variable to conduct this geostatistical approach.



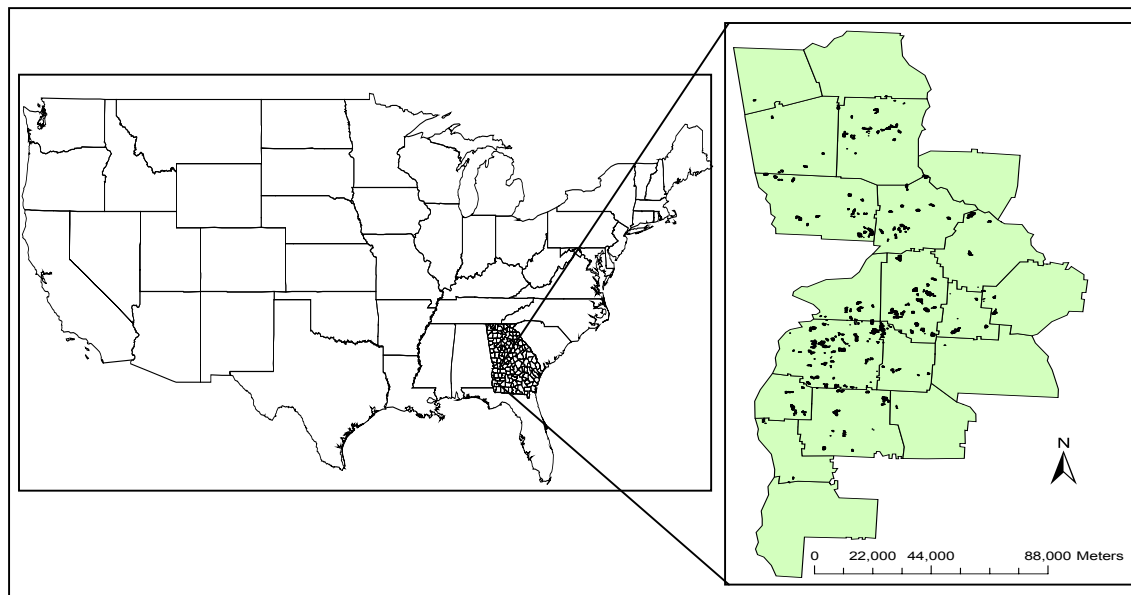
**Figure 1.** A systematic geostatistical approach for predicting forest parameters using remotely sensed data.

In addition to analyzing spatial characteristics of an integrated GIS with ground and remote sensing data, it is also necessary to analyze nonspatial data; for example, the selection of band combinations and data reduction of remotely sensed imagery. What is the association between the response variable and independent variables i.e., the remotely sensed data? Distribution tests may be needed. However, the kriging equations are derived without being based on any distributional assumptions (Myers, 1996). Correlation diagnostics are important for multivariable geostatistics. The variogram models are fitted to check spatial autocorrelation and dependence. Cross variograms need fitting if multivariable geostatistical approaches are conducted. Additionally, it is important to check whether a spatial trend exists in the data of the response variable. Both universal kriging and regression kriging are efficient to incorporate the trend in geostatistical predictions.

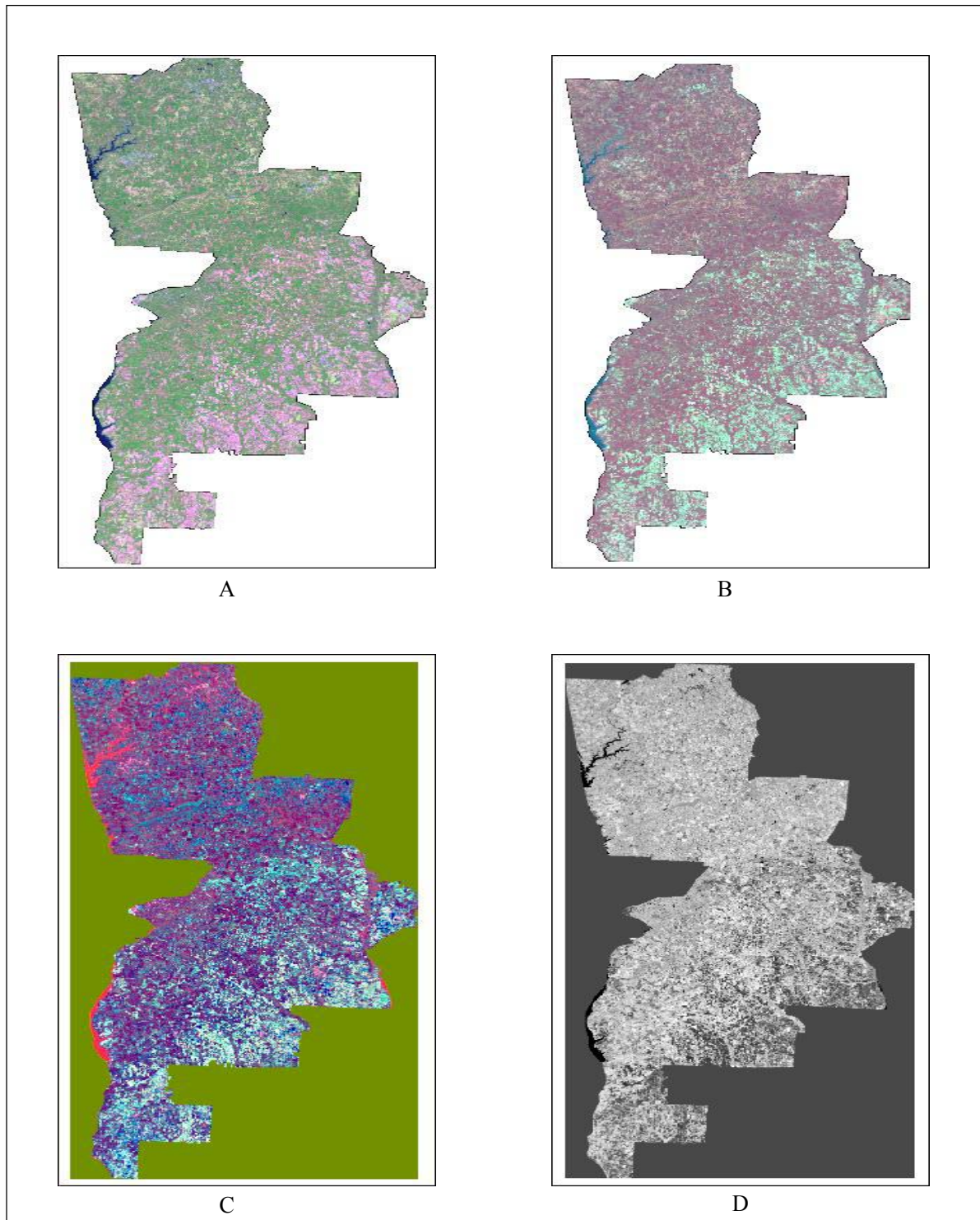
## Data Sources

### Ground data

Ground data covering 20 counties in west Georgia were inventoried in 1999 (Figure 2). These locations of ground data were collected using differential Global Positioning System (DGPS) units with errors of about  $\pm 5$  m. The coordinates of the ground data were converted to the Universal Transverse Mercator to match those of the Landsat ETM+ images (Figure 3). There were 2822 ground records used in this study with a mean of  $13.99 \text{ m}^2/\text{ha}$  and a range from  $0.038$  to  $29.84 \text{ m}^2/\text{ha}$ . The basal area and dominant height were measured, and volume of trees was calculated according to tree species. Basal area of pines is used as the only response variable in this study. Basal area at the Landsat pixel level (30 m) is predicted for the un-inventoried areas in these 20 counties.



**Figure 2.** The study area includes 20 counties in the State of Georgia. The ground inventory locations are indicated as the dark dotted places in these 20 counties.



**Figure 3.** Landsat ETM+ images used for pine basal area prediction. A, a 543 band combination; B, a 432 band combination; C, the three PCs images; D, the NDVI images.

### **Remote sensing data**

Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images (Path/Row: 19/37 and 19/38) acquired on 10 September 1999 from the USGS Earth Resource Observation System Data Center were used in this research. Atmospheric conditions were clear at the time of image acquisition,

and the data had been corrected for the radiometric and geometric distortions of the images to the standard Level 1G before delivery. Two Landsat images covering this study area were masked after the geometric corrections using U.S. Geological Survey (USGS) digital orthophoto quarterquads (DOQQs) as the sources of control (RMSE is less than 10 m). This resulted in a 4449 pixel by 9010 row 6-band (i.e., 1, 2, 3, 4, 5, and 7) image for analysis.

#### *Band Combinations*

Band 1 of Landsat images contributes little for vegetation analysis. Studies indicate that as the leaf coverage changes from 0% to 11.9%, 43.2%, and 87.6%, very little change occurs in the reflectance of band 1 (0.4-0.5  $\mu\text{m}$ ) (Short, 1999). The differences of reflectance increase from 0.5 to 0.8  $\mu\text{m}$  as leaves change. The differences of reflectance in the mid-infrared ranges are very close to the differences in the near infrared ranges. Band 7 of Landsat images is not used as an independent variable. Bands 2, 3, 4 and 5 are used, and 432 and 543 band combinations are applied to estimate pine basal area.

#### *Principal Component Analysis*

Principal component analysis (PCA) is the most frequently used technique for remote sensing data reduction. Generally, remotely sensed data, such as Landsat images, are highly correlated among the adjacent spectral bands (Barnsley, 1999). The Landsat bands are transformed into orthogonal principal components (PC). The first PC contains the largest percentage of data variation, and the second PC contains the second largest variance of the data, and so on. The higher the PC is numbered, the less useful information the PC contains. In this research, the six Landsat ETM+ bands used (i.e., band 1, 2, 3, 4, 5, and 7) were processed using PCA, and the first three PCs were applied for pine basal area analysis because they accounted for more than 95% total variance.

#### *Normalized Difference Vegetation Index*

In this study, the normalized difference of vegetation index (NDVI) is used for pine basal area estimation. NDVI is based on a ratio of the near infra red (NIR) and the red channels, and the standard equation for NDVI is as in equation 1.

$$NDVI = \frac{NIR - Re d}{NIR + Re d} \quad (1)$$

Healthy forests reflect strongly in the near-infrared portion of the spectrum while absorbing strongly in the visible red. On the other hand, soil, bare ground, and rock show near equal reflectance in both the near-infrared and red portions and have NDVI values close to 0 while water bodies have the opposite trend to vegetation and the index is negative. The NDVI image can significantly enhance the discrimination of vegetation cover from other surface cover types. The values of NDVI generally range from 0.05 for sparse vegetation cover to 0.7 for dense vegetation cover (Tucker, 1979). It not only measures both the amount of green vegetation and vegetation health in an area, but it also is a basic indicator of changes in vegetation over space and time. It

has been extensively applied as a proxy for leaf area index (Tucker, 1979), vegetation biomass (Seller, 1987), and net primary production (Goward et al, 1985). Therefore, NDVI indicates the spatial characteristics of forest stand development, especially the density and health of trees. It has been proven to be an efficient indicator in detecting and quantifying large-scale changes in plant and ecosystem processes (Braswell et al. 1997; Myneni et al. 1997).

## Methodology

### Correlation analysis

Correlation analysis was applied to measure the strength of association between the response variable and the independent variables. Pearson's product-moment correlation ( $r_{xy}$ , equation 2)

coefficient and the Pearson partial correlation ( $r_{xy \cdot z_1 z_2}$ , equation 3) coefficient were used to

measure the association between the response variable and the independent variables. Pearson's product-moment correlation measures the association without considering the correlation contributions from other independent variables. The Pearson partial correlation measures the strength of a relationship between two variables while controlling the effects of two additional variables. Therefore, it is called the second-order partial correlation indicating the partial correlation between  $x$  and  $y$  controlling for both  $z_1$  and  $z_2$ .

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (2)$$

$$r_{xy \cdot z_1 z_2} = \frac{r_{xy \cdot z_1} - r_{xz_2 \cdot z_1} r_{yz_2 \cdot z_1}}{\sqrt{(1 - r_{xz_2 \cdot z_1}^2)(1 - r_{yz_2 \cdot z_1}^2)}} \quad (3)$$

### Geostatistical approach

Geostatistical methods are based on the theory of regionalized variables (Matheron, 1965), which makes the assumption that data are observations of stochastic variables. We can consider a spatial variable as a realization of a random function represented by a stochastic model.

One of the key steps in geostatistical modeling is the semivariogram, a function describing the spatial dependence of the spatial variable. The semivariogram has been used widely in remote sensing to determine spatial structures (Curran, 1988; Warren, et al., 1990; Atkinson & Lewis, 2000). Based on the semivariogram, the geostatistical process derives optimal linear unbiased spatial prediction methods (i.e., kriging) by minimizing mean-squared prediction error. However, the assumptions of stationarity, which often are not met by the field-sampled data sets, and the requirement of a large dataset to define the spatial autocorrelation, result in the limitations of univariate kriging. Fortunately, geostatistical methods also provide optimal prediction methods

using auxiliary data. Large volumes of auxiliary data for forest research are available now, such as remote sensing data. Incorporating the auxiliary data, co-kriging and regression kriging, as described below, can increase prediction accuracy. The gstat package (Pebesma, 2005) is mainly referenced for variogram and kriging methods as follows.

## Variograms

### *Direct variogram*

The direct variogram generally is computed from equation (4),

$$\gamma(h) = \frac{1}{2N} \sum_{i=1}^N [Z(x_i) - Z(x_{i+h})]^2 \quad (4)$$

where  $x_i$  is a data location,  $h$  is a vector of distance,  $Z(x_i)$  is the data value of one kind of attribute at location  $x_i$ ,  $N$  is the number of data pairs for a certain distance and direction of  $h$  units. The equation is used for determining the spatial autocorrelation of the univariate variable.

### *Cross variograms*

A typical cross variogram is calculated as in equation 5, and is applied for the joint spatial variability between two kinds of spatial variables. It is defined as half of the average product of the lag distance relative to the two variables  $Z$  and  $Y$ .

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} \{[Z(x_i) - Z(x_{i+h})] * [Y(x_i) - Y(x_{i+h})]\} \quad (5)$$

When the direct and cross variogram models are fitted, they also can guarantee that the fitted models follow the linear model of coregionalisation (Goovaerts, 1997). This ensures the cross covariance matrices are always positive. Calculations and visualizations of directional variograms, variogram clouds, and identification through interactive examination in the variogram cloud were finished using the gstat package (Pebesma, 2004).

## Kriging

### *Ordinary kriging and universal kriging*

Ordinary kriging (OK) is identical to multiple linear regression with a couple of important differences. The ordinary kriging model is as in equation 6.  $Z(s_0)$  is the value to be interpolated at location  $s_0$ ,  $z(s_i)$  are the sampled values at their locations, and  $\lambda_i$  are the weights to be assigned to each sampled value. Universal kriging is applied when a trend exists. Universal kriging is often fitted using a polynomial equation, which is similar to the equation 6 to analyze the trend across the study area.

$$Z(s_0) = \sum_{i=1}^n \lambda_i z(s_i) \quad (6)$$

### *Cokriging*

For forest applications, a few studies using remote sensing data have been conducted using the geostatistical approach. Dungan et al. (1994) and Dungan (1998) applied co-kriging and a stochastic simulation method for forest management using synthetic remote sensing datasets.

Co-kriging (CoK) is an extension of kriging, and is a method for estimating one or more variables of interest using data from several variables by incorporating not only spatial correlation but also inter-variable correlation. Co-kriging is a very versatile and rigorous statistical technique for spatial point estimation when both primary and auxiliary attributes are available. It is defined as in equation 7. If each component of  $z(s_0)$  satisfies the intrinsic hypothesis (Journel and Huijbregts, 1978), then equation 5 is unbiased if

$$Z(s_0) = \sum_{j=1}^n z(s_j) \Lambda_{j\bullet} \quad (7)$$

$$\sum_{j=1}^n \Lambda_{j\bullet} = I \quad (8)$$

where  $I$  is an identity matrix  $= [1, 0, \dots, 0]^T$  and  $T$  indicates a transpose, and  $\Lambda_{j\bullet}$  are the weights associated with prediction. Equation 7 is

$$\sum_{\phi=1}^v \Gamma(s_i, s_j) + \Psi = \Gamma(s_i, s_0) \quad i = 1, \dots, n \quad (9)$$

where  $z(s_j)$  is the vector  $z_1(s_j) \dots z_m(s_j)$ .  $\Gamma(s_i, s_j)$  and  $\Gamma(s_i, s_0)$  are the cross variograms, and  $\Psi$  is the Lagrange Multiplier for  $i$  from 1 to  $n$ .

According to the sample relations between the primary variable and the auxiliary variables, CoK could be described in several ways as follows. The most efficient way to predict the primary variable is to use the auxiliary variables to cokrig it into dense grid locations. This is named heterotopic cokriging (Wackernagel, 1994). Isotopic cokriging requires that data on both the target variable and co-variables be measured at all sample locations. A variant of both is generalized cokriging (Myers, 1982) that involves simultaneous prediction of all the correlated variables into more dense locations. The complete case is the case where the covariates and the primary variable do not share any common locations. A more general type applied using remote sensing data is collocated cokriging, where covariates are available at all interpolation locations, although the primary variable is available at only a few locations. When CoK is compared to univariate kriging, no new concept is added, but there is heavier notation associated with having several variables (Goovaerts, 1997).

### *Regression kriging*

Regression kriging (RK) is a hybrid method that combines either a simple or multiple-linear regression model (or a variant of the generalized linear model (GLM) and regression trees) with

kriging (Odeh et al., 1995; Goovaerts, 1997). In the process of RK, kriging with uncertainty introduces the regression residuals (i.e., the model uncertainty) into the kriging system, which is then applied directly to predict the primary variable. The predictions are combined from two parts; one is the estimation obtained by regressing the primary variable on the auxiliary variables; the second part is the residual estimated from the ordinary kriging. Regression kriging is estimated as follows:

$$\hat{Z}_{rk}(s_0) = \hat{m}(s_0) + \hat{\ell}(s_0) \quad (10)$$

$$\hat{Z}_{rk}(s_0) = \sum_{k=0}^v \hat{\beta}_k * q_k(s_0) + \sum_{i=1}^n \omega_i(s_0) * \ell(s_i) \quad q_0(s_0) = 1, \quad i = 1, \dots, n \quad (11)$$

where  $\hat{\beta}_k$  are trend model coefficients, optimally estimated using generalized least squares;

$\omega_i$  are weights determined by the semivariance function, and  $\ell$  are the regression residuals. In

the gstat package, univariate kriging and multivariable kriging are applied for pine basal area prediction (Pebesma, 2004, 2005).

### Model Evaluation

In this study, different geostatistical models are developed and applied for pine basal area prediction. There are always discrepancies between true and predicted values. It is necessary to validate the models and check which is more efficient. For this geostatistical approach, two methods for assessing models are applied. One method is cross validation, which is used to validate whether the model fits the training data. The second method is validation based on random samples outside of the training data set. We developed 200 random points to check which model is more efficient in spatial predictions of pine basal area.

There are many different measures for checking discrepancies and each has its advantages and weaknesses. Details about forecast evaluation were discussed by Murphy and Katz (1985). Typically, four criteria, standard deviation (SD), bias error (BE), root mean square error (RMSE), and mean-absolute error (MAE) are used to directly compare forecast and observation. Standard deviation is the measure of dispersion from the mean of a particular parameter as illustrated by equation 12.

$$SD = \left[ \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X})^2 \right]^{1/2} \quad (12)$$

where  $N$  is the size of the sample,  $X_n$  is the sample values and  $\bar{X}$  is the mean of the sample. The bigger the SD, the larger the dispersion of the estimations is from the mean. For the error term, SD typically is used to measure the extent that forecast error differs from the mean. In this study, the SD of errors (SDe) is computed to analyze dispersions of errors across the whole study area.

Bias error is used to measure whether the model under-forecasts or over-forecasts a parameter

and is defined in the equation:

$$BE(X) = \frac{1}{N} \sum_{n=1}^N (X_f - X_o) \quad (13)$$

where  $N$  is the total number of comparisons,  $X_f$  is the forecast value, and  $X_o$  is the observed value. A positive BE indicates a tendency to overpredict while a negative BE implies under predictions.

The square-root of the individual squared differences between forecast and observation is root mean square error (RMSE). It is defined in equation 14.

$$RMSE(X) = \left[ \frac{1}{N} \sum_{n=1}^N (X_f - X_o)^2 \right]^{1/2} \quad (14)$$

Mean-absolute error is the average of the absolute value of the difference between forecast and observation as defined in equation 15. Mean-absolute error values near or equal to 0 indicate perfect or almost perfect forecasts. This measure is not as heavily weighted towards large differences in forecast comparisons as with RMSE.

$$MAE = \frac{1}{N} \sum_{n=1}^N |X_f - X_o| \quad (15)$$

## Results

### Correlations between Pine Basal Area and Predictors

Predictors are grouped into four groups: a 432 band combination; a 543 band combination; a three-PCs combination; and an NDVI image. The general Pearson correlation coefficients were calculated and summarized in Table 1. Considering the absolute values of these coefficients, for the correlations between pine basal area and different independent variables, PC2 has the highest correlation, the second one is NDVI, the third one is band5, and then, band3, PC1, band2, band4, and PC3.

Since different combinations of predictors were used, the Pearson partial correlation coefficients were calculated and tested in the combinations of bands and PCs in order to better understand the associations between pine basal area and the predictors (Table 2). In the 432 band combination, band 3 and band 4 have similar degree correlations but in different directions; one is positive, and another is negative; band 2 is little correlated with the pine basal area, and the coefficient is not significantly different from 0. In the 543 band combination, band 4 and band 5 have similar correlations with pine basal area. However, band 4 is positively correlated, and band 5 is negatively correlated. Band 3 is little correlated with pine basal area. PC2 is highly correlated with pine basal area. The coefficient of PC1 is much smaller. The correlation between PC3 and pine basal area might be little, since its P value is around the boundary of 0.05 and therefore

statistically means the partial correlation coefficient is close to 0.

	PINEBA	Band2	Band3	Band4	Band5	NDVI	PC1	PC2	PC3
PINEBA	1								
Band2	-0.3917	1							
Band3	-0.5417	0.8364	1						
Band4	0.3456	0.1221	-0.0724	1					
Band5	-0.5964	0.8067	0.9312	-0.0488	1				
NDVI	0.6365	-0.6517	-0.8794	0.5202	-0.8187	1			
PC1	-0.5195	0.8623	0.9384	0.1197	0.9766	-0.7417	1		
PC2	-0.6520	0.7129	0.9022	-0.3508	0.9448	-0.9269	0	1	
PC3	-0.0315	-0.1852	-0.1287	-0.7872	-0.3163	-0.2450	0	0	1

**Table 1.** Pearson correlation matrix for the variables analyzed for pine basal area estimation. The second column indicates the correlations between pine basal area (PINEBA) and predictors of four Landsat ETM+ bands, three PCs, and NDVI. The values from column 3 to column 10 indicate some independent variables also are highly correlated, and Pearson partial correlation need conducting to understand the real contributions of independent variables to the estimations of PINEBA.

	234 band combination			345 band combination		
	Band2	Band3	Band4	Band3	Band4	Band5
$r_{xy}$	0.0129	-0.3350	0.3436	0.0828	0.3997	-0.3429
$P$ value	0.4976	<.0001	<.0001	<.0001	<.0001	<.0001

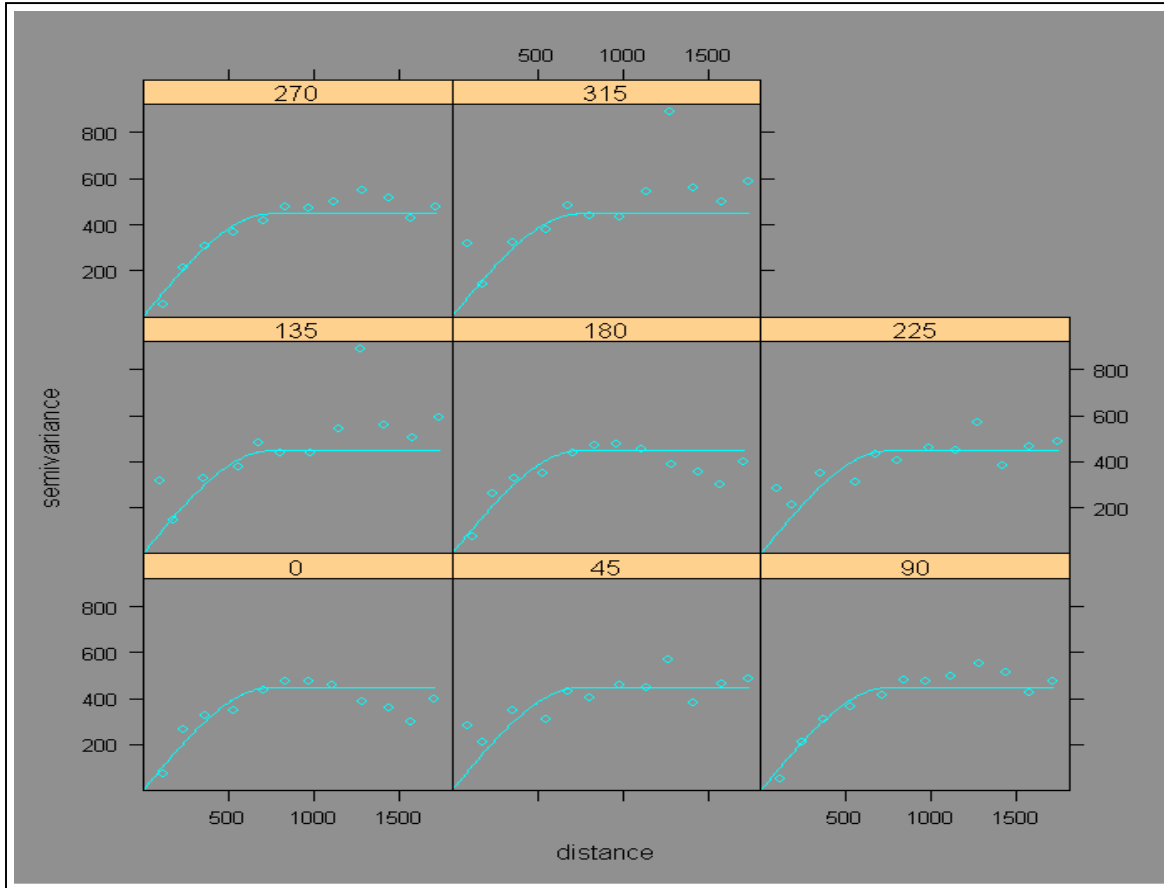
**Table 2.** Partial correlations eliminating effects due to correlations from other predictors. Pearson partial correlation coefficients ( $r$ ) were calculated by eliminating effects due to the correlations between pine basal area and the other two variables in the band combinations. For example, the correlation between pine basal area and band2 is 0.0129 when the contributions from band3 and band4 were removed. This correlation is not significantly different from 0, since the  $P$  value is 0.4976 which is much bigger than 0.05.

### Variograms and Spatial Dependence

Variograms were used to spatially analyze the surface properties of pine basal area. Based on the variogram cloud, the empirical semivariogram model was created. The different types of semivariogram models used to fit the points include exponential, Gaussian, circular, spherical, tetraspherical, pentaspherical, Hole effect, K-Bessel, and J-Bessel models. The spherical model had the best fits and was selected as the theoretical model applied for spatial predictions. The fit of the spherical model has a nugget of 5, a partial sill of 450, and a range of 750. Also, there was no obvious trend existing among the pine basal area across the study area.

The characteristics of the semivariogram also may be affected by the directions, which result from a special geographic phenomenon. For example, a certain kind of species exists and crosses the area in a certain direction. It is therefore necessary to check anisotropy. Semivariogram analyses at directions 0, 45, 90, 135, 180, 225, 270, and 315 were conducted and indicated similar

spatial dependence at these eight directions (Figure 4). It is not necessary to analyze anisotropic effects in spatial predictions.



**Figure 4.** Semivariogram modeling effects of eight different directions

### Assessment of Pine Basal Area Estimation

We first applied Univariate kriging (i.e., OK and UK) to estimate the pine basal area using 2822 ground inventory points. The UK was used to check whether it is effective compared to the OK, though there was no obvious trend of pine basal area existing across the study area. Four types of co-kriging were applied using the 432 band combination, the 543 band combination, NDVI, and PCs as the auxiliary data. At last, four groups of regression kriging were conducted using the 432 band combination, the 543 band combination, NDVI, and PCs as predictors.

The results were evaluated using cross validation (Table 3). Bias errors using the kriging methods indicated the values of BE were close to 0, and almost unbiased estimations of pine basal area were obtained. For RMSE, there was not much difference between OK, UK, and the four kinds of co-kriging. However, the RMSEs of the estimations using regression kriging were much smaller than those from OK, UK, and co-kriging. In order to further assess these geostatistical approaches, 200 random sample points outside of the training dataset were selected and used to compare these kriging methods (Table 4). The regression kriging methods had the smallest BE, MAE, RMSE, and SDe, which indicated that regression kriging was more efficient than other

kriging methods. Pine basal area predictions based on RK resulted in the prediction BE of 27.9~31.5% of the mean (13.99 m<sup>2</sup>/ha), the prediction MAE of 39.3~42.1% of the mean, the prediction RMSE of 63.5~68.6% of the mean, and the prediction SDe of 59.3~62.1% of the mean using the 200 random points outside the training datasets.

	OK	UK	CoK432	CoK543	CoKndvi	CoKPCs	RK432	RK543	RKndvi	RKPCs
BE	-0.076	-0.078	-0.099	-0.100	-0.095	-0.095	-0.078	-0.067	-0.066	-0.070
RMSE	11.310	11.290	10.970	11.000	11.010	11.020	7.020	7.000	7.220	6.890

**Table 3.** Model evaluation using cross validation. Ordinary kriging (OK), universal kriging (UK), Co-kriging (Cok), and regression kriging (RK) are used to predict basal area. CoK432 means using the 432 band combination as predictors to krig the basal area, likewise CoK543, CoKndvi, CoKPCs, RK432, RK543, RKndvi, and RKPCs; bias error (BE) and root mean square error (RMSE) are used to measure the discrepancy between observations and predictions.

	OK	UK	CoK432	CoK543	CoKndvi	CoKPCs	RK432	RK543	RKndvi	RKPCs
BE	10.120	10.130	4.990	4.980	4.760	4.660	4.460	4.010	4.432	3.964
RMSE	13.320	13.390	10.550	10.320	10.560	10.010	9.655	8.980	9.601	9.161
SDe	8.660	8.770	9.300	9.260	9.310	9.210	8.583	8.601	8.700	8.280
MAE	10.330	10.470	6.310	6.290	6.310	6.280	5.929	5.502	5.900	5.727

**Table 4.** Model and forecast evaluation using validation based on random samples. Stand deviation of errors (SDe), mean-absolute errors (MAE), BE, and RMSE are used to measure the discrepancy between observations and predictions. Other notations are the same as Table 3.

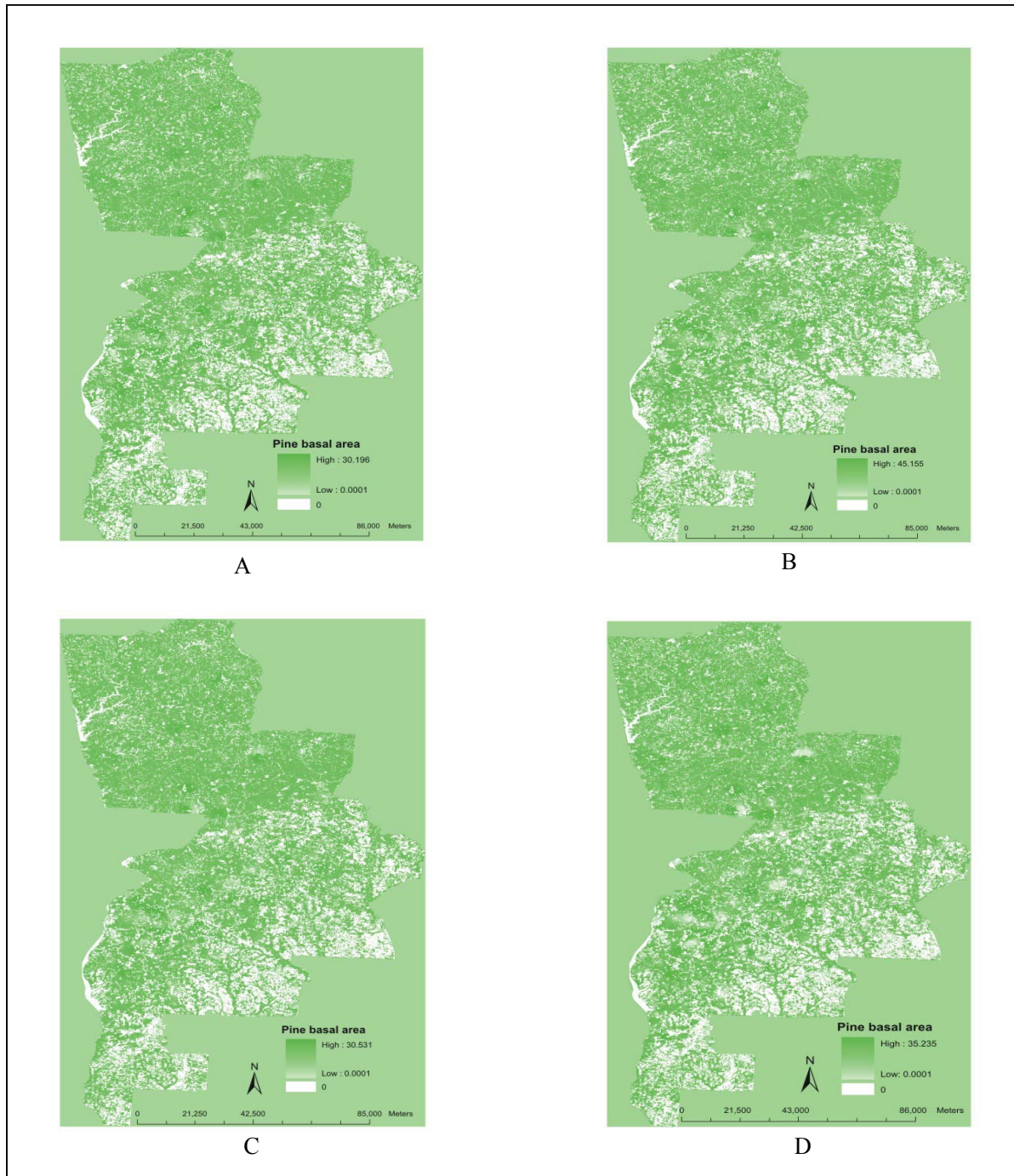
### Pine Basal Area Mapping Using Regression Kriging

Regression kriging was the best approach to predict pine basal area using Landsat ETM+ images. The results of regression kriging were transformed and used to map the pine basal area at these 20 counties using ERDAS Imagine<sup>®</sup> and ArcGIS9.1. The pine basal areas were mapped based on the four types of regression kriging using the 432 band combination, the 543 band combination, NDVI, and PCs as predictors (Figure 5). The standard deviations of errors were also mapped in order to indicate the spatial characteristics of errors of pine basal area estimations (Figure 6). Using the 432 band combination, we obtained relatively smaller standard errors of predicted pine basal area across the whole study area than those from the 543 band combination, three PCs and NDVI.

## Discussions

Challenges still exist in the field of large area forest inventory using remotely sensed data (Tokola et al. 1996, Trotter et al. 1997, Holmström & Fransson 2003). Spatial diversity of forest stands and landscape makes the spatial prediction of forest parameters a major challenge, although the remote sensing data are highly associated with forest features. For example, forest stands may

have very similar values of biomass/carbon but have different spectral characteristics because of differences in species. The differences of spectral characteristics between plantations and natural stands might exist although the stands have many of the same characteristics, such as same species, same age, and same density. These differences will add noise when the prediction models are fitted based on the associations between remotely sensed data and ground-inventoried data.



**Figure 6.** Pine basal area estimations using regression kriging with Landsat ETM+ data as predictors. A, using bands 2, 3, and 4 as predictors; B, using bands 3, 4, and 5 as predictors; C, using NDVI as predictors; D, using three PCs as predictors.



**Figure 7.** Mapping standard errors of spatial predictions of pine basal area from regression kriging. A, standard errors using bands 2, 3, 4 as predictors; B, standard errors using bands 3, 4, 5 as predictors; C, standard errors using the band of NDVI as predictors; D, standard errors using PCs as predictors.

Berterretche et al. (2005), Tuominen et al.(2003), and Zhang et al (2004) applied geostatistical models to estimate forest variables, leaf area index, and classify forest lands based on remote sensing data. Compared to their studies, multivariable kriging (i.e., RK in this study) is robust and results in relative smaller errors. Multivariable kriging can be applied for almost all kinds of forest parameters. Also, either numerical or categorical data can be used in the process of

kriging, i.e., any kind of variable can be used as auxiliary data or predictors.

Remote sensing data and ground inventory data are collected and stored in different data structures. The discrepancy between remotely sensed data and ground sampling data might be the source of big errors in forest predictions (Tokola et al. 1996, Gilbert & Lowell 1997). The ground inventory data are usually collected at the forest plot level or forest stand level. The plot size may be from several meters to 10 or 20 meters. The stand size may be from 10 meters to dozens of meters, and the stands are assumed to be homogenous. Therefore, some ground data may be finer than remote sensing data in spatial resolution, but generally, remote sensing data has a finer spatial resolution than ground inventory data. This may result in some noise added to the geostatistical modeling and cause bias errors and mean-absolute errors.

## Conclusions

The systematic approach of geostatistical prediction and mapping developed by integrating remote sensing, ground inventory, and GPS data provides a new way to spatially estimate forest parameters using remotely sensed data. It has many applications in forest or natural resource management. Forest metrics, such as stand density, dominant height, species, stand age, forest health conditions, the probability of forest fire, biomass, carbon, and so on, can be incorporated in the model. They can be estimated spatially at finer spatial resolution using remotely sensed data with higher spatial resolution.

Providing finer spatial information is essential for large area timber, biomass, and carbon budget management and planning. Kriging is an optimum method for spatial interpolation. Regression kriging is the most powerful one among the different kriging methods in this research. It was used to predict the pine basal area at 30m for these 20 counties (about 35000 km<sup>2</sup>) using only 2822 ground inventory data points. Four groups of independent variables are used in RK. The 543 band combination resulted in the smallest BE, RMSE, MAE, and had a relatively smaller SDe. Therefore, Compared with OK, UK and CoK using different auxiliary data, RK resulted in the smallest BE, RMSE, SDe, and MAE. Regression kriging using the NDVI as the predictor could be the best method for pine basal area predictions if computation is considered for large area basal area inventory, since it has only one independent variable and can significantly reduce computation time as compared with other band combinations. For other forest parameters, such as dominant height, timber volume, or biomass/carbon, other band combinations, such as PCs or NDVI need to be applied again to check which will result in better estimations.

More research is needed to demonstrate whether the geostatistical approach is more or less efficient than other methods used for large area forest inventory, such as K nearest neighbor methods using remotely sensed data. This will further demonstrate the efficiency and usefulness of this geostatistical approach for forest inventory and management.

## References

- Ardö, J. 1992. Volume quantification of coniferous forest compartments using spectral radiance recorded by Landsat Thematic Mapper. *International Journal of Remote Sensing* 13 : 1779-1786.
- Atkinson, P.M. and P. Lewis. 2000. Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences* 26: 361-371.
- Atkinson, P.M., R. Webster, and P.J. Curran. 1994. Cokriging with airborne MSS imagery. *Remote Sensing of Environment* 50: 335-345.
- Barnsley, M.J., 1999. Digital remote sensing data and their characteristics, *Geographical Information Systems: Principles, Techniques, Applications, and Management* (Second Edition).
- Berterretche, M., A.T. Hudak, W. B. Cohen, T. K. Maierperger, S.T. Gower, and J. Dungan. 2005. Comparison of regression and geostatistical methods for mapping Leaf Area Index (LAI) with Landsat ETM+ data over a boreal forest. *Remote Sensing of Environment* 96: 49-61.
- Braswell, B.H., D.S. Schimel, E. Linder, and B. III. Moore. 1997. The response of global terrestrial ecosystems to interannual temperature variability. *Science* 278: 870-872.
- Chica-Olmo, M. and F. Abarca-Hernandez. 2000. Computing geostatistical image texture for remotely sensed data classification. *Computer & Geosciences* 26: 373-383.
- Cressie, N.A.C. 1993. Statistics for spatial data. John Wiley & Sons, New York.
- Curran, P.J. and P.M. Atkinson. 1998. Geostatistics and remote sensing. *Progress in Physical Geography* 22: 61-78.
- Curran, P.J., 1988. The semivariogram in remote sensing: an introduction. *Remote Sensing of Environment* 24: 493-507.
- Dungan, J.L., D.L. Peterson and P.J. Curran. 1994. Alternative approaches for mapping vegetation quantities using ground and image data. In: Michener, W., Stafford, S. & Brunt, J. (eds.). *Environmental information management and analysis: ecosystem to global scales*. Taylor and Francis, London, UK. pp. 237-261.
- Dungan, J.L. 1998. Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing* 19: 267-285.
- Short, N.M. 1999. The remote sensing tutorial.  
[Http://www.fas.org/irp/imint/docs/rst/Sect3/Sect3\\_1.html](http://www.fas.org/irp/imint/docs/rst/Sect3/Sect3_1.html). Accessed May 1, 2005.
- Foody, G.M. and D.S. Boyd. 1999. Fuzzy mapping of tropical land cover along an environmental gradient from remotely sensed data with an artificial neural network. *Journal of Geographical Systems* 1: 23-35
- Foody, G.M. 2000. Mapping land cover from remotely sensed data with a softened feed forward neural network classification. *Journal of Intelligent and Robotic Systems* 29: 433-449.
- Franco-Lopez, H., A.R. Ek, and M.E. Bauer. 2001. Estimation and mapping of forest stand density, volume and cover type using k-nearest neighbors method. *Remote Sensing of Environment* 77: 251-274.

- Gilbert, B. and K. Lowell. 1997. Forest attributes and spatial autocorrelation and interpolation: effects of alternative sampling schemata in the boreal forest. *Landscape and Urban Planning* 37: 235-244.
- Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York.
- Goward, S.N., C.J. Tucker. and D.G. Dye. 1985. North American vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer. *Vegetatio* 64: 3-14.
- Holmström, H. and J.E.S. Fransson. 2003. Combining remotely sensed optical and radar data in kNN estimation of forest variables. *Forest Science* 10: 409-418.
- Lark, R.M., 1996. Geostatistical description of texture on an aerial photograph for discriminating classes of land cover. *International Journal of Remote Sensing* 17: 2115-2133.
- Matheron, G. 1965. Les Variable Regionalisees et leur Estimation. Masson, Paris.
- Moeur, M. and A.R. Stage. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. *Forest Science* 41(2): 337-359.
- Murphy, A.H. and R.W. Katz. 1985. Probability, statistics, and decision making in the atmospheric sciences. Boulder, Colo: Westview Press.
- Myers, D.E. 1982. Matrix formulation of cokriging. *Mathematic Geology* 14: 249-157.
- Myers, B., 1996. [Http://www.ai-geostats.org/archives/1996/AI-11-96/0039.html](http://www.ai-geostats.org/archives/1996/AI-11-96/0039.html). Accessed May 1, 2006.
- Myneni R.B., C.D. Keeling, C.J. Tucker, G. Asrar, and R.R. Nemani. 1997. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature* 386: 698-702.
- Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma* 67: 215-226.
- Odeh, I.O.A. and A.B. McBratnery. 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. *Geoderma* 97: 237-245.
- Pebesma, E. J. 2004. Multivariable Geostatistics in S: the Gstat Package. *Computer & Geosciences* 30: 683-691.
- Pebesma, E.J. 2005. The Gstat Package. <http://cran.r-project.org/doc/packages/gstat.pdf>. Accessed November 25, 2005.
- Tatem, A.J., Lewis, H.G., Atkinson, P.M. & Nixon, M.S. 2001. Land cover mapping from remotely sensed images at the sub-pixel scale using a Hopfield neural network. *IEEE Transactions on Geoscience and Remote Sensing* 39: 781-796.
- Tokola, T., Pitkänen, S., Partinen, S. & Muinonen, E. 1996. Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. *International Journal of Remote Sensing* 17: 2333-2351.
- Tomppo, E. 1991. Satellite imagery-based national inventory of Finland. *International Archives of Photogrammetry and Remote Sensing* 28: 7-1, 419-424.
- Trotter, C.M., J.R. Dymond, and C.J. Goulding. 1997. Estimation of timber volume in a coniferous plantation forest using Landsat TM. *International Journal of Remote Sensing* 18: 2209-2223.

- Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8: 127-150.
- Tuominen, S., S. Fish, and S. Poso. 2003. Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventories. *Canadian Journal of Forest Research* 33: 623-634
- Warren, B.C., T.A. Spies, and G.A. Bradshaw. 1990. Semivariograms of digital imagery for analysis of conifer canopy structure. *Remote Sensing of Environment* 34: 167-178.
- Wackernagel, H., 1994. Multivariate spatial statistics. *Geoderma* 62: 83-92.
- Zhang, C., S.E. Franklin, and M.A. Wulder. 2004. Geostatistical and texture analysis of aribo-re-acquired images used in forest classification. *International Journal of Remote Sensing* 25: 859-865.