

True versus perturbed forest inventory plot locations for modeling: a simulation study

John W. Coulston, Kurt H. Riitters, Ronald E. McRoberts, Greg A. Reams, and William D. Smith

Abstract: USDA Forest Service Forest Inventory and Analysis plot information is widely used for timber inventories, forest health assessments, and environmental risk analyses. With few exceptions, true plot locations are not revealed; the plot coordinates are manipulated to obscure the location of field plots and thereby preserve plot integrity. The influence of perturbed plot locations on the development and accuracy of statistical models is unknown. We tested the hypothesis that the influence is related to the spatial structure of the data used in the models. For ordinary kriging we examined the difference in mean square error based on true and perturbed plot locations across a range of spatial autocorrelations. We also examined the difference in mean square error for regression models developed with true and perturbed plot locations across a range of spatial autocorrelations and spatial resolutions. Perturbing plot locations did not significantly influence the accuracy of kriging estimates, but in some situations linear regression model development and accuracy were significantly influenced. Unless the independent variable has high spatial autocorrelation, only coarse spatial resolution data should be used to develop linear regression models.

Résumé : Les données des placettes du programme d'analyse et d'inventaire forestier du service des forêts du département de l'agriculture des États-Unis sont largement utilisées pour inventorier les ressources ligneuses, évaluer la santé des forêts et analyser les risques environnementaux. À quelques exceptions près, l'emplacement réel de ces placettes n'est pas révélé; leurs coordonnées sont manipulées pour cacher leur localisation sur le terrain et ainsi préserver leur intégrité. De ce fait, l'influence de la modification de l'emplacement des placettes sur le développement et la précision des modèles statistiques est inconnue. Nous avons testé l'hypothèse que cette influence est reliée à la structure spatiale des données utilisées dans les modèles. Pour le krigeage ordinaire, nous avons examiné la différence entre l'erreur quadratique moyenne obtenue avec l'emplacement réel et celle obtenue avec l'emplacement modifié pour une gamme d'autocorrélations spatiales. Nous avons aussi examiné la différence entre l'erreur quadratique moyenne des modèles de régression basés sur l'emplacement réel et l'emplacement modifié pour une gamme d'autocorrélations spatiales et de résolutions spatiales. La modification de l'emplacement des placettes n'a pas influencé significativement la précision des estimations du krigeage mais, dans certaines situations, le développement et la précision du modèle de régression linéaire ont significativement été influencés. Ainsi, à moins que la variable indépendante ait une autocorrélation spatiale élevée, seules les données à faible résolution spatiale devraient être utilisées pour développer les modèles de régression linéaire.

[Traduit par la Rédaction]

Received 6 May 2005. Accepted 28 October 2005. Published on the NRC Research Press Web site at <http://cjfr.nrc.ca> on 22 March 2006.

J.W. Coulston.¹ North Carolina State University, Department of Forestry and Environmental Resources, Box 8008, Raleigh, NC 27695, USA.

K.H. Riitters and W.D. Smith. USDA Forest Service, Southern Research Station, 3041 Cornwallis Road, Research Triangle Park, NC 27709, USA.

R.E. McRoberts. USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108, USA.

G.A. Reams. USDA Forest Service, Washington Office, 1601 North Kent Street, Arlington, VA 22209, USA.

¹Corresponding author (e-mail: jcoulston@fs.fed.us).

Introduction

The USDA Forest Service Forest Inventory and Analysis Program (FIA) collects data on tree and forest attributes using a design assumed to produce a random equal-probability sample. These data are used for many purposes, including timber inventories, forest health assessments, risk assessments, and predictive modeling of forest attributes. To preserve landowner privacy and field-plot integrity, true plot locations are not available to the general public. To enable data access while preserving privacy, FIA has proposed several methods of manipulating plot coordinates. This paper examines the effects of these manipulations on the development and accuracy of two types of statistical models.

FIA field plots are part of a three-part sample of forests in the United States. The base FIA sample is a 27× intensification

of the Environmental Monitoring and Assessment Program (EMAP) isotropic sampling grid (White et al. 1992). The nominal sampling intensity is one plot per 2403 ha (Brand 2004). The sampling grid is triangular and each sample point is represented by a hexagonal Thiessen polygon. Each hexagon contains one randomly located FIA field plot (Reams et al. 2005).

Because plots are located on both public and private properties, private owners' awareness of the inventory and granting access for measurements are essential. For that reason, the true plot locations have always been confidential. Prior to 2002, field-plot locations were revealed within approximately 1.6 km of the true location. New methods of perturbing plot locations are needed because of recent legislation designed to improve data access while protecting landowner privacy (Smith 2002).

There is no national standard for perturbing plot locations, but there are guidelines that may be satisfied at the regional level using different techniques. One method currently used is to perturb plot locations by manipulating plot coordinates and "swapping" (i.e., exchanging) data among plots. Plot coordinates are manipulated by randomly shifting the x and y coordinates of the true locations for all plots. Swapping plot attributes involves the exchange of observations among some plots. Plot coordinates are usually only shifted within a single county, and plot attributes are only swapped if the plots are sufficiently similar (Lister et al. 2005). This ensures that county-level estimates are not influenced. However, perturbing plot locations influences the spatial characteristics of the data and therefore can influence the development and accuracy of spatially explicit predictive models.

Many spatial statistical models have been developed with FIA data. For example, Jenkins et al. (2001) developed models to predict biomass and net primary production for 0.5° cells tiling the mid-Atlantic region of the United States. Moisen and Frescino (2002) compared five predictive modeling techniques using FIA data as response variables, with explanatory variables derived from satellite imagery. Coulston et al. (2003) used ordinary kriging to predict potential ozone injury and assess ozone injury risk to ozone-sensitive northeastern tree species. Morin et al. (2003) used FIA field plot data 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). The interpolated surface was then intersected with other spatial data and used to assess the potential susceptibility of eastern forests to *P. ramorum*.

An earlier case study (Coulston et al. 2006) examined the influence of perturbed plot locations on the accuracy of ordinary kriging and residual kriging estimates of forest biomass. For residual kriging, a regression model was developed to predict forest biomass using percent forest cover and leaf area index (LAI; estimated from MODIS satellite imagery) on a per-plot basis. The model residuals were then kriged, and the final biomass estimate was the sum of the regression model estimate plus the predicted residual from kriging. For both models, there were no statistically significant differences ($\alpha = 0.05$) between the accuracy of the models developed with the true FIA plot locations and the accuracy of the models developed with the perturbed FIA plot locations. However, it was noted that the influence of perturbed plot

locations on the accuracy of ordinary kriging estimates is related to the spatial autocorrelation of the data. Furthermore, because the explanatory variable in the regression model was obtained by intersecting continuous raster data with discrete plot locations, the influence on model development and accuracy was related to the spatial resolution and autocorrelation of the explanatory data.

The general objective of this study was to develop practical guidelines for developing spatial models using FIA plot data with perturbed locations. We systematically address key modeling issues through simulations to evaluate (1) the influence of perturbing FIA plot locations on the accuracy of ordinary kriging estimates across a range of spatial autocorrelation and (2) the influence on the development and accuracy of linear regression models across a range of spatial resolutions and spatial autocorrelations of the explanatory variable.

Methods

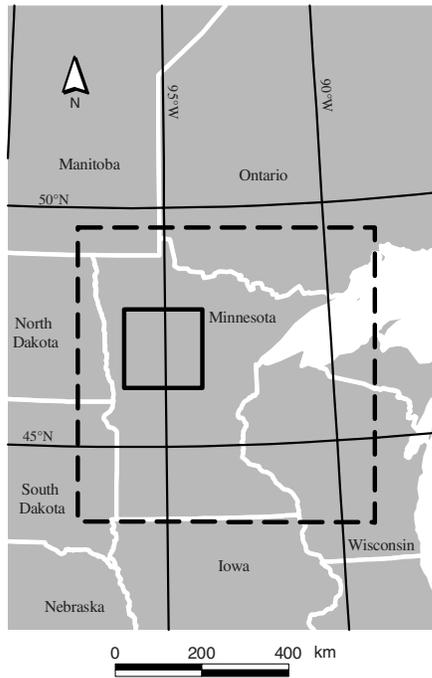
Database development

True and perturbed plot locations, along with plot measurements, were obtained for 12 730 FIA plots in Minnesota. Plot locations were perturbed according to procedures used by the regional FIA program. Here the term "perturbed" is used to indicate both plot coordinate manipulation and the exchange of plot attributes. We randomly extracted 100 test plots for model comparisons. Models were developed without the test plots, and predictions were made for each test plot. This allowed us to compare the accuracy of the models based on true locations with that of models based on perturbed locations.

There are many continuous forestry variables of interest in modeling exercises (e.g., forest biomass, percent forest cover). For example, Coulston et al. (2006) examined the influence of using perturbed plot locations on predictive models of forest biomass. However, in this case rather than arbitrarily selecting one particular forestry variable, we used data sampled from simulated maps that were normally distributed with mean zero and unit variance ($N(0,1)$). Several forestry variables can be approximately normally distributed (e.g., basal area, growth) or can be normalized, and all variables can be standardized to mean zero and variance one by subtracting the mean from each observation and then dividing that value by the standard deviation. Also, using normally distributed data ensured that we did not violate the statistical assumptions about linear regression errors.

The simulated maps were created at spatial resolutions approximating satellite imagery (30–1000 m) and climate-model output (2000 m). We used the R statistical package GaussRF function (Schlather 1999) to simulate Gaussian random fields with known spatial characteristics. A random field is the two-dimensional equivalent to the one-dimensional stochastic process (Schlather 2001) and can be described by its variogram (see Isaaks and Srivastava 1989). In variogram models, the "nugget" refers to the y intercept (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), the "scale" refers to the portion of semivariance explained by distance, and the "range" is the distance at which the semivariance reaches 95% of the sill.

Fig. 1. Location and relative areal extent of simulated raster maps. The solid line represents the 30 m spatial resolution maps, and the broken line represents all other maps.



The strength of the spatial autocorrelation can be quantified by a measure similar to the nugget-to-sill ratio proposed by Kravchenko (2003):

$$[1] \quad c = \frac{v_1}{v_0 + v_1}$$

where, c is the spatial autocorrelation, v_0 is the nugget, v_1 is the scale, and $v_0 + v_1$ is the sill.

We used a spherical variogram model to generate raster maps for Minnesota with mean zero and variance one for five spatial resolutions $r = (30, 250, 500, 1000, 2000 \text{ m})$ and eight spatial autocorrelations $c = (0.01, 0.05, 0.15, 0.35, 0.55, 0.75, 0.95, 0.99)$. To generate the eight levels of spatial autocorrelation v_0 and v_1 were allowed to change but the sill ($v_0 + v_1$) and range remained constant. The range was constant at 20 km, and the sill was constant at one. The final set of simulations included 40 raster maps of continuous data ranging from 30 to 2000 m in spatial resolution and with spatial autocorrelation $c = 0.01$ to $c = 0.99$. All maps, except the 30 m resolution maps, encompassed Minnesota (Fig. 1). (Because of computer limitations, raster maps with more than 6000 cells per side could not be generated.) Random sampling of these maps produced an $N(0,1)$ variable for each map. Figure 2 shows an example of the simulated 500 m resolution data for three levels of spatial autocorrelation and the histograms based on sampling these maps using the FIA plot locations.

Ordinary kriging

Ordinary kriging (see Isaaks and Srivastava 1989) is a standard interpolation technique used to estimate values at unmeasured locations. We used the $r = 250 \text{ m}$ spatial resolution maps with spatial autocorrelations of $c = (0.15, 0.35,$

$0.55, 0.75, 0.75, 0.99)$ to examine the influence perturbed plot locations have on the accuracy of kriging estimates. We did not use maps with spatial autocorrelation of $c = (0.01, 0.05)$ because it is unlikely that kriging will be used if the spatial autocorrelation is that low. Each of the six maps were intersected with the true FIA plot locations to create a data set where each record had the true plot location, the perturbed plot location, and a value extracted from each of the six simulated maps. After removing the 100 randomly selected test plots, empirical semivariograms were calculated for each level of spatial autocorrelation based on the true plot locations and the perturbed plot locations. The empirical variograms based on the true plot locations were then modeled using a spherical model (Isaaks and Srivastava 1989) and nonlinear regression. For the perturbed plot locations, the empirical variograms were modeled using spherical, Gaussian, and exponential models. Multiple models were used to account for possible differences in model selection as a result of not knowing the real model a priori. The resulting parameter estimates were applied in ordinary kriging to estimate values at each true location of the test plots. Next we calculated the mean squared error (MSE) of prediction:

$$[2] \quad \text{MSE}_p = \frac{1}{n} \sum_{i=1}^{100} (k - \hat{k}_p)^2$$

where $p = 0$ denotes true locations, $p = 1$ denotes perturbed locations, k is the actual value from intersection, and \hat{k}_p is the estimate from kriging for the $n = 100$ test plots. We used an analysis of variance (ANOVA) to test the null hypothesis that $\text{MSE}_0 = \text{MSE}_1$ for each level of spatial autocorrelation and variogram model.

Linear regression

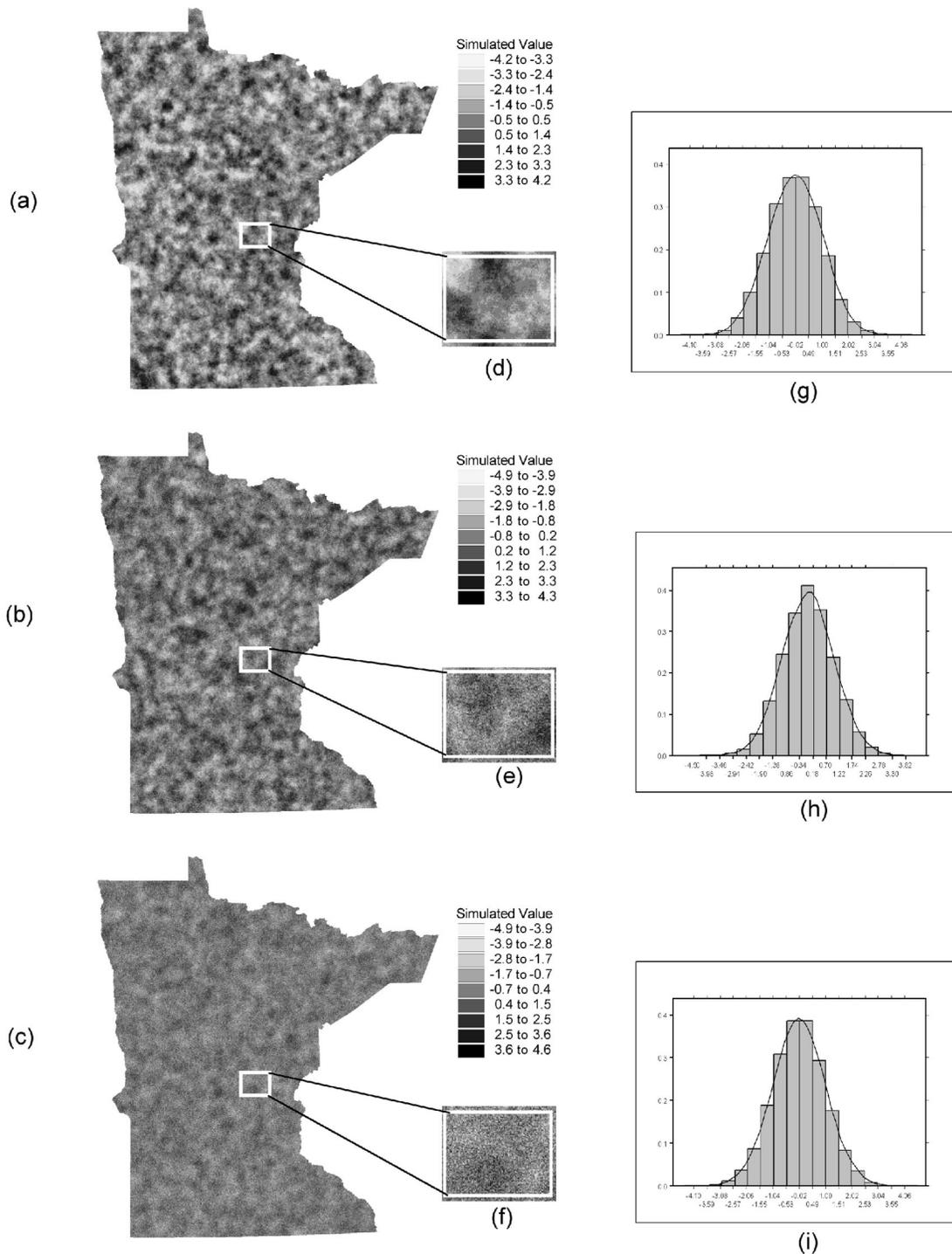
In a regression model, the dependent variable may come from field-plot observations and the independent variable may come from a raster map. The objective of this analysis is to examine the influence of perturbed plot locations on regression model development and accuracy, for different levels of spatial resolution and spatial autocorrelation. We also tested whether the influence is related to the correlation between dependent and independent variables. For this purpose, we constructed several data sets such that the dependent variables had known correlations with the independent variables.

The following procedures were used to construct data sets of dependent variables representing the field-plot observations. Let D be a field-plot observation (dependent variable), and let I be a value for that location from a raster map (independent variable). We simulated D values from the I values by adding a controlled amount of variance to the I values as follows. Note that for the case of perfect correlation ($\rho = 1.0$), D is a linear function of I . For the case of less than perfect correlation ($\rho < 1.0$), D can be derived from I by the following equation:

$$[3] \quad D_m = a + \left(\frac{Z}{L_m} \right) I$$

where a is an arbitrary constant, Z is a normally distributed random variable with mean zero and variance one, and L_m is a factor used to control the resulting correlation between I

Fig. 2. Examples of the simulated $r = 500$ m resolution data for spatial autocorrelations (a) $c = 0.95$, (b) $c = 0.55$, and (c) $c = 0.15$. The white box shows a $4\times$ zoom for (d) $c = 0.95$, (e) $c = 0.55$, and (f) $c = 0.15$. Note that as c decreases, the spatial pattern approaches randomness and the maps have a “salt and pepper” appearance. The histograms were created by sampling each map with the FIA plot locations for (g) $c = 0.95$, (h) $c = 0.55$, and (i) $c = 0.15$.



and D . Note that when $L_m = 3$, the correlation between I and D is approximately 0.95. For $L_m = 2.0, 1.5$, and 1.21 , the resulting correlations are approximately 0.89, 0.84, and 0.77, respectively. These correlations, in turn, represent linear regression models for which the R^2 values are 0.9, 0.8, 0.7, and 0.6, respectively. We use the subscript m to denote these four “true R^2 ’s”.

For each of the simulated maps representing different levels of spatial resolution and spatial autocorrelation (described earlier), eq. 3 was applied to generate four simulated plot-level dependent variables. After intersection with maps of true and perturbed plot locations, each record in the resulting data set had values from the four dependent variables (corresponding to different levels of m), the value of the “true” in-

Table 1. Mean square errors (MSE) from kriging estimates and probability of a greater F statistic under the null hypothesis $MSE_{\text{true locations}} = MSE_{\text{perturbed locations}}$ for each level of spatial autocorrelation c and each variogram model (spherical, Gaussian, and exponential).

c	True locations		Perturbed locations					
	Spherical		Spherical		Gaussian		Exponential	
	MSE		MSE	$P > F$	MSE	$P > F$	MSE	$P > F$
0.15	0.747		0.748	0.993	0.747	0.999	0.751	0.972
0.35	0.739		0.731	0.956	0.749	0.953	0.741	0.990
0.55	0.436		0.425	0.882	0.441	0.954	0.419	0.819
0.75	0.446		0.450	0.966	0.459	0.886	0.458	0.895
0.95	0.258		0.289	0.559	0.287	0.519	0.296	0.487
0.99	0.205		0.217	0.808	0.239	0.547	0.221	0.737

dependent variable (based on the true plot location) and the value of the “perturbed” independent variable (based on the perturbed plot location) for each level of spatial resolution and spatial autocorrelation.

After removing the 100 randomly selected test plots from the data sets, regression models were developed using ordinary least squares and the linear model

$$[4] \quad D_{rcmp} = a_{rcmp} + b_{rcmp}I_{rcp} + \epsilon$$

where a_{rcmp} is the intercept for spatial resolution r , spatial autocorrelation c , true R^2 m , and for $p = 0$ (true locations) and $p = 1$ (perturbed locations); b_{rcmp} is the slope for spatial resolution r , spatial autocorrelation c , true R^2 m , and for $p = 0$ and $p = 1$; I_{rcp} is the value of the independent variable based on intersection for spatial resolution r and spatial autocorrelation c using $p = 0$ and $p = 1$; and ϵ is the uncorrelated error $\sim N(0, \sigma^2)$.

A total of 320 regression models were developed. The regression models were then applied to the 100 test plots to calculate \hat{D}_{rcmp} . We calculated the mean square error of prediction as

$$[5] \quad MSE_p = \frac{1}{n} \sum_{i=1}^{100} (D_{rcm} - \hat{D}_{rcmp})^2$$

and tested the null hypothesis that $MSE_0 = MSE_1$ for each level of r , c , and m .

Results

Ordinary kriging

Perturbed plot locations did not influence the accuracy of kriging estimates for any level of spatial autocorrelation regardless of the variogram model used for the perturbed locations. While it did cause a 1%–4% increase in the nugget (increase in local-scale variation), the null hypothesis that $MSE_{\text{true locations}} = MSE_{\text{perturbed locations}}$ was not rejected ($\alpha = 0.05$) (Table 1). The largest difference between MSEs was at the $c = 0.95$ spatial autocorrelation level when the exponential variogram model was used; however, this difference was not statistically significant.

Linear regression

Perturbing plot locations influenced both the development and accuracy of linear regression models. For comparisons, we arbitrarily identified the circumstances where perturbed

locations reduced the model R^2 to less than 0.40. The effects of perturbed locations were most evident for regression models developed from explanatory data with relatively fine resolution (e.g., $r = 30$ m) and low spatial autocorrelation (e.g., $c = 0.01$) (Table 2). However, the effect of perturbed locations on each combination of spatial resolution and spatial autocorrelation was also influenced by the strength of the relationship (i.e., the true R^2). In general, as the true R^2 decreased, statistically significant differences ($\alpha = 0.05$) in MSE were less likely to occur and model R^2 's < 0.40 were more likely to occur.

For the true $R^2 = 0.90$ scenario, model R^2 's fell below 0.40 when the spatial autocorrelation was $c \geq 0.55$ except for spatial resolutions of $r = (1000, 2000$ m) (Table 2). Models were more likely to have R^2 's ≥ 0.40 for this scenario. However because this scenario was based on a relatively strong relationship between the dependent and independent variables, significant differences ($\alpha = 0.05$) in MSE between estimates based on true plot locations and estimates based on perturbed plot locations were more likely to occur. This point was illustrated by the $r = (30, 250$ m) resolutions where significant differences ($\alpha = 0.05$) in MSE were found for all levels of spatial autocorrelation except for $c = (0.95, 0.99)$.

When the true $R^2 = 0.80$, model R^2 's below 0.40 and statistically significant difference ($\alpha = 0.05$) in MSE occurred under similar levels of r and c (Table 2). This overlap included all simulations where $r = 30$ m and the spatial autocorrelation was $c \leq 0.75$. Simulations where $r = (250, 500$ m) and spatial autocorrelation was $c \leq 0.55$ were also included in this overlap. Conversely, when the spatial resolution was 2000 m, the only case where the $R^2 < 0.40$ was for the $c = 0.05$ spatial autocorrelation level. However, there was no statistical difference ($\alpha = 0.05$) in MSE under this situation.

For the true $R^2 = 0.70$ and true $R^2 = 0.60$ scenarios with spatial resolutions $r < 1000$ m the results were the same. Only data with a very high spatial autocorrelation $c > 0.75$ produced models with an $R^2 \geq 0.40$ for the $r = (30, 250, 500$ m) resolutions. However, for these resolutions statistically significant differences ($\alpha = 0.05$) in MSE were only found when the spatial autocorrelation was $c \leq 0.35$.

Discussion

Our results suggest that perturbing plot locations does not significantly influence the accuracy of kriging estimates, but the development and accuracy of simple linear regression

Table 2. Regression model R^2 and the probability of a greater F statistic under the null hypothesis $MSE_{\text{true locations}} = MSE_{\text{perturbed locations}}$ for each dependent variable, spatial resolution r , and spatial autocorrelation c .

r (m)	c													
	0.01		0.05		0.15		0.35		0.55		0.75		0.95	
	Model R^2	$P > F$												
True $R^2 = 0.60$														
30	0.00	0.000*	0.00	0.000*	0.01	0.001*	0.07	0.003*	0.20	0.659	0.26	0.854	0.48	0.810
250	0.00	0.000*	0.00	0.000*	0.02	0.004*	0.08	0.009*	0.18	0.226	0.31	0.827	0.48	0.861
500	0.02	0.002*	0.03	0.035*	0.05	0.045*	0.12	0.036*	0.22	0.244	0.34	0.709	0.51	0.939
1000	0.14	0.102	0.15	0.129	0.17	0.137	0.24	0.543	0.32	0.652	0.42	0.844	0.51	0.952
2000	0.30	0.749	0.29	0.792	0.31	0.810	0.38	0.908	0.43	0.837	0.47	0.906	0.52	0.948
True $R^2 = 0.70$														
30	0.00	0.000*	0.00	0.000*	0.01	0.000*	0.08	0.000*	0.23	0.326	0.31	0.625	0.56	0.794
250	0.00	0.000*	0.00	0.000*	0.02	0.000*	0.10	0.001*	0.21	0.087	0.36	0.656	0.56	0.808
500	0.02	0.000*	0.03	0.003*	0.05	0.008*	0.14	0.004*	0.25	0.107	0.40	0.574	0.59	0.912
1000	0.17	0.023*	0.17	0.030*	0.20	0.030*	0.28	0.314	0.37	0.462	0.49	0.762	0.60	0.930
2000	0.35	0.564	0.34	0.624	0.37	0.656	0.44	0.790	0.50	0.767	0.54	0.861	0.61	0.925
True $R^2 = 0.80$														
30	0.00	0.000*	0.01	0.000*	0.02	0.000*	0.09	0.000*	0.27	0.038*	0.37	0.259	0.65	0.783
250	0.00	0.000*	0.01	0.000*	0.02	0.000*	0.11	0.000*	0.24	0.009*	0.42	0.339	0.65	0.704
500	0.02	0.000*	0.03	0.000*	0.06	0.001*	0.16	0.000*	0.29	0.019*	0.46	0.344	0.68	0.854
1000	0.19	0.001*	0.20	0.001*	0.23	0.001*	0.32	0.071	0.43	0.185	0.56	0.596	0.69	0.884
2000	0.41	0.258	0.40	0.330	0.43	0.369	0.51	0.532	0.57	0.630	0.63	0.766	0.70	0.879
True $R^2 = 0.90$														
30	0.00	0.000*	0.01	0.000*	0.02	0.000*	0.09	0.000*	0.30	0.000*	0.42	0.013*	0.73	0.824
250	0.00	0.000*	0.01	0.000*	0.03	0.000*	0.12	0.000*	0.57	0.000*	0.47	0.027*	0.73	0.469
500	0.03	0.000*	0.04	0.000*	0.07	0.000*	0.18	0.000*	0.32	0.000*	0.52	0.070	0.76	0.706
1000	0.22	0.000*	0.22	0.000*	0.26	0.000*	0.36	0.001*	0.49	0.009*	0.63	0.269	0.77	0.766
2000	0.45	0.014*	0.45	0.036*	0.49	0.042*	0.57	0.101	0.64	0.349	0.71	0.533	0.79	0.758

*Rejection of the null hypothesis at or below the $\alpha = 0.05$ level.

models is significantly influenced in some situations. We can suggest some general guidelines for simple linear regression with an FIA plot attribute as the dependent variable and an independent variable obtained by intersecting continuous raster data with perturbed plot locations. First, coarse spatial resolution (e.g., 1000–2000 m) data should be used when the spatial autocorrelation of the explanatory variable is less than 0.75; finer resolution raster data (e.g., 30–500 m) should only be used if the spatial autocorrelation is greater than 0.75. Second, if fine-resolution models to predict plot attributes are to be developed from maps exhibiting low spatial autocorrelation, then it is necessary to have true plot locations or to use one of the FIA Spatial Data Services Centers (http://www.fs.fed.us/ne/fia/spatial/index_ss.html) to perform the intersection using true plot locations.

Lister et al. (2005) described how to exchange plot attributes using a similarity value based on Euclidean distance between plots, forest-type group, and productivity class. This ensures that only plots with some degree of similarity have their attributes swapped. In general, only a small percentage of plots have their attributes swapped (McRoberts et al. 2005). However, for forest characteristics such as biomass that are related to forest-type group and productivity class, the effect of exchanging data is probably less than our analysis would indicate. For other variables such as tree damage and lichen communities that are not strongly related to forest type and

productivity class, our results represent a worse-case scenario because there is no relationship between the dependent variables and the variables used to determine swapping.

Our approach could be extended to consider other spatial data types such as categorical raster data, continuous vector data, and categorical vector data. Our findings should apply to predictive models (e.g., Moisen and Frescino 2002), where FIA plot data are the dependent variables and the independent variables are derived from spatial intersection of ancillary data. The two most important characteristics of the ancillary data are the spatial resolution and the spatial autocorrelation. In the case of vector data, the minimum mapping unit corresponds to the spatial resolution. In the case of categorical data, “contagion” is analogous to spatial autocorrelation. Plots with perturbed locations have a higher probability of being assigned a “correct” value or similar value during intersection when the ancillary data have a relatively coarse spatial resolution, or a large minimum mapping unit (vector maps), or a high degree of contagion (categorical raster maps).

In summary, FIA plot data are valuable for examining a wide range of environmental topics. Plot confidentiality is required by law and is necessary to ensure long-term data collection on private land. The perturbed plot locations that are available should be used with caution. In some cases (e.g., ordinary kriging) using perturbed plot locations does not significantly affect accuracy. In other cases (e.g., linear

regression) using perturbed plot locations can influence both model accuracy and model development when ancillary data are intersected with perturbed plot locations. The results presented here should be used as guidelines to decide when perturbed plot locations are sufficient for modeling purposes and when the FIA Spatial Data Services Centers should be used to perform spatial overlays.

Acknowledgements

The research described in this article was supported in part by a cooperative agreement between North Carolina State University and the US Forest Service.

References

- Brand, G. 2004. Forest Inventory and Analysis sampling hexagons. FIA fact sheet series. Available from http://fia.fs.fed.us/library/fact-sheets/data-collections/Sampling_hexagons.pdf [cited 11 April 2005].
- Coulston, J.W., Smith, G.C., and Smith, W.D. 2003. Regional assessment of ozone sensitive tree species using bioindicator plants. *Environ. Monit. Assess.* **83**: 113–127.
- Coulston, J.W., Reams, G.A., McRoberts, R.E., and Smith, W.D. 2006. Practical considerations when using perturbed forest inventory plot locations to develop spatial models: a case study. *In Proceedings of the Sixth Annual Forest Inventory and Analysis Symposium, 21–24 September 2004, Denver, Colo.* Edited by R.E. McRoberts, G.A. Reams, P.C. Van Deusen, and W.H. McWilliams. USDA For. Serv. Gen. Tech. Rep. WO-70. pp. 81–86.
- Isaaks, E.H., and Srivastava, R.M. 1989. *An introduction to applied geostatistics.* Oxford University Press, New York.
- Jenkins, J.C., Birdsey, R.A., and Pan, Y. 2001. Biomass and NPP estimation for the mid-Atlantic region (USA) using plot-level forest inventory data. *Ecol. Appl.* **11**: 1174–1193.
- Kravchenko, A.N. 2003. Influence of spatial structure on accuracy of interpolation methods. *Soil Sci. Soc. Am. J.* **67**: 1564–1571.
- Lister, A., Scott, C.T., King, S.L., Hoppus, M., Butler, B., and Griffith, D. 2005. Strategies for preserving owner privacy in the national information management system of the USDA Forest Service's Forest Inventory and Analysis unit. *In Proceedings of the Fourth Annual Forest Inventory and Analysis Symposium, 19–21 November, 2002, New Orleans, La.* Edited by R.E. McRoberts, G.A. Reams, P.C. Van Deusen, and W.H. McWilliams. USDA For. Serv. Gen. Tech. Rep. NC-252. pp. 163–166.
- McRoberts, R.E., Holden, G.R., Nelson, M.D., Liknes, G.C., Moser, W.K., Lister, A.J., King, S.L., LaPoint, E.B., Coulston, J.W., Smith, W.B., and Reams, G.A. 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. *J. For.* **103**: 275–279.
- Moisen, G.G., and Frescino, T.S. 2002. Comparing five modelling techniques for predicting forest characteristics. *Ecol. Model.* **157**: 209–225.
- Morin, R.S., Gottschalk, K.W., and A.M. Liebhold. 2003. Potential susceptibility of eastern forests to sudden oak death, *Phytophthora ramorum*. 2003 Forest Health Monitoring working Group Meeting, Monterey, Calif. Available from <http://www.na.fs.fed.us/spfo/fhm/posters/posters03/sod.pdf> [cited 11 April 2005].
- Reams, G.A., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F., and Moisen, G.G. 2005. The forest inventory and analysis sampling frame. *In The enhanced forest inventory and analysis program: national sampling design and estimation procedures.* Edited by W.A. Bechtold and P.L. Patterson. USDA For. Serv. Gen. Tech. Rep. SRS-80. pp. 11–26.
- Schlather, M. 1999. Introduction to positive definite functions and to unconditional simulation of random fields. Lancaster University, Lancaster, UK. Tech. Rep. ST-99-10. Available from <http://fhm.fs.fed.us/posters/posters03/sod.pdf> [cited 1 March 2006].
- Schlather, M. 2001. Simulation and analysis of random fields. *R News*, **1**(2): 18–20. Available from http://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf [cited 11 April 2005].
- Smith, W.B. 2002. Forest inventory and analysis: a national inventory and monitoring program. *Environ. Pollut.* **116**: 233–242.
- White, D., Kimerling, A.J., and Overton, W.S. 1992. Cartographic and geometric component of a global sampling design for environmental monitoring. *Cartogr. Geogr. Info. Syst.* **19**: 5–22.