

# USING GEOSPATIAL METHODS FOR DERIVATION OF FINE SPATIAL RESOLUTION FOREST INVENTORY FROM GROUND INVENTORY DATA AND LANDSAT IMAGERY

Qingmin Meng, Chris J. Cieszewski, R. Lowe  
Warnell School of Forest Resources, University of Georgia, Athens, GA 30602 USA

## ABSTRACT

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 geospatial 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 Landsat ETM 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.

**KEYWORDS:** Kriging, regression kriging, GIS, remote sensing.

## 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

---

*In* Prisley, S., P. Bettinger, I-K. Hung, and J. Kushla, eds. 2006. Proceedings of the 5<sup>th</sup> Southern Forestry and Natural Resources GIS Conference, June 12-14, 2006, Asheville, NC. Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA.

boundaries in the field with an accuracy rate of  $\pm 5$  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.

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 paper. 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 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.

## METHODS

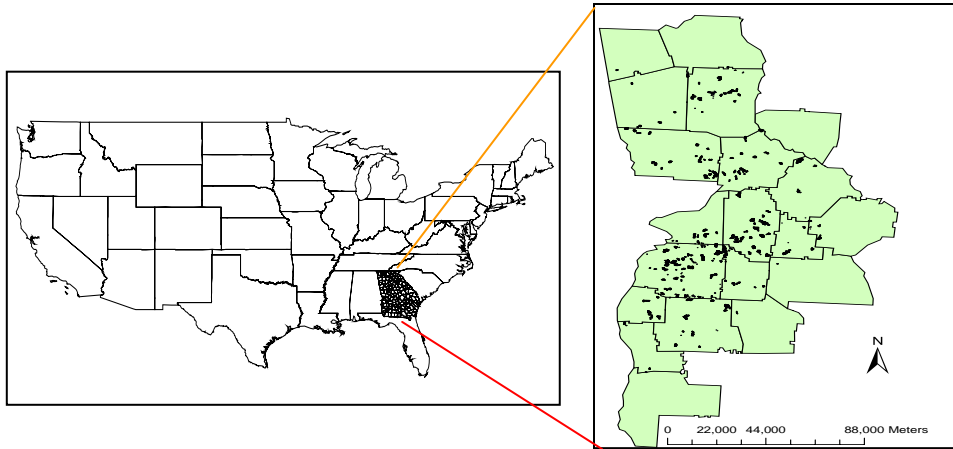
### Data Sources

Ground data covering 20 counties in west Georgia were inventoried in 1999 (Figure 1). 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. There were 2822 ground records used in this study with a mean basal area of 13.99 m<sup>2</sup>/ha and a range from 0.038 to 29.84 m<sup>2</sup>/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.

### 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. These two Landsat images covering this study area were masked after the geometric corrections. This resulted in a 4449 pixel by 9010 row 6-band (i.e., 1, 2, 3, 4, 5, and 7) image for analysis.



**Figure 1. 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.**

#### Band Combinations

Band 1 of Landsat images contributes little for vegetation analysis. The differences of reflectance increase from 0.5 to 0.8  $\mu m$  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 were used, and 4-3-2 and 5-4-3 band combinations were applied to estimate pine basal area.

#### Principal Components

Principal component analysis (PCA) is the most frequently used technique for remote sensing data reduction. 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. 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 vegetation index (NDVI) was used for pine basal area estimation. NDVI was 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.

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

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 zero while water bodies have the opposite trend to vegetation and the index is negative. It has been extensively applied as a proxy for leaf area index (Tucker, 1979), vegetation biomass, 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).

### Correlation analysis

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. The Pearson partial correlation is used to indicate 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 estimating the semivariogram. 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. 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.

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.

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  $Y$ .

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

### Ordinary kriging and universal kriging

Ordinary kriging (OK) is similar to multiple linear regression but has 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, and a popular way of universal kriging is to fit a polynomial equation similar as 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. It is defined as in equation 7.

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

If each component of  $z(s_0)$  satisfies the intrinsic hypothesis, then equation 5 is unbiased if

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

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  $n$ .

### 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, 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. The gstat package is used for univariate kriging, CoK and RK (Pebesma, 2004, 2005).

### Model Evaluation

Four criteria including standard deviation (SD), bias error (BE), root mean square error (RMSE), and mean-absolute error (MAE) are used to directly compare forecast and observation.

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

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 estimates are 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 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. Mean-absolute error is calculated as equation 15.

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

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

## RESULTS & DISCUSSION

### Correlation Analysis

Predictors are grouped into four groups: a 4-3-2 band combination; a 5-4-3 band combination; a three-PCs combination; and an NDVI image. 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 band 5, and then, band 3, PC1, band 2, band 4, 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. In the 4-3-2 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 zero. In the 5-4-3 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 zero.

### Variograms

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. Semivariogram analyses at directions 0, 45, 90, 135, 180, 225, 270, and 315 were conducted and indicated similar spatial dependence at these eight directions. It is not necessary to analyze anisotropic effects in spatial predictions.

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 4-3-2 band combination, the 5-4-3 band combination, NDVI, and PCs as the auxiliary data. At last, four groups of regression kriging were conducted using the 4-3-2 band combination, the 5-4-3 band combination, NDVI, and PCs as predictors.

The results were evaluated using cross validation (Table 1). Bias errors using the kriging methods indicated the values of BE were close to zero, 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, validation based on 200 random sample points outside of the training dataset were selected and used to compare these kriging methods (Table 2). 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.

**Table 1. 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 4-3-2 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	-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 2. 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 1.**

	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

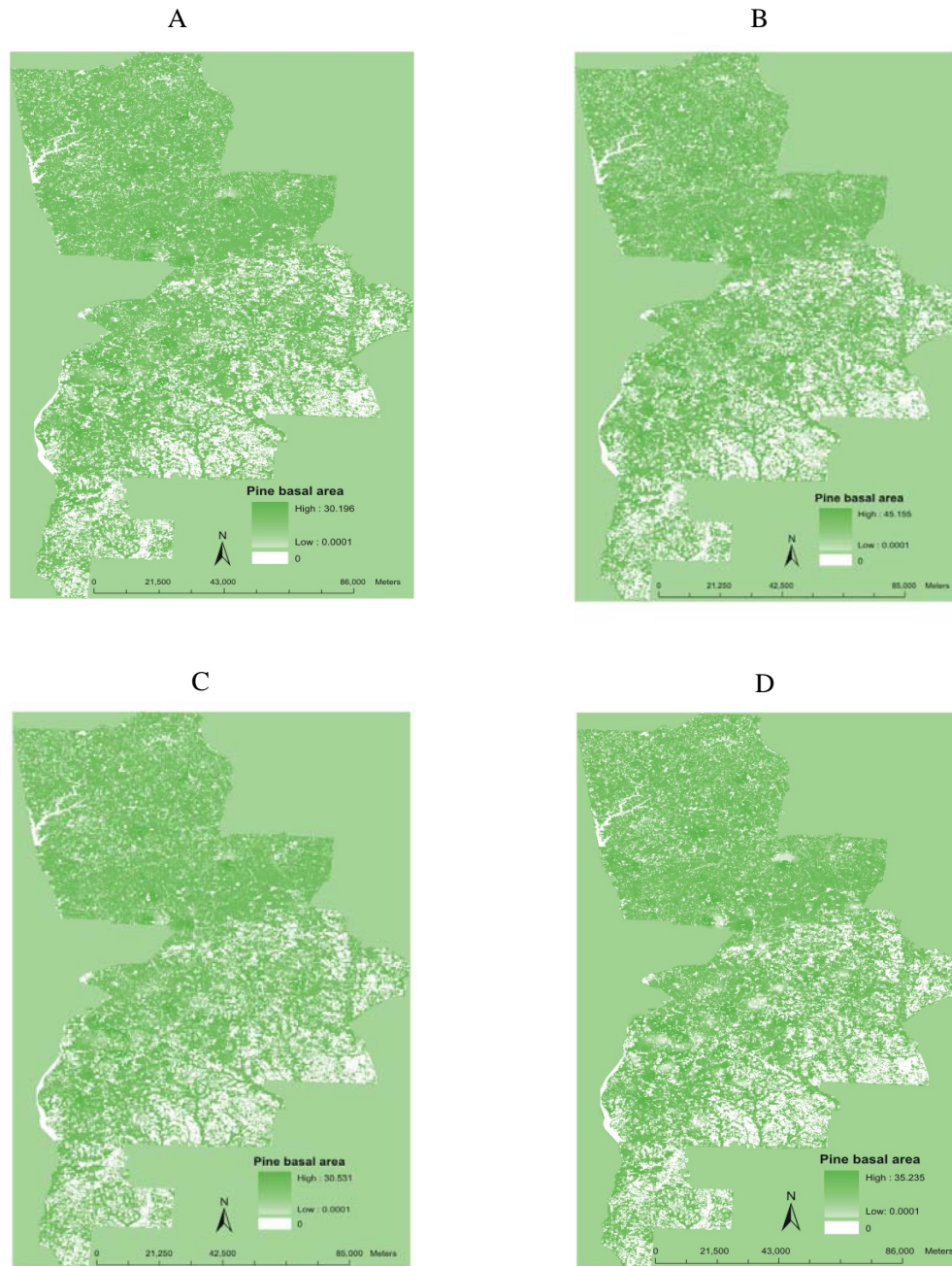
### Pine Basal Area Mapping Using Regression Kriging

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 2).

## CONCLUSIONS

The systematic approach of geostatistical prediction and mapping developed by integrating remote sensing, ground inventory and GPS data in this study 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; RK using the 5-4-3 band combination is the best method for pine basal area predictions. 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 best estimations.



**Figure 2. 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.**

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

- Atkinson, P.M. and P. Lewis. 2000. Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences*. 26: 361-371.
- Braswell, B.H., D.S. Schimel, E. Linder, and B. Moore III. 1997. The response of global terrestrial ecosystems to interannual temperature variability. *Science*. 278: 870-872.
- 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: *Environmental information management and analysis: ecosystem to global scales*, 237-261, Michener, W., S. Stafford, & J. Brunt (eds.). London, UK: Taylor and Francis..
- Dungan, J.L. 1998. Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing*. 19: 267-285.
- Goovaerts, P. 1997. *Geostatistics for natural resources evaluation*. New York, NY: Oxford University Press.
- 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.
- Matheron, G. 1965. *Les Variable Regionalisees et leur Estimation*. Masson, Paris.
- 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.
- 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.

Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*. 8: 127-150.

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.