Practical Considerations When Using Perturbed Forest Inventory Plot Locations To Develop Spatial Models: A Case Study

John W. Coulston[1], **Gregory A. Reams**[2], **Ronald E. McRoberts**[3], **and William D. Smith**[4]

Abstract.—U.S. Department of Agriculture Forest Service Forest Inventory and Analysis plot information is used in many capacities including timber inventories, forest health assessments, and environmental risk analyses. With few exceptions, actual plot locations cannot be revealed to the general public. The public does, however, have access to perturbed plot coordinates. The influence of perturbed plot coordinates on the development of spatial models is unknown. We examined the influence by comparing the accuracies of two spatial models for predicting forest biomass, ordinary kriging and residual kriging. We developed each model using the actual coordinates and 10 independent perturbations of the actual coordinates. We tested for differences in accuracy using analysis of variance. No statistically significant difference in accuracy was found. The results represent only a small portion of the possible outcomes, however. We suggest a simulation study to examine the spatial range of influence that plot coordinate perturbation has on model accuracy.

Introduction

The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture (USDA) Forest Service collects data on tree and forest attributes using a quasi-systematic sample. These data are used for many purposes including timber inventories, forest health assessments, and risk assessments. Because of privacy issues, actual plot locations cannot be revealed to scientists outside the FIA program or the general public. FIA has implemented several methods for perturbing plot locations to protect plot integrity and ensure landowner privacy. Although the perturbed plot locations are available to the public, the effects of the perturbations on the accuracy of spatial models are unknown.

Before 2002, FIA field plot locations perturbed within 1.6 km of the actual locations were available to the general public. Although currently no national standard exists for perturbing plot coordinates, guidelines that may be satisfied at the regional level using different techniques are available. One method currently used is to randomly shift plot locations and swap data among plots. In this article, we use the term "perturbed" to denote both the random shift in plot location and the swapping of plot attributes. Plot perturbation influences the spatial characteristics of the data and, therefore, can influence the accuracy of spatial models.

Spatial models and FIA data are widely used in environmental assessments. For example, Morin *et al*. (2003) used FIA field plot data, perturbed plot locations, and median indicator kriging to interpolate a surface of percent forest basal area of species susceptible to *Phytophthora ramorum* (a fungus-like organism that causes Sudden Oak Death). This interpolated surface was then intersected with other spatial data and used to assess the potential susceptibility of Eastern forests to *Phytophthora ramorum*. Coulston *et al.* (2003) used ordinary kriging to predict potential ozone injury at FIA phase 3 (formerly forest health monitoring) plot locations and assess ozone injury risk to ozone-sensitive Northeastern tree species. This analysis was conducted using the centers of the sampling hexagons (White *et al.* 1992) as plot locations rather than the actual plot locations.

[1] Research Assistant Professor, North Carolina State University, Department of Forestry, Box 8008, Raleigh, NC 27695. E-mail: jcoulston@fs.fed.us.
[2] National Program Leader, U.S. Department of Agriculture (USDA), Forest Service, National Headquarters, 1601 North Kent Street, Arlington, VA 22209. E-mail: greams@fs.fed.us.
[3] Mathematical Statistician, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108. E-mail: rmcroberts@fs.fed.us.
[4] Assessment Coordinator, USDA Forest Service, Southern Research Station, 3041 Cornwallis Road, Research Triangle Park, NC 27709. E-mail: bdsmith@fs.fed.us.

Spatial models generally rely on the relationship among observations by distance and direction (e.g., kriging). More complicated spatial models may further rely on ancillary data that are intersected with plot data (e.g., residual kriging). The objective of this study was to examine the influence of FIA plot coordinate perturbations on the accuracy of two spatial models for predicting forest biomass. The first model was ordinary kriging of forest biomass, and the second model was residual kriging in which forest biomass was predicted using percent forest and leaf area index (LAI) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data.

Methods

Plot-level estimates of percent forest land use and forest biomass were obtained for 3,914 FIA plots in Minnesota. The plot locations were randomly perturbed 10 different times in accordance with the procedures used by the FIA program of the North Central Research Station, USDA Forest Service. Perturbing plot locations entails randomly shifting the x and y coordinates of the actual locations for all plots, and swapping plot attributes (e.g., tree volume m^3ha^{-1}) entails exchanging coordinates among a proportion of plots. These manipulations are usually done within a county, and plot attributes can be swapped only if the plots are sufficiently similar (e.g., same forest type). The data set consisting of the percent forest land use and forest biomass estimates and the actual plot locations is denoted REAL, while the 10 data sets consisting of the estimates and the perturbed plot locations are denoted REPS. Before the spatial models were developed, we randomly extracted 180 plots (approximately 5 percent) from the data set. Average plot biomass for these extracted plots was 26.6 tons/acre, and the standard deviation was 19.3 tons/acre. The biomass models were then developed without these plots, model predictions were made for the 180 plots, and the accuracy of the models with and without plot coordinate perturbations (i.e., for the REAL and REPS data sets) was compared.

Because ordinary kriging is a central technique in this analysis, we provide a brief overview. (For more details, see Cressie 1993 or Isaaks and Srivastava 1989.) Ordinary kriging is a standard interpolation technique with a minimum of three steps required to estimate values at unmeasured locations. First, the empirical semivariogram is calculated; second, the empirical semivariogram is modeled; and third, parameter estimates obtained from the modeled semivariogram are used to predict values at unmeasured locations. The semivariance between values for a particular lag distance h is

$$\gamma(h) = \frac{1}{2N(h)} \sum \left(v_j - v_i\right)^2$$

where N is the number of pairs (i,j), and $v_i - v_j$ is the difference between the values of pair (i,j). A semivariogram is a graph of semivariance by distance class. Several model types may be used to model the empirical semivariogram, including the Gaussian model, wave model, power model, and exponential model. Most variogram models can be characterized by three parameters: the nugget, sill, and range. The nugget refers to the y-intercept of the modeled semivariogram and is a function of microscale variation or measurement error. The sill refers to the maximum value of semivariance (i.e., the total variation in the data), and the range is the distance at which the semivariance reaches 95 percent of the sill.

After the semivariogram has been modeled, ordinary kriging can be used to estimate values at unsampled points. Ordinary kriging is a weighted average such that

$$\hat{V}_0 = \sum_{i=1}^{n} w_i V_i$$

where $\hat{V}_0$ is the estimate at unmeasured location 0, w_i is the weight for the ith observation, and V_i is the value of the ith observation. The weights sum to 1 and are determined by minimizing the overall estimation error. The estimation variance is

$$S_0^2 = w_i \gamma\left(s_i - s_0\right) + \lambda$$

where $\gamma(s_i - s_0)$ is the modeled semivariance for the distance between s_i and s_0, and λ is the Lagrange multiplier from solving the linear system of equations for minimum estimation error.

We predicted forest biomass using kriging at each of the extracted 180 plots using the REAL and REPS data sets. To accomplish this, we first examined the sample variograms for

each of the 11 data sets. Second, we modeled the sample variograms with the power model

$$\gamma(h) = C_1(h)^a$$

In this model, a is dimensionless and dictates the shape of the variogram, and C_1 has the same dimension as the variance. The parameters C_1 and a were estimated using weighted nonlinear regression where the weight was inversely proportional to distance and semivariance. The logic behind this weighting was that small semivariance values near distance 0 have the most importance for kriging. This weighting is similar to the weighting proposed by Cressie (1985). Third, we used the ordinary kriging equation to predict biomass values at the 180 locations extracted before model development using the REAL and REPS data sets.

Delhomme (1978, 1979) first proposed combining regression and kriging. In our study, we used residual kriging for which a regression model was developed to predict forest biomass using percent forest and LAI. The model residuals were then kriged. The percent forest values were collected in the field for each plot, and the LAI values were obtained by intersecting a 1-km resolution map of LAI with the REAL and REPS data sets. The model was developed empirically with general form

$$E(Bm) = \exp\left(cP_f + g\sqrt{P_f LAI}\right), \qquad (1)$$

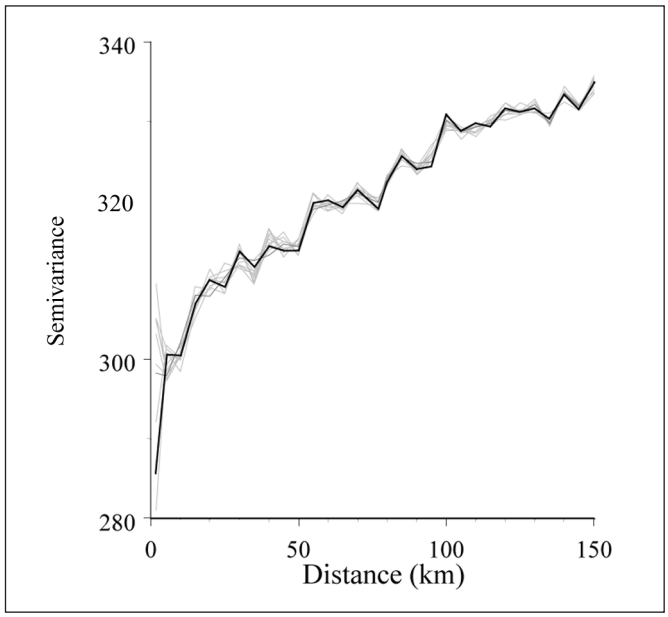
where E(Bm) is the statistical expectation of forest biomass (tons/acre), exp(.) is the exponential function, P_f = percent forest land use, c = percent forest parameter, LAI = leaf area index derived from MODIS satellite imagery, and g = parameter for adjusted LAI. To solve model (1), we first transformed it into its linear form by taking the natural logarithm of each side. Next, we used ordinary least-squares to estimate each parameter. The linear model was then back-transformed, and the semivariance of the residuals was examined and modeled with the power variogram model. We then used ordinary kriging to predict the residual for the prediction for each of the 180 plots extracted from the analysis. The final predicted value of forest biomass was the sum of the predicted value from model (1) and the predicted residual from kriging. This method was applied to the REAL and REPS data sets.

We used analysis of variance to examine the influence that plot coordinate perturbation had on the accuracy of the spatial models. Specifically, we tested for differences in mean error and mean squared error among results for the 180 plots extracted from the data. If we observed an overall difference, we then examined the reason for the observed difference using Tukey's studentized range test.

Results

We visually inspected the empirical variograms of forest biomass for differences. Plot coordinate perturbation had the greatest influence on semivariance values between plots closer than approximately 1,900 m (fig. 1); i.e., the plot coordinate perturbation changed correlations among observations for plots separated by relatively short distances. The total variation and range of spatial autocorrelation were relatively uninfluenced which was expected because relatively small shifts in plot locations should influence only local variability. We used a power variogram model to develop the theoretical variogram. The power model does not technically have a sill and range, but the plot coordinate perturbation did influence the parameter estimates $\hat{C}_1$ and $\hat{a}$.

Figure 1.—*Empirical semivariogram for the REAL data set (solid dark line) and the REPS data sets (solid gray lines).*



No statistically significant difference exists in mean error or mean square error among kriging estimates based on the REAL and REPS data sets (table 1). The mean error of estimates based on the REAL data set was 1.192 tons/acre, which fell between the high and low limits from the REPS data sets. The estimates based on the REAL data set had the highest mean square error. Because no statistically significant difference exists between estimates based on the REAL and REPS data sets, they all fell in the same Tukey grouping.

We developed a regression model to predict forest biomass based on percent forest and LAI. The linearized model had $R^2 = 0.88$, $\hat{c} = 2.59$, and $\hat{g} = 0.23$. Standardized regression coefficients were used to compare the influence of each predictor variable when the variables are measured in different units (SAS 1999). The standardized regression coefficients were $\hat{c} = 0.811$ and $\hat{g} = 0.127$, suggesting that the model was most heavily influenced by P_f. Figure 2 shows the nonlinear form of model (1). The parameter estimates were slightly different for models developed from the REPS data sets. The range of estimates for the *c* parameter was 2.59–2.73, and the range of estimates for the g parameter was 0.16–0.23. All the regression models based on the REPS data sets had $R^2 \approx 0.88$ which was similar to that obtained for the regression model based on the REAL coordinates.

No statistically significant difference exists in mean error or mean square error among residual kriging estimates based on the REAL and REPS data sets (table 1). Residual kriging using the REAL data set had a mean error of 1.201 tons/acre and a mean squared error of 316.83 (tons/acre)^2. The highest mean error from the REPS data sets was 1.308 tons/acre, and the highest mean squared error was 321.12 (tons/acre)^2. The lowest mean error and mean squared error were 1.032 tons/acre and 311.36 (tons/acre)^2, respectively. Because no statistically significant difference exists between estimates based on the REAL and REPS data sets, they all fell in the same Tukey grouping.

Figure 2.—*Predicted biomass based on model (1).*

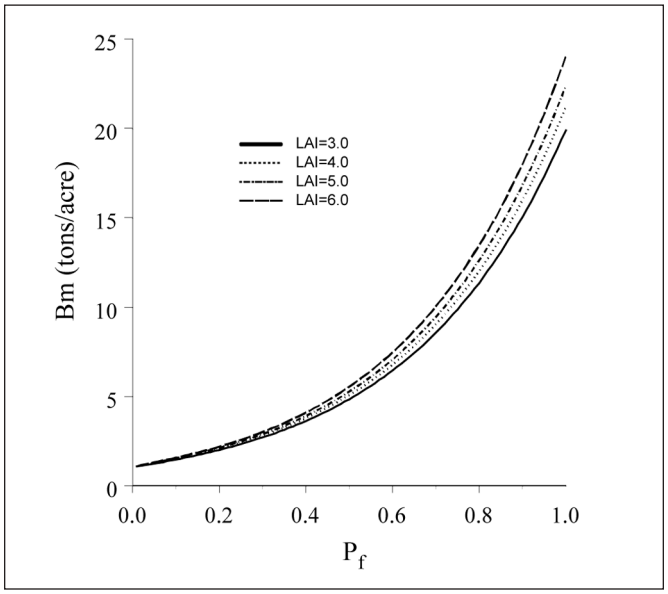


Table 1.—*Mean prediction error and mean squared prediction error for estimates of biomass based on the REAL and REPS data sets.*

Data	Kriging estimates Mean error	Residual kriging estimates Mean squared error	Mean error	Mean squared erro
	tons/acre	*(tons/acre)*^{2}	*tons/acre*	*(tons/acre)*^{2}
REAL	1.192	343.07	1.201	316.83
REPS01	1.211	339.24	1.032	315.43
REPS02	1.218	339.13	1.247	321.12
REPS03	1.184	338.40	1.212	213.97
REPS04	1.190	340.29	1.221	316.48
REPS05	1.234	339.07	1.308	315.61
REPS06	1.179	341.23	1.083	315.07
REPS07	1.213	338.27	1.174	316.12
REPS08	1.159	338.40	1.123	316.04
REPS09	1.203	341.68	1.189	316.85
REPS10	1.177	338.73	1.281	311.36

Discussion

In this study, the plot coordinate perturbations did not influence the accuracy of the spatial models. Two characteristics of the data, however, may have contributed to this outcome. First, the biomass variable had a weak spatial structure based on the variogram. Second, the regression model was based on percent forest estimates from the field and LAI estimates based on MODIS imagery. Based on the standardized regression coefficients, the percent forest variable had the highest weight in the model.

We considered forest biomass to exhibit a weak spatial structure because the proportion of the semivariance explained by distance was relatively small. We can examine the strength of the spatial structure in many ways. With the power variogram model, the C_1 parameter typically has estimates between 0 and 2 (SAS 1996). When C_1 approaches 0, the semivariogram approaches a horizontal line. Our estimate was $\hat{C}_1 = 0.033$, which suggests a weak spatial structure; i.e., biomass has large variance and exhibits little spatial correlation, even at small distances. When the empirical semivariogram exhibits a horizontal linear structure (i.e., slope = 0 or horizontal line), the best linear unbiased predictor is the average. When the spatial structure is weak, the kriging equation will produce estimates close to the global average. We hypothesize that when the variable of interest has a weak spatial structure, the plot coordinate perturbations have a minimal effect on the accuracy of kriging estimates because estimates approach the global average.

The regression model developed for this study was most heavily influenced by the percent forest variable. This variable was collected in the field so that each plot, regardless of plot coordinate perturbation, had the actual field estimate for percent forest. The LAI variable was obtained by intersecting the imagery with the plot locations. The MODIS data were 1-km in resolution which matched well with the plot coordinate perturbation, because 95 percent of the perturbed plot locations were within 0.8 km of the actual plot location. Also, LAI estimates were adjusted by the percent forest in model (1). We suggest that the influence of plot coordinate perturbation on the accuracy of residual kriging depends on the resolution and the spatial autocorrelation of the intersected information.

For spatial models developed by intersecting ancillary data (e.g., regression kriging, residual kriging, mixed models), the two most important characteristics of the ancillary data are the resolution and the spatial autocorrelation. The resolution of the ancillary data is important because the probability that a plot will be assigned incorrect information during intersection decreases with decreasing resolution. For example, plots will more likely be assigned the correct value from intersection when the resolution of the ancillary data is 5 km as opposed to 30 m. The autocorrelation of the ancillary data is also important because plots will more likely be assigned a value similar to the correct value when high autocorrelation exists. For example, if the resolution of the ancillary data is 30 m, and the spatial autocorrelation is zero (i.e., a random spatial pattern), the probability of assigning the correct value to the plot is very low. If the resolution of the ancillary data is 30 m, and the spatial autocorrelation is large, however, the probability of assigning the correct value to the plot is much greater.

Conclusions

The objective of this study was to examine the influence of plot coordinate perturbation on the accuracy of kriging estimates and residual kriging estimates. For the cases we examined, no statistically significant influence on accuracy exists. Generalizations should be made with caution due to the potential influence of the following factors:

1. **Spatial structure in the variable of interest.** A weak spatial structure should produce little effect, while a strong spatial structure may produce a larger effect.
2. **Spatial resolution of ancillary data.** Coarse spatial resolution decreases the probability of assigning incorrect ancillary data values to a plot, while fine spatial resolution increases the probability.
3. **Spatial autocorrelation of ancillary data.** High spatial autocorrelation in ancillary data decreases the probability of large errors in the assignment of ancillary data value to a plot, while low spatial autocorrelation increases the probability.

We suggest that these topics be further investigated using simulated variables of known spatial structure.

Literature Cited

Coulston, J.W.; Smith, G.C.; Smith, W.D. 2003. Regional assessment of ozone sensitive tree species using bioindicator plants. Environmental Monitoring and Assessment. 83: 113–127.

Cressie, N. 1985. Fitting variogram models by weighted least squares. Mathematical Geology. 17: 563–586.

Cressie, N. 1993. Statistics for spatial data. New York: John Wiley & Sons. 900 p.

Delhomme, J.P. 1978. Kriging in hydrosciences. Advanced Water Resources. 1: 251–266.

Delhomme, J.P. 1979. Spatial variability and uncertainty in groundwater flow parameters: a geostatistical approach. Water Resources Research. 15: 269–280.

Isaaks, E.H.; Srivastava, R.M. 1989. An introduction to applied geostatistics. New York: Oxford University Press. 561 p.

Morin, R.S.; Gottschalk, K.W.; Liebhold, A.M. 2003. Potential susceptibility of eastern forests to Sudden Oak Death, *Phytophthora ramorum*. 2003 forest health monitoring working group meeting. http://www.fhm.fs.fed.us/posters/posters03/sod.pdf. (20 May 2005).

SAS Institute, Inc. 1996. SAS/STAT® technical report: spatial prediction using the SAS® system. Cary, NC: SAS Institute, Inc. 80 pp.

SAS Institute, Inc. 1999. SAS/STAT® users guide, version 8. Cary, NC: SAS Institute, Inc. 3884 p.

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