Strategies for Preserving Owner Privacy in the National Information Management System of the USDA Forest Service's Forest Inventory and Analysis Unit

Andrew Lister, Charles Scott, Susan King, Michael Hoppus, Brett Butler, and Douglas Griffith[1]

Abstract.—The Food Security Act of 1985 prohibits the disclosure of any information collected by the USDA Forest Service's FIA program that would link individual landowners to inventory plot information. To address this, we developed a technique based on a "swapping" procedure in which plots with similar characteristics are exchanged, and on a "fuzzing" procedure in which the geographic locations of the plots are randomly perturbed by 805 m. A simulation experiment was performed to assess the effects of fuzzing and swapping. Our results indicate the procedures can provide meaningful information and comply with the law. Further refinements of the technique are ongoing.

The USDA Forest Service's Forest Inventory and Analysis (FIA) program is responsible for conducting a national forest inventory (Gillespie 1999). FIA uses a network of tens of thousands of ground plots to collect information on the quantity, quality, composition, location, and other characteristics of the forests and on land ownership and use on these plots. Many of the forested plots are or soon will be georeferencedusing a global positioning system (GPS), making the data in the FIA database appealing to land managers and scientists interested in using FIA data in a spatial context. Historically, FIA did not divulge exact locations of ground plots to protect landowner privacy and to protect the integrity of the sample. FIA attempted to accommodate data consumers and adhere to the existing security policies by performing in-house analyses.

In 2000, the Department of the Interior and Related Agencies Appropriations Act (H.R. 3423) amended the Food Security Act of 1985 (H.R. 2100) to include FIA in a list of activities that may not make data available to the public if the owner of the land on which the data were collected can be identified. Since the FIA data are referenced with GPS, and ownership maps are freely available to the public in county tax offices, making public the plot data with GPS or digitized location is tantamount to revealing the owner's name and thus violating the law.

In addition to addressing legal concerns, maintaining privacy of the plot locations is essential to FIA's mission. If the plot location were freely available, individuals could either intentionally or unintentionally alter the ecological conditions on the plot, impacting the integrity of data that are collected the next time the plot is measured (in 5 to 10 years). Furthermore, the value of the data is degraded if it is felt that land managers might intentionally alter land management around FIA plots to affect (or avoid affecting) the survey.

Nonetheless, FIA wants to assist users in utilizing the spatial nature of the FIA data while preserving privacy. To reach this goal, we developed a technique whereby the plot coordinate data are slightly altered (fuzzed) and some of the plot data are exchanged (swapped). The purpose is to maintain the functional value, or "ecological signal" of the data while introducing enough uncertainty to decouple the plot-landowner relationship. We then tested the effects of this fuzzing and swapping on the calculation of average board-foot volume within circles of various sizes. The goal of the experiment was to characterize the distribution of errors that data consumers might get when using fuzzed and swapped data.

Methods

The geographic location data collected on 2,037 plots in Maine between 1999 and 2001 were fuzzed using ArcView GIS software (ESRI, Redlands, CA 92373) such that each "new" plot

[1]Forester, Project Leader, Operations Research Analyst, Research Forester, Supervisory Forester, and Forester, respectively, U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA 19073. Phone: 610–557–4038; e-mail: alister@fs.fed.us.

site was located on land within the same county in a random direction by up to 805 m from its original location.

To perform the swap, forested plots on private land were placed into groups based on ownership: forest industry, nonindustrial corporate, other nonindustrial private, and nonindustrial individual. If there were not at least three unique owners within each group within a county, the groups were combined as follows: forest industry with nonindustrial corporate, and other nonindustrial private with nonindustrial individual. If there were still fewer than three owners, all categories were combined into a single "private lands" category. If there were not at least three owners in this private lands group, adjacent counties were combined until all criteria were met.

From within these groups, 12.5 percent of the plots were chosen for exchange with ecologically similar plots within the same group to produce a 25-percent swap. The Euclidean distance-based similarity measure was calculated using the following equation:

$$\text{Similarity Value} = (\text{northing}_a - \text{northing}_b)^2 + (\text{easting}_a - \text{easting}_b)^2 + (\text{forest type group}_a - \text{forest type group}_b)^2 + (\text{productivity class}_a - \text{productivity class}_b)^2$$

where *a* is a plot in the group selected for exchange, and *b* is a plot in the group not in the original selection but still in the same group and county.

Smaller values indicate more similarity. The similarity-defining variables for swapping were chosen because they are static; it would be undesirable to swap plots based on characteristics that would likely change between inventories. Northing and easting are Albers Equal Area coordinates in meters, forest type group is an FIA tree species group identification number that ranges from mostly conifers at the low end to mostly deciduous species at the high end, and productivity class is a value measured in the field by FIA crews that is based on site index, which is the relationship between a representative tree's height and its age.[2]

To test the results of this procedure, a simulation experiment was performed. ArcView GIS software was used to create 1,000 randomly located circles with radii of 5, 10, and 20 km in Maine. Within each circle, the average board-foot volume (bfv) of the unperturbed plots and that of the fuzzed and swapped plots was calculated by summing the bfv of each tree located on each plot within the circle. For each circle, the absolute difference (AD) between pre- and post-fuzzed and swapped bfv averages was calculated, and histograms of ADs were constructed. Scatterplots were created to visually assess the relationship between unperturbed averages and fuzzed and swapped averages, and simple linear regressions were calculated for the data in these scatterplots to describe the relationships.

Results and Discussion

The histograms of ADs from the 5-, 10-, and 20-km circles are shown in figure 1. The means and coefficients of variation of the ADs (in parentheses) for the 5-km, 10-km, and 20-km circles are, respectively, 877.3 (179 percent), 478.8 (102 percent), and 207.0 (86 percent). The degree of skewness (Υ) decreased with increasing circle radius (Υ_5=3.2, Υ_{10}=1.4, Υ_{20}=1.3).

There were about 30 plots in the 20-km circles, 8 in the 10-km circles, and 2 in the 5-km circles. The 5-km circles have the largest percentage of ADs in the lowest error category and the largest range, followed by the 10-km and then the 20-km circles. This is because, for smaller circles, the bfv averages in the tails of the unperturbed data's distribution are very susceptible to either no change or extreme change after fuzzing and swapping, leading to either very small or very large ADs. For the larger circles, however, swapping more likely will occur from within a circle than from without, and the smaller perimeter/area ratio lowers the chances of plots being fuzzed into or out of a circle.

The scatterplots of the unperturbed versus the fuzzed and swapped bfv averages are shown in figure 2. The confidence intervals for the parameters of the regression lines whose equations are shown on figure 2 are shown in table 1. The slopes and intercepts of all regression lines (fig. 2) indicate that the bfv averages calculated from fuzzed and swapped plots are overestimated for low bfv averages and underestimated for high bfv averages. The y intercepts of the regression lines

[2] USDA Forest Service. 2000. Forest inventory and analysis national core field guide, volume 1: field data collection procedures for phase 2 plots, version 1.4. USDA Forest Service, internal report. On file at USDA Forest Service, Washington Office, Forest Inventory and Analysis, Washington, DC.

Figure 1.—*Histograms of absolute differences in board-foot volume (bfv) (AD=abs(unperturbed average bfv – fuzzed and swapped bfv)) obtained within circles of 20-, 10-, and 5-km radius. These graphs represent the effects that the fuzzing and swapping procedure had on data retrievals. N=1000.*

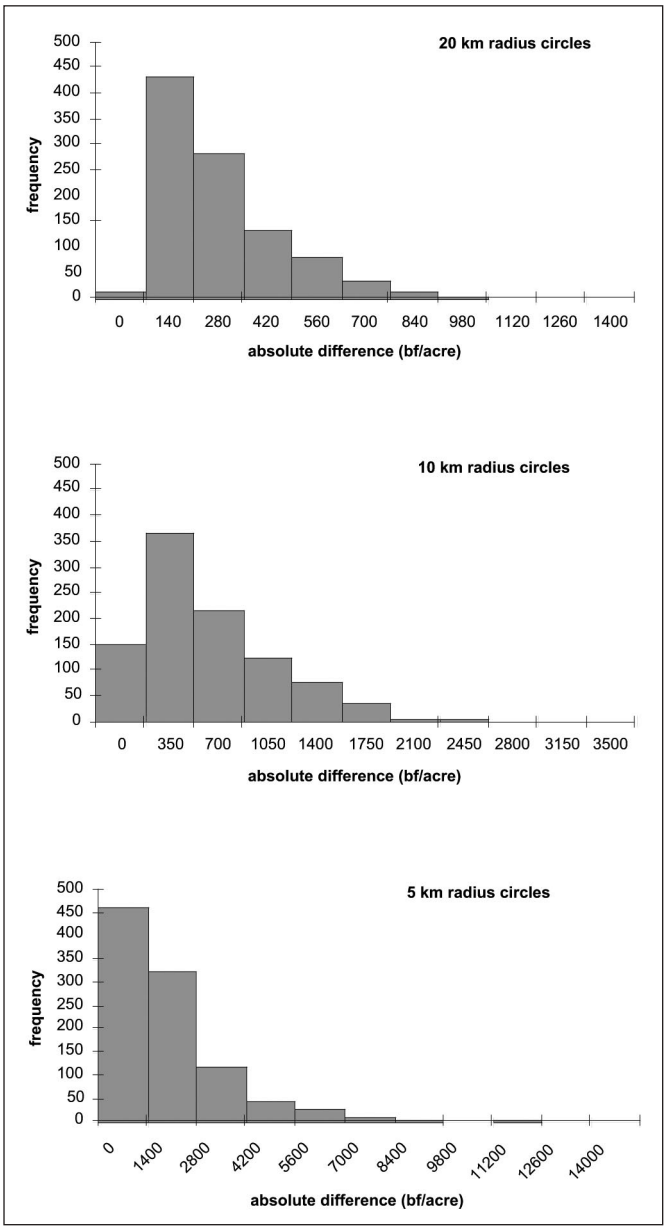

Figure 2.—*Scatterplots and simple linear regression lines and equations of fuzzed and swapped bfv averages versus unperturbed bfv averages within circular areas of various radii. N=1000.*

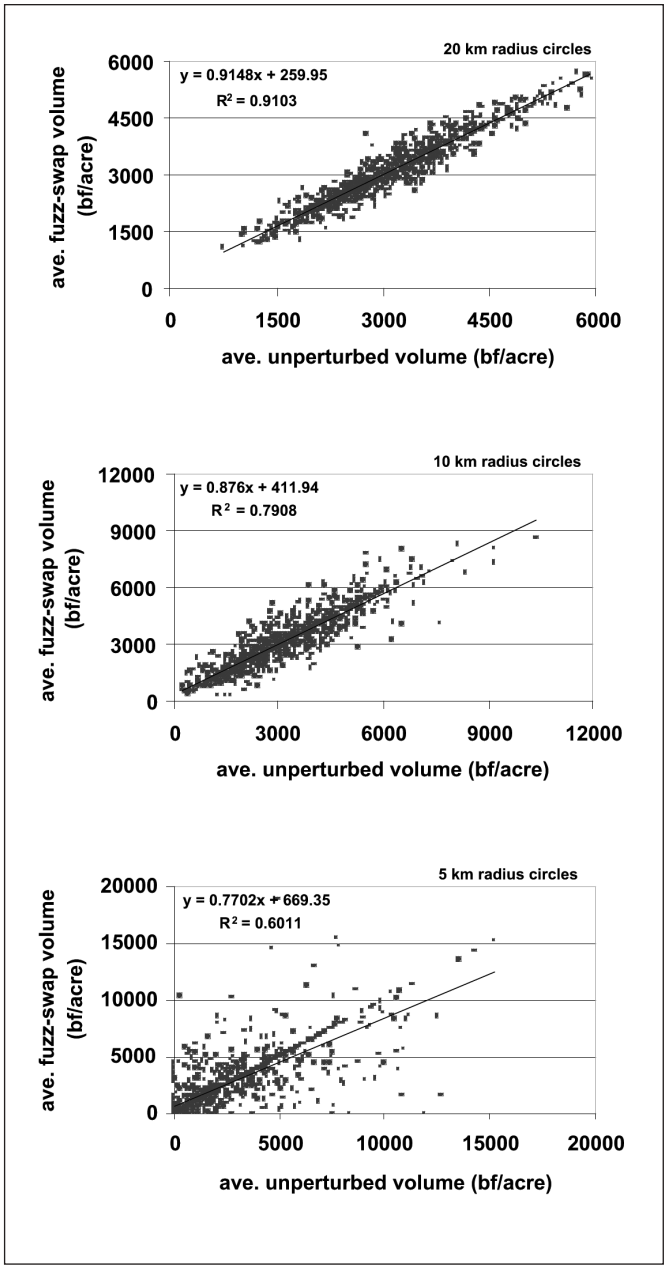

Table 1.—*Ninety-five-percent confidence intervals of the coefficients of the simple linear regression lines that describe the scatterplots of unperturbed vs. fuzzed-swapped plot data for three circle radii*

	5 kilometer			10 kilometer			20 kilometer		
	coeff	lower	upper	coeff	lower	upper	coeff	lower	upper
Slope	0.77	0.73	0.81	0.88	0.85	0.90	0.91	0.90	0.93
Intercept	669.4	514.2	824.5	412.0	314.0	509.8	260.0	202.4	317.5

decrease with increasing circle radius (fig. 2); none of the 95-percent confidence intervals for the intercept contain 0 (table 1). The smaller circles have a larger range of ADs (fig. 1); this forces the y intercept of the regression lines away from zero for the smaller circles. It is noteworthy that there are no zero values found in the scatterplots for the 10- or 20-km circles. There are, however, 1,000 points defining the trajectory of the regression line, making us feel comfortable with our use of the y intercept as a statistic that describes these scatterplots.

All of the 95-percent confidence intervals for the slopes of the regression lines fall below 1 (table 1). The slopes and R^2 values for the regression lines approach 1 with increasing circle radius (fig. 2). This is because, for larger circles, the averaging effects of the fuzzing and swapping tend to act uniformly throughout the entire distribution of values, lowering the variance of the ADs and the deviation of the regression line from a slope of 1.

Across all circle sizes, the bfv averages in the tails of the distribution will always tend toward the sample mean after fuzzing and swapping occurs. In general, the bfv average of a plot moving into a circle with a bfv average in one of the tails of the distribution will be nearer to the sample mean value than it will be to that of the other plots in that circle.

Our results might be difficult to generalize. For example, there is no guarantee that other plot attributes will have the same or less variability than average bfv. Larger data retrievals will be less subject to large fluctuations in summary values than will small ones because a smaller percentage of the total number of plots will be affected by fuzzing. Likewise, data retrievals within different shaped areas might affect summaries more than circular retrievals due to the effects of the geometric complexity of landscape patterns. However, we see no reason to believe that the same principles that governed our current results will not hold with other variables.

In conclusion, the fuzzing and swapping technique outlined here shows great promise. It was conceived as a way to provide useful data to interested parties outside of FIA without violating the law, compromising the ecological integrity of the plots, or introducing concerns about treatment bias. An effort to maintain the functional value of the data is made by conducting geographic fuzzing within a short distance and by swapping plots with similar ecological conditions. The results of our simulation experiment suggest that for average bfv within user-defined circular areas of various sizes, the functional value of the data is kept relatively high, i.e., the fuzzing and swapping technique does not change the fundamental quality of the data dramatically. The functional value is highest for retrievals containing more plots. The fuzzed and swapped data will be most useful for consumers interested in creating summaries over large areas, or for those interested in producing their own coarse-scale graphical representations of the occurrence of FIA-measured attributes. Furthermore, correlative studies with other spatial layers also can be conducted as long as the analyst understands the impact of the slight loss of the data's functional value.

Literature Cited

Gillespie, A.J. 1999. Rationale for a national annual forest inventory program. Journal of Forestry. 97(12): 16–20.