The effect of blurred plot coordinates on interpolating forest biomass: a case study

J.W. Coulston[1] and G.A Reams[2]

[1] Department of Forestry
North Carolina State University
Southern Research Station
Forestry Sciences Laboratory
3041 Cornwallis Road
Research Triangle Park, NC 27709
Phone 919-549-4071
jcoulston@fs.fed.us

[2] USDA Forest Service
Southern Research Station
Forestry Sciences Laboratory
3041 Cornwallis Road
Research Triangle Park, NC 27709
Phone 919-549-4010
greams@fs.fed.us

Abstract

Interpolated surfaces of forest attributes are important analytical tools and have been used in risk assessments, forest inventories, and forest health assessments. The USDA Forest Service Forest Inventory and Analysis program (FIA) annually collects information on forest attributes in a consistent fashion nation-wide. Users of these data typically perform interpolations with the kriging or inverse distance weighting methods which requires the coordinates of each FIA plot. However because of privacy issues, FIA uses two methods to manipulate plot locations to insure landowner privacy. The influence these manipulations have on the accuracy of interpolated surfaces is unknown. We investigated the influence by comparing actual and interpolated estimates of forest biomass created from data with manipulated coordinates for three interpolation techniques. We found that kriging consistently under-performed the inverse distance and Thiessen polygon methods. Overall the inverse distance method performed best. We suggest using the inverse distance method for spatial interpolation of FIA data with blurred plot coordinates when relatively little spatial autocorrelation exists.

Keywords: forest inventory and analysis, spatial statistics, cross-validation, Food security act of 1985

1. Introduction

The USDA Forest Service Forest Inventory and Analysis Program (FIA) collects data on tree and forest attributes using a systematic sample. These data are used for many purposes including timber inventories, forest health assessments, and risk assessments. However because of privacy issues actual plots location can not be revealed to scientists outside the FIA program. FIA proposed two methods to manipulate plot locations to insure landowner privacy. They are "fuzzing" and "fuzzing and swapping". Fuzzing refers to a random shift in the x,y coordinate of the actual plot location. Fuzzing and swapping refers to a random shift in the x,y coordinate of the actual plot location and a random swapping of plot attributes between plots (*e.g.*, tree volume m^3ha^{-1}). The effects these manipulations have on the accuracy of spatial interpolations are unknown.

FIA field plots are one part of a 3 phase sample of forests in the United States. The base FIA sample is a 27x intensification of the EMAP isotropic sampling grid (White *et al.* 1992) and covers the entire United States. The

nominal sampling intensity is approximately 1 plot per 2430 ha (Brand 2004). The EMAP and FIA sampling grids are triangular and each grid point is represented by a Thiessen polygon which is hexagonal (Figure 1). FIA used various sampling designs prior to the EMAP sampling grid therefore the location of the FIA field plot within the hexagon is not necessarily the center. The actual location is based on whether there was an existing field plot in the hexagon (FIA or Forest Health Monitoring). If an existing plot was located in the hexagon, then it was chosen. If several existing field plots were present, the one closest to the hexagon center was used. If no existing field plots were present, then a new sample location was selected based on a random azimuth and distance from hexagon center (Reams *et al*. In review).

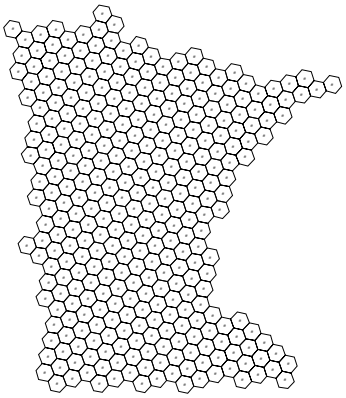

Figure 1. The EMAP base grid for Minnesota and the Thiessen polygon for each plot.

Field plots are located on both public and private properties. Public and private owner's awareness of the FIA program and granting access to the forest land is essential to the program. For this reason, the exact plot coordinates are kept confidential. The confidentiality policy is consistent with the Privacy Act of 1974 and the 1985 Food Security Act. However, new legislation includes further provisions for confidentiality of FIA information (Smith 2002). Prior to 2002, FIA field plot locations were available to the general public outside of the FIA program with fuzzed locations within approximately 1.6 km of the actual location. Currently the proposed measures to insure confidentiality are (1) to fuzz 95 percent of plot location to within 0.8 km, (2) fuzz the remaining 5 percent within 1.6 km of the actual locations and (3) to swap 1-20 percent of the plots. Fuzzing and particularly swapping can influence the spatial characteristics of the data.

Spatial statistics are widely used in forestry, ecology, and other disciplines. Many applications of spatial statistics in these disciplines use FIA data. For example, 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). This interpolated surface was then intersected with other spatial data and used to assess the potential susceptibility of Eastern Forests to *Phytophthora ramorum*. In this example, Morin *et al.* (2003) used fuzzed FIA field plot locations. Coulston *et al.* (2003) used ordinary kriging to predict potential ozone injury at FIA plot locations and assess ozone injury risk to ozone sensitive northeastern tree species. This analysis was conducted using the hexagon centers rather than actual FIA plot locations. Coulston *et al.* (2004) used Thiessen polygons to assess air pollution in forest ecosystems of the United States using FIA biomonitoring data with fuzzed plot locations.

The objectives of this study are (1) to examine the influence that plot coordinate manipulations (fuzzing and swapping) have on interpolated surfaces of tree biomass using data from Minnesota as an example, (2) to investigate whether external users could develop meaningful spatial models based on fuzzed and swapped plot locations. To accomplish this we compared the Thiessen polygon, inverse distance squared weighting, and kriging interpolators using the original plot locations. We then compared the three interpolators using fuzzed and fuzzed and swapped replications of the original dataset.

2. Methods

2.1 Fuzzing and swapping

We followed the suggested FIA protocols to fuzz and swap plot locations. The main restriction was that fuzzing and swapping occurred at the county level. Plots from one county were not fuzzed into another county. Likewise, plots from one county were not swapped with plots from another county. This FIA protocol ensured that county level summaries were valid regardless of the plot coordinate manipulations.

Fuzzed plot coordinates were generated for each of the 3735 currently measured plots in Minnesota (Figure 2). To accomplish this, we randomly selected 5 percent of the plots and fuzzed their coordinates by randomly generating an offset distance and randomly generating an offset azimuth. The offset distance was between 0 km and 1.6 km and the offset azimuth fell between 0° and 360°. For the remaining 95 percent of the of the plots, the offset distance was between 0 km and 0.8 km and the offset azimuth was between 0° and 360°. If the random offset placed the plot in a different county then another random offset was generated for that plot. We created 30 replications of fuzzed plot locations.

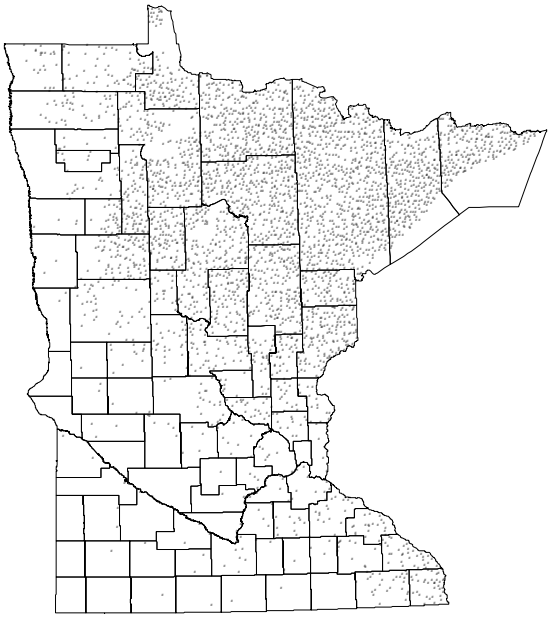

Figure 2. Fuzzed locations of the 3735 FIA plots used in this analysis.

We used 25 of the fuzzed plot replications to generate the fuzzed and swapped replications. Five replications of fuzzed plot coordinates with each of 1, 5, 10, 15, and 20 percent swap were created. To swap the desired percentage of plots (*e.g.* 20 percent) we selected all plots within a county. Twenty percent of the plots were randomly selected and added to a list. Suppose the list was indexed from 0 to n-1 where n was the total number of plots in the list and i was the index number. Then each i plot was swapped with the (n-1) – i plot in the list for $n > 1$. This swapping was done for each county for each level of swapping (1, 5, 10, 15, 20 percent). In summary, we had 5 replications of fuzzed plot coordinates, 5 replication of fuzzed and 1 percent swapped plot coordinates, 5 replication of fuzzed and 5 percent swapped plot coordinates, 5 replication of fuzzed and 10 percent swapped plot coordinates, 5 replication of fuzzed and 15 percent swapped plot coordinates, and 5 replication of fuzzed and 20 percent swapped plot coordinates.

2.2 Interpolation techniques

Thiessen polygons (NN), also known as polygonal declustering, and inverse distance weighting (IDW) were the simplest interpolation techniques we implemented (see Isaaks and Srivastava 1989 for *e.g.*). To interpolate a

surface from point data using Thiessen Polygons, the area of influence each point represents is determined. For example, in Figure 1 the hexagonal cells are the area of influence for each plot of the triangular grid. The Thiessen polygons are then assigned the same value as the point they represent. When this technique is used to predict a point, it is simply the nearest neighbor method. Inverse distance weighting is also a nearest neighbor method and a low-pass filter in the context that the surface created is relatively smooth. Adjusting the power to which the separation distance is raised and/or the number of nearest neighbors will influence the smoothness of the interpolation. We used both the NN and IDW (based on the twelve nearest neighbors) interpolation techniques to predict biomass (t ac^{-1}) in Minnesota based on the real coordinates and each fuzzed and swapped replication.

Kriging is also a low-pass filter however this interpolation technique requires several steps to implement. These steps include computing the empirical semi-variogram, modeling the empirical semi-variogram, and applying the kriging equations. Second order stationarity (constant mean and variance) is a requirement when using this technique. We used a square-root transformation to normalize the biomass data, calculated the empirical semi-variogram, and checked for anisotropy. We selected the power model based on the form of the empirical semi-variogram.

$$\gamma(h)=C_1(h)^a \quad (1)$$

where,
$\gamma(h)$ = the semi-variance at distance h
C_1 = model parameter related to the total semi-variance
a = model parameter related to the range of autocorrelation.

We fit the power model using weighted non-linear regression where the weight was inversely proportional to distance and semi-variance. The logic behind this weighting was that small semi-variance values near distance 0 have the most importance when kriging. This is similar to the weighing proposed by Cressie (1985). We used kriging and a power covariance model to interpolate biomass (t ac^{-1}) in Minnesota based on the real coordinates and each of the 30 fuzzed and swapped replications.

2.3 Comparing the results

Cross-validation was used to examine each interpolation technique (NN, IDW, and kriging) using actual plot locations. Cross-validation is a standard statistical technique where each observation is sequentially removed from the dataset and an estimate is calculated for the observation that was removed. To compare the results we used

$$x_p=1/n(\Sigma(v_o-v_p)^2) \quad (2)$$

where,
x_p = the average squared prediction error
n=number of observations
v_o=observed value
v_p=predicted value.

In the case of the fuzzed and fuzzed and swapped replications, we predicted a value at each actual plot location for each replication and interpolation technique. We then used equation (2) to compare observed versus predicted values for each of the interpolation techniques.

3. Results

As expected, the kriging interpolator performed better than either the IDW or NN methods (Figure 3) using the real coordinates. IDW had a similar but slightly higher x_p. However, both the kriging and IDW displayed bias by over-predicting small values and under-predicting large values (Figure 4). NN did not display this bias but was the worst of the interpolators examined using cross validation.

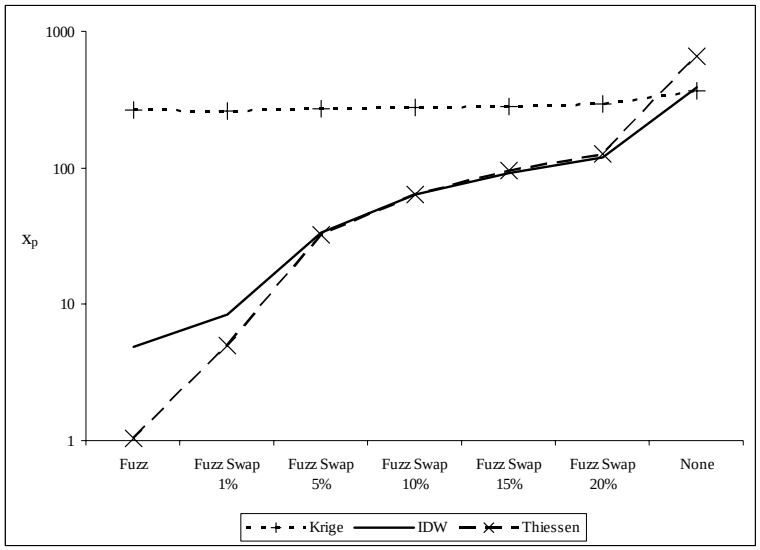

Figure 3. The average squared prediction error (x_p) for each interpolation technique based on the amount of fuzzing and swapping. Results for the real plot locations are labeled "none". The x_p value for the "none" category represents how well the interpolation technique predicted unknown value. All other categories represent how closely each interpolation technique estimated the biomass at the actual plot location. The values of the "none" category are not comparable with the values of the other categories.

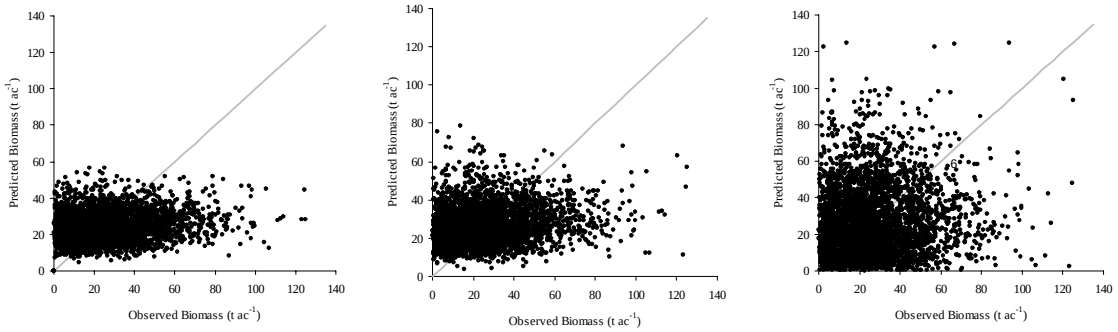

Figure 4. Observed versus predicted values of forest biomass (t ac^{-1}) from the cross-validation analysis using the real plot locations for (a) kriging, (b) IDW, and (c) NN.

Using the fuzzed and fuzzed and swapped replications to predict values at each actual plot location we found that the best interpolator depended on the amount of coordinate manipulation. For example, when only fuzzing was done, the NN interpolator had the lowest x_p but when the fuzz with 15% swap scenario was examined, the IDW performed better (Figure 3). The NN interpolator performed better for all scenarios except fuzz with 15% and 20% swap. Interestingly, the kriging interpolations did not perform as well as the NN and IDW interpolations for any amount of fuzzing and swapping.

4. Discussion

The bias of the interpolated biomass values created using IDW and kriging on the original data (no coordinate manipulation) was somewhat expected. We examined why this occurred using the empirical variogram created with the real plot coordinates (Figure 5). The power model without a nugget effect was used to model the empirical variogram however, if we had used another model (*e.g.* exponential, spherical, Gaussian) we would have specified a nugget effect. In variogram terminology, a nugget effect is used to explain micro-scale variation in the data and the sill refers to the total semi-variance which is approximately the variance of the dataset. The difference between the nugget and the sill represents the semi-variance explained by distance. If we had used an

exponential model with a nugget effect, the nugget would have been approximately 2.5 and the sill would have been 3.2. This means that only 22% of the total variance in the dataset was attributable to distance. When the empirical variogram displays a linear structure with 0 slope the best estimator is the average. Because of the high frequency spatial pattern, and the limited amount of information provided by the variogram, both the IDW and kriging interpolators tended to predict values close to the average for each location. This in turn over predicted low values and under predicted high values.

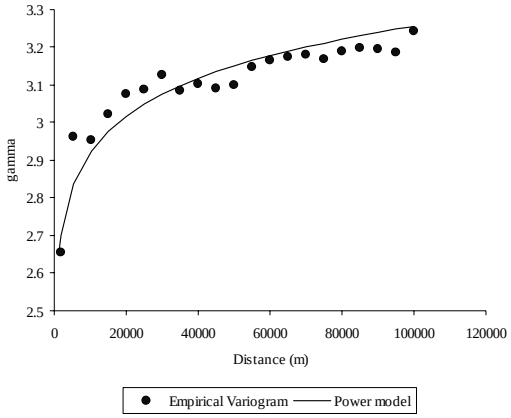

Figure 5. Empirical and modeled variogram calculated using the real plot locations and the square-root of forest biomass.

The fuzzing and swapping did influence the overall spatial structure of the dataset as illustrated by the variogram analysis. On most occasions, the estimated C_1 and a parameters from the fuzzed and fuzzed and swapped replications were outside of the 95% confidence interval of the parameters based on actual plot locations (Table 1). In fact, both parameters were within the 95% confidence interval in only 5 of the 30 replications. As mentioned above, the semi-variance at short lag distance is important when kriging a surface. The coordinate manipulations had the largest influence on semi-variance values at short lag distances (Figure 6). As expected, this was particularly evident for the fuzz and swap 20% scenario but was also visible in all fuzzing and swapping scenarios.

Table 1. The number of times the parameters for the power variogram model (equation 1) of the replications (fuzzed and swapped) fell within the 95% confidence intervals of the parameters calculated using the original dataset (no coordinate manipulations). For each column, the maximum number is 30.

	C_1	a	
95% C.I	(1.788 - 1.9724)	(0.0425 - 0.0531)	
	Number of replications within 95 % C.I.		
	C_1	a	Both parameters
Fuzz	1	1	1
Fuzz Swap 1%	2	2	2
Fuzz Swap 5%	0	1	0
Fuzz Swap 10%	1	1	0
Fuzz Swap 15%	1	1	1
Fuzz Swap 20%	1	1	1
	6	7	5

The NN interpolation technique had the lowest x_p for all scenarios except the fuzz with 15% swap and fuzz with 20% swap. Under these relatively high degrees of swapping, the IDW performed better. While the NN preformed well, the implication is that if you use the Thiessen polygon interpolation method you have a better chance of the predicted value actually occurring in the polygon. However, the resolution of the interpolated surface is only as fine as the sampling intensity. The IDW method can be used to interpolate surfaces at a higher resolution than the sample. In this case, the interpolated value at the actual (even if unknown) plot locations would be relatively close and the interpolated values at un-sampled location would be better than the NN estimate.

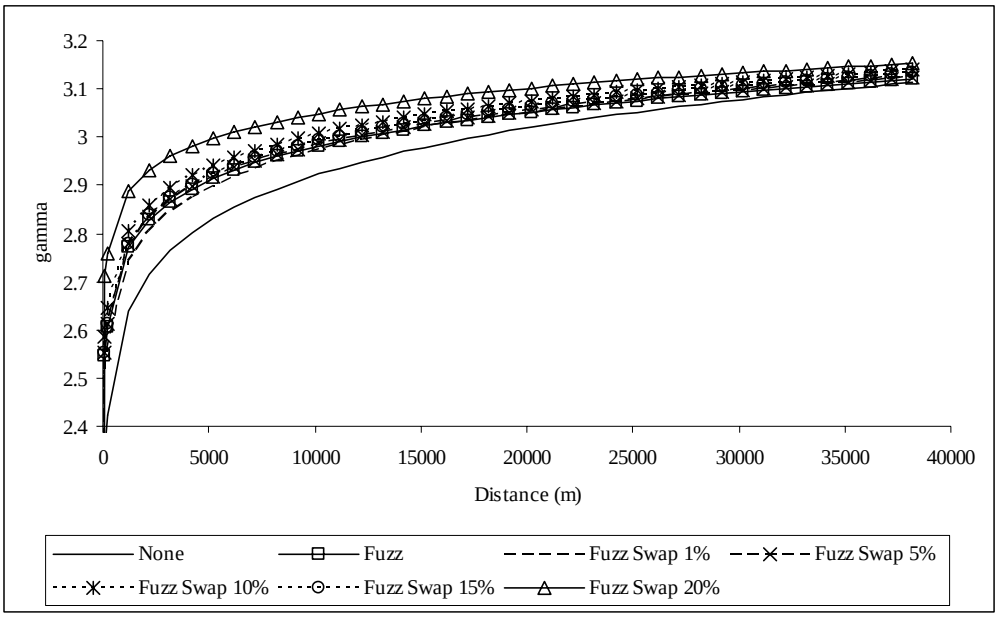

Figure 6. Average variogram for each level of fuzzing and swapping. Results using the real plot locations are labeled "none".

This analysis only addresses one variable (biomass) however there are many FIA plot attributes of interest for interpolation. Some of these attributes may have a more, or less, pronounced spatial structure. To examine the influence blurred plot coordinates has on the larger suite of FIA variables, simulations could be used. For example, one could generate spatial surfaces with known covariance functions and apply the same analysis presented here. We suggest that this topic should be investigated further. However in the case of variables that exhibit a weak spatial structure, such as forest biomass, we suggest the IDW interpolator. We have shown that as the degree of fuzzing and swapping increased the usefulness of FIA program for spatial modeling decreases. To make the program more useful to users and maintain plot confidentiality, perhaps the best alternative is to have FIA generate the interpolated surface. However the resources and infrastructure are currently limited for this to happen on a production scale.

References

Brand, G. 2004. Forest Inventory and Analysis Sampling Hexagons. FIA fact sheet series. Online url: [http://fia.fs.fed.us/library/Factsheets/Sampling_hexagons.doc]

Cressie, N. 1985. Fitting Variogram Models by Weighted Least Squares. *Mathematical Geology* 17 pp. 563-586.

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

Coulston, J.W., Riitters, K.H., Smith, G.C. 2004. A Preliminary Assessment of Montreal Process Indicators of Air Pollution for the United States. *Environmental Monitoring and Assessment*. In Press.

Isaaks, E. H. and Srivastava, R. M. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, NY, 561 p.

Morin, R.S., K.W. Gottschalk, and A.M. Liebhold. 2003. Potential Susceptibility of Eastern Forests to Sudden Oak Death, *Phytophthora ramorum*. 2003 Forest Health Monitoring working Group Meeting. Monteray, CA. Online url: [http://www.na.fs.fed.us/spfo/fhm/posters/posters03/sod.pdf]

Reams, G.A., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F., Moisen, G.G. In review. The forest inventory and analysis sampling frame. In Bechtold, W.A., and Patterson, P.L. (eds). The enhanced forest inventory and analysis program: national sampling design and estimation procedures. General Technical Report. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station.

Smith, W.B. 2002. Forest inventory and analysis: a national inventory and monitoring program. *Environmental Pollution* 116, pp. 233-242.

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