Adding uncertainty to forest inventory plot locations: effects on analyses using geospatial data

Alexia A. Sabor, Volker C. Radeloff, Ronald E. McRoberts, Murray Clayton, and Susan I. Stewart

Abstract: The Forest Inventory and Analysis (FIA) program of the USDA Forest Service alters plot locations before releasing data to the public to ensure landowner confidentiality and sample integrity, but using data with altered plot locations in conjunction with other spatially explicit data layers produces analytical results with unknown amounts of error. We calculated the potential error from using altered location data in combination with other data layers that varied in mean map unit size. The incidence of errors associated with the use of altered plot locations exhibited a strong inverse relationship to the mean map unit size of the other data sets used in the analyses. For a 30 m × 30 m resolution land cover map, plot misclassification rates ranged from 32% to 66%, whereas only 1%–10% of plots were misclassified for ecological subsection data (mean polygon size 9067 km^2). Housing density data derived from the US Decennial Census (mean polygon size = 5.7 km^2) represented an intermediate condition, with 5%–70% of data points misclassified when altered plot locations were used. These analyses demonstrate the impacts of altering FIA plot locations and represent an important step toward making the FIA database more helpful to a broad variety of end users.

Résumé : Le Service des forêts du département de l'Agriculture des États-Unis modifie la localisation des placettes avant de rendre publiques les données de son programme d'inventaire et d'analyse forestiers pour préserver la confidentialité des propriétaires fonciers et l'intégrité des échantillons. Mais l'utilisation de données provenant de placettes dont la localisation a été modifiée, conjointement avec d'autres couches de données spatialement explicites, produit des résultats analytiques qui comportent des erreurs dont l'ampleur est inconnue. Nous avons calculé l'erreur potentielle reliée à l'utilisation de données de localisation modifiées combinées à d'autres couches de données dont la taille moyenne de l'unité cartographique était variable. Il y avait une forte relation inverse entre l'incidence des erreurs associées aux placettes dont la localisation avait été modifiée et la taille de l'unité cartographique des autres données utilisées dans les analyses. Pour une carte de couverture végétale avec une résolution de 30 m × 30 m, le taux d'erreur de classification variait de 32 % à 66 % alors que pour les données de sous-sections écologiques dont la taille moyenne des polygones était de 9067 km^2, seulement 1 %–10 % des placettes étaient mal classées. Les données de densité résidentielle dérivées du recensement décennal américain, dont la taille moyenne des polygones est de 5,7 km^2, représentent une situation intermédiaire avec 5 % – 70 % d'erreur de classification lorsque les placettes dont la localisation a été modifiée étaient utilisées. Ces analyses démontrent les impacts de la modification de la localisation des placettes du programme d'inventaire et d'analyse forestiers et constituent un pas important pour rendre la base de données de ce programme plus utile à une grande variété d'utilisateurs.

[Traduit par la Rédaction]

Introduction

As contemporary organizations gather, analyze, and share large quantities of personal and household data, maintaining the privacy of their subject groups is becoming increasingly important (Muralidhar and Sarathy 2005). In particular, the confidentiality of personal information collected electronically by businesses, the medical industry, and the government has become a topic of great concern (Chen and Rea 2004; O'Herrin et al. 2004). Both private and public institutions increasingly struggle to balance their obligation to protect the privacy of the individuals who are the source of these data against users' needs for accurate information (Domingo-Ferrer et al. 2004). To minimize the possibility of disclosing personal information, data-collecting institutions and agencies may apply a variety of masking procedures (Brand 2002; Domingo-Ferrer et al. 2004; Lechner and Pohlmeier 2004). Data masking necessarily involves some information loss, but the magnitude of such losses and their potential effects on the accuracy of data analyses are unknown.

Received 20 October 2006. Accepted 8 March 2007. Published on the NRC Research Press Web site at cjfr.nrc.ca on 23 November 2007.

A.A. Sabor[1] and V.C. Radeloff. Department of Forest Ecology and Management, University of Wisconsin, 1630 Linden Drive, Madison, WI 53706, USA.
R.E. McRoberts. North Central Research Station, USDA Forest Service, 1992 Folwell Avenue, Saint Paul, MN 55108, USA.
M. Clayton. Department of Statistics, University of Wisconsin, 1210 Dayton Street, Madison, WI 53706, USA.
S.I. Stewart. North Central Research Station, USDA Forest Service, 1033 University Avenue, Suite 360, Evanston, IL 60201, USA.

[1] Corresponding author (e-mail: aasabor@wisc.edu).

doi:10.1139/X07-067

The Forest Inventory and Analysis (FIA) database of the USDA Forest Service provides an excellent case study of the conflicts that arise as agencies attempt to balance users' needs while maintaining data privacy and sample integrity. FIA collects georeferenced data on vegetation and other natural features on approximately 128 000 forested plots, providing the only comprehensive source of inventory information on private and public forest land in the United States. Data collected on each plot from the 1970s to the present are available in tabular format on the Internet (http://www.fia.fs.fed.us/tools-data/data/). This database is of great potential value to land managers, consultants, researchers, and others interested in the scale and pattern of forest change over time and space. For example, FIA data have been used to examine rates of timber harvest (Munn et al. 2002), monitor the effects of climate change (Iverson and Prasad 1998; Stolte 2001), predict tree species distribution (Schwartz et al. 2001), and assess damage caused by natural disasters (Faust et al. 1994).

Although the geographic coordinates and landowner information included in the FIA database allow spatially explicit analyses, the program has long been concerned that disclosing precise plot locations could compromise sample integrity. First, such disclosure may attract other activities that either intentionally or unintentionally affect plot composition (e.g., damaged trees, trampled vegetation, and compacted soils), thereby altering inventory results. Second, disclosures of exact plot locations may make public certain proprietary information on growth and yield or management practices, resulting in landowners' refusing to allow repeated FIA assessments on their property (McRoberts et al. 2005). In 2000, this concern was formalized when the US Congress, as part of the Interior and Related Agencies Appropriations Act (H.R.3423), mandated that FIA plot location and ownership data receive confidential treatment. This language prevents FIA from disclosing sample locations to individuals outside the program in such a way that individual land ownership and other proprietary information could be determined with certainty. Moreover, the risk of revealing landowners' personal information has grown as the FIA program increasingly works with state agencies, universities, other federal agencies, and contractors to implement fieldwork, analysis, reporting, and monitoring.

Because of the new legislation and the variety of program partners, a new policy regarding the direct or indirect release or disclosure of personal information pertaining to plot ownership had to be developed. Thus, the Web-available plot locations released by FIA are now altered in two ways. The majority of FIA plots undergo perturbation, in which the plot coordinate data are altered but still are located within a 1.6 km radius of the true plot location. A much smaller subset of privately owned plots also undergoes swapping, in which the plot location data are first perturbed and then exchanged with data from other plots similar in both ownership and ecological condition (Lister et al. 2005). Users do not have any way of discerning either the extent to which plot locations have been perturbed within the 1.6 km radius or exactly which plot locations may have undergone swapping.

The FIA program's intent is to maintain the ecological validity of its data while decoupling plot–landowner information by adding uncertainty to plot locations. Prior studies concluded that perturbing and swapping have minimal effects on analyses of variables included in the FIA database if the area of interest (AOI) is large enough. For example, a study by Lister et al. (2005) showed that adding uncertainty to FIA plot locations had steadily decreasing effects on multiplot estimates of board foot volume (1 board foot = 2.359 737 dm^3) as circular AOIs increased from 5 to 20 km in radius. McRoberts et al. (2005) similarly found that perturbing and swapping had negligible effects for design-based estimation of forest attributes included in FIA when the radii of circular AOIs exceeded 30 km. Many users, however, are interested in examining the relationships between FIA data and other spatially explicit data, either in raster format or containing irregularly shaped polygons, on finer scales that reflect typical private ownerships or correspond to community interests in political or economic activities. Errors incurred when conducting such analyses using perturbed FIA plot location data may or may not be similar to those incurred when using circular AOIs or when examining only data included in the FIA database. For example, Coulston et al. (2006) found that the extent to which perturbed FIA plot locations influence the development and accuracy of linear regression models is significantly affected by the cell size and spatial autocorrelation among cells of the raster data sets containing the independent variables. Thus, there are many unanswered questions about the utility of altered FIA plot location data for ecological research.

FIA Spatial Data Services (SDS) was created to facilitate the connection between user-generated geospatial data to FIA's true geospatial information to generate derived products that comply with the confidentiality law (USDA Forest Service 2004). Although SDS Centers play a valuable role in meeting the needs of those who wish to use FIA data, many users will find SDS Centers too geographically distant to visit themselves, and SDS will face limitations in their ability to address all users' requests within a reasonable time frame. Therefore, our objective was to quantify the amount of error introduced by using FIA data with altered plot locations in conjunction with other data sets so that researchers can evaluate whether perturbed FIA data are suitable for conducting certain kinds of ecological research or answering management questions. To do so, we chose three data sets that represent a range of map unit sizes, are widely available, and are likely to be useful in answering a broad variety of research questions: (*i*) a 30 m × 30 m land cover classification, (*ii*) census partial block group data with polygon sizes ranging from <0.01 to 1640 km^2, and (*iii*) ecological subsection data with polygon sizes ranging from 469 to 80 600 km^2.

Methods

Study area

The study region includes Michigan, Minnesota, and Wisconsin, an area covering 494 014 km^2. This area is characterized by cold, snowy winters and warm, humid summers with precipitation evenly distributed throughout the year. A gradual transition zone, defined by temperature, frontal movement, and vegetation extends from north-central Minnesota to southeastern Wisconsin and then across the Lower Peninsula of Michigan (Stearns 1997).

Within the study area, approximately 210 000 km^2 are designated as forest land, defined by the USDA Forest Service as a minimum land area of 0.41 ha in size that is at least 10% stocked by forest trees of any size or that formerly had such tree cover and that is not currently developed for a nonforest use (Bechtold and Patterson 2005). Predominant forest types in these three states include maple–beech–birch, aspen–birch, spruce–fir, and oak–hickory (Shifley and Sullivan 2002). This region encompasses a wide range of land cover types, varies greatly in housing density, and includes many ecological subsections, making it particularly suitable for a study of this type.

Data sources

Forest attributes

Our analyses were conducted using FIA plot location data collected from Michigan, Minnesota, and Wisconsin during 2000–2003. FIA field survey plots occur at an intensity of approximately one 0.41 ha plot per 1200 ha in Minnesota and Wisconsin and one plot per 800 ha in Michigan (McRoberts 2006). Field survey personnel collect quantitative and qualitative data on stand condition, land use, ownership, timber volume, tree species, and tree condition (Miles et al. 2001). We used FIADB version 1.7 downloadable data files from the National FIA Database Retrieval Web site (http://ncrs2.fs.fed.us/4801/fiadb/index.htm) to obtain the publicly available perturbed and swapped plot coordinates. Exact plot coordinates were obtained and analyzed at the USDA Forest Service SDS Center in St. Paul, Minnesota.

Land cover

We used the USGS National Land Cover Data (NLCD) derived from Landsat thematic mapper (TM) satellite imagery ca. 1992. The TM images, combined with supporting information such as topography, census, agricultural statistics, soil characteristics, and other land cover maps, have been classified into a hierarchical, 21 class land cover scheme applied consistently over the United States at a 30 m × 30 m resolution (Vogelmann et al. 2001). Eighteen of the 21 cover classes occur within the study region, with major land cover classifications in the study region including herbaceous cultivated (36.3%), forested upland (27.0%), water (20.8%), and wetlands (12.9%). Developed area accounts for 1.7% of land cover, whereas all other cover types account for less than 2.0% of the total surface area of this region.

We also aggregated NLCD data into eight broader categories (e.g., coniferous, deciduous, and mixed forest all became simply “forest”) and calculated the mean area of all patches of contiguous pixels formed by grouping pixels of like classes into homogeneous landscape units using an eight-neighbor rule. When we did this, the mean patch size across all categories was 0.41 ± 96.35 km^2 (mean ± SD).

Housing density

Housing density for the year 2000 was estimated using US Decennial Census data at the partial block group (PBG) level via methods developed by Hammer et al. (2004). Because of concerns about privacy and sampling error, certain data are released only for aggregations of census blocks (block groups). However, block groups are divided by a variety of political boundaries, such as congressional districts and minor civil divisions, which permit division into multiple partial block groups. PBGs have a mean size one-tenth that of block groups and, therefore, provide a much better spatial resolution while including the complete array of population and housing attribute information available at the block group level. Sizes of PBGs in the study region range from <0.01 to 1640 km^2 (mean = 5.7 km^2), while housing densities range from 0.0 to 16 945 units/km^2 (mean = 71.45 units/km^2).

Ecological subsections

The National Hierarchical Framework of Ecological Units divides the country into progressively smaller areas of land and water based on physical and biological characteristics and ecological processes (Cleland et al. 1997). Ecological subsection boundaries are typically delineated by discrete changes in surficial geology (Great Lakes Ecological Assessment 2004). For our analyses we used ecological subsections as delineated by USDA Forest Service ECOMAP (McNab and Avers 1994), which included 90 ecological subsections ranging in size from 469 to 80 600 km^2 (mean = 9067 km^2).

Data analyses

Individual FIA plots were classified as perturbed or swapped based on the linear distance between true and altered plot locations. Those plots having a linear distance of ≤1.6 km between true and altered plot coordinates were categorized as perturbed, whereas those with linear distances >1.6 km were considered swapped. This threshold was chosen based on the maximum extent to which plot coordinates are perturbed and was intended to ensure that the subset of data categorized as swapped did not include any data points that were merely perturbed. Information on land cover type, housing density, and ecological subsection was associated with each true FIA plot location and its perturbed or swapped counterpart in a geographic information system (GIS). Information on land cover, housing density, or ecological subsection was missing from some plot locations because perturbation or swapping moved the plot location outside the study region or into a water body. These plot records were eliminated, as were duplicate plot records, yielding 21 498 records for perturbed plots and 491 records for swapped plots. In general, perturbed and swapped data were analyzed separately. In some instances, however, all 21 989 plots were analyzed together to determine whether there were any differences in the results, because users will not be able to differentiate between perturbed and swapped plot location data when using FIA data available on the Internet. In those instances, both the results from individual and combined analyses are reported.

We graphed housing densities derived using true plot location data against those derived using perturbed and swapped FIA locations on a log–log scale to examine whether there was a linear relationship between these two sets of results. Data points that did not fall on a straight line were examined on a map to determine whether the locations of these plots exhibited any distinctive spatial pattern, such as clustering around public lands or water bodies. In addition, we created residual plots comparing housing densities at true and altered plot locations to assess whether there was any bias to estimates derived using altered plot locations. We used Pearson's correlation coefficients and their *p* values to evaluate the linearity and strength of the relationship between true, perturbed, and swapped locations both by state and over the entire study area and performed paired *t* tests to assess differences in mean housing density values among these data. Correlation coefficients and paired *t* tests for housing densities derived using perturbed plot coordinates were analyzed at the county and ecological subsection level, but such analyses were not possible using swapped plot locations because of insufficient sample sizes.

Because NLCD and ecological subsection data are categorical rather than continuous, we could not perform the same types of analyses on these data as on housing density. Instead, we calculated the percentage of changes that occurred in our results when using either true or perturbed FIA plot locations. In addition, we computed simple kappa (κ) coefficients for NLCD and ecological subsection data. Kappa is a measure of agreement between two categorical data sets that assumes a value between 0 and 1, with 1 being complete agreement between data sets. Kappa is positive whenever the observed agreement exceeds chance agreement, with its magnitude reflecting the strength of agreement (Cohen 1960). NLCD data were analyzed using both the full suite of 18 land cover classes in the study region and the eight broader aggregations of these 18 classes.

Results

In general, the magnitude of the effects of perturbing and swapping FIA plot coordinates strongly depended on the mean map unit size of the comparative data set used in the analyses. Here, we review our results in order from finest to coarsest scale data layers.

National land cover data

When we used the full set of 18 land cover types, 51.5% of perturbed plots and 66.8% of swapped plots exhibited differences in land cover when compared with those derived using true FIA plot coordinates (Tables 1 and 2). When compared with land cover types derived using true plot locations, kappa coefficients for perturbed data (κ = 0.36) and for swapped data (κ = 0.16) indicated considerable lack of agreement between these data sets. Although many of these land cover type changes occurred between closely related cover types, such as coniferous and deciduous forest, the use of aggregated land cover categories still resulted in the misclassification of 32.7% of perturbed plots (κ = 0.50) and 51.7% (κ = 0.36) of swapped plots (Tables 3 and 4), again indicating a strong lack of agreement between these data sets. Upon examining a map, we saw no spatial patterning among plots that changed NLCD categories due to the added uncertainty in the FIA plot coordinate data.

Data analyses combining the perturbed and swapped plot location data sets yielded results similar to those for perturbed data alone, with land cover type changes occurring 51.7% of the time when we used all 18 land cover types and 33.9% of the time when we used aggregated land-cover categories.

Housing density

Graphs of log-transformed housing density for true versus perturbed or swapped coordinates demonstrated a distinctly linear relationship but included considerable scatter (Figs. 1 and 2). When we mapped plot locations for those points that fell along the axes (i.e., exhibited a housing density of zero for either the true or perturbed or swapped locations but not both), it was apparent that many of these plots occurred in areas exhibiting high spatial heterogeneity in housing density, such as PBGs with no houses intermixed with PBGs containing 5–64 houses/km^2.

Residual plots for both perturbed and swapped plot locations (Figs. 3 and 4) exhibited the following lines of points: (*i*) down the *y* axis due to instances in which actual housing densities equaled zero and the value derived from altered plot locations was greater and (*ii*) along the diagonal due to instances where actual housing densities were greater than zero, but the value derived from altered plot coordinates was zero. There was no apparent bias in estimates based on altered plot locations in cases where neither the actual or estimated housing density was greater than zero.

Housing densities derived using true FIA plot locations were highly correlated (Pearson $R = 0.68$, $P < 0.0001$) with those derived using perturbed locations. Approximately 85% of all plots exhibited a housing density difference of ≤0.5 units/km^2 between true and perturbed coordinates. In about 5% of cases, however, differences of >10 housing units/km^2 resulted from using perturbed plot coordinates (Fig. 5). Similarly, housing densities derived using swapped plot locations were significantly correlated with those derived using true coordinates (Pearson R = 0.41, P < 0.0001), although less strongly than those derived using perturbed plot location data. Swapping resulted in nearly 41% of all plots exhibiting a housing density difference of ≤0.5 units/km^2 and approximately 17% exhibiting differences of >10 housing units/km^2 as compared with the housing densities derived using true plot coordinates (Fig. 6).

Housing density data derived using perturbed or swapped coordinates exhibited very similar distributions in relation to true coordinates (Figs. 7 and 8). Paired *t* tests confirmed that there were no significant differences between the means for these data sets either when all plots in the study area were included in the analysis or when tests were performed on data for individual states. When aggregated at the county level, housing densities derived using perturbed FIA plot locations were significantly different at the α = 0.05 level from those derived using true plot locations only 3% of the time. Likewise, when housing density was aggregated to the level of ecological subsection, values derived using true plot locations differed significantly only 7% of the time from those derived from perturbed plot locations. When mapped, there was no clear relationship between the amount or spatial patterning of housing development that could explain the significant differences in housing densities at the county level.

Data analyses combining the perturbed and swapped data sets indicated that approximately 84% of all plots exhibited housing density differences of ≤0.5 units/km^2, a figure comparable to that for perturbed data alone. Similarly, paired *t* tests showed no significant differences between mean housing densities using the combined perturbed and swapped data sets versus true plot location data when examined across

Table 1. Confusion matrix for 18 land cover categories using perturbed Forest Inventory and Analysis (FIA) plot locations (n = 21 498).

	NLCD category for true plot locations																	
NLCD category for perturbed plot locations	OW	LIR	HIR	CIT	RSC	QMG	TB	DF	EF	MF	SL	GH	PH	RC	SG	URG	WW	EHW	User's accuracy
OW	0.06	*				*		0.01	0.01	0.01			<0.01	<0.01			0.01	0.02	0.21
LIR	0.02	0.34	0.26	0.11		*		<0.01	*	*		*	<0.01	<0.01		0.24	*	*	0.32
HIR	*	0.14	0.26	0.10				<0.01		*		*	*	<0.01		0.09		*	0.27
CIT	0.01	0.07	0.13	0.22		*		<0.01	*	*			0.01	<0.01	*	0.06	<0.01	*	0.22
RSC					*			*	*			*							0.40
QMG		*	*			0.39		*	*	*			*	*				*	0.52
TB							0.14	0.01	0.01	0.01	*	*	*	*			<0.01	*	0.13
DF	0.30	0.11	0.14	0.15	*	0.23	0.28	0.49	0.24	0.37	0.34	0.37	0.18	0.09	0.10	0.13	0.18	0.16	0.45
EF	0.09	*		*		*	0.09	0.05	0.31	0.13	*	0.06	0.02	0.01	*	*	0.05	0.02	0.32
MF	0.07	*	*	*		*	0.14	0.07	0.13	0.18	*	0.06	0.01	<0.01	*		0.05	0.02	0.18
SL	*							<0.01		*	*			*			0.01	*	0.08
GH	*	*					*	0.02	0.02	0.01		0.20	0.01	<0.01			0.01	0.01	0.14
PH	0.13	0.08	*	0.09		*	*	0.09	0.03	0.04	*	0.08	0.35	0.15	0.13	0.11	0.05	0.13	0.35
RC	0.10	0.12	0.07	0.18		*	*	0.12	0.06	0.04		0.10	0.34	0.67	0.25	0.18	0.05	0.17	0.69
SG	*	*						<0.01	*				0.01	0.01	0.38	*	*	0.03	0.44
URG	*	0.03	0.07	0.06				<0.01	*	*			<0.01	<0.01	*	0.15	*	*	0.16
WW	0.14	0.04	*	0.03		*	14.00	0.11	0.15	0.17	0.23	0.07	0.04	0.02	0.04	*	0.53	0.17	0.51
EHW	0.05	*	*	0.03		*	*	0.02	0.03	0.02	*	*	0.03	0.02	0.07		0.06	0.25	0.27
Producer's accuracy	0.06	0.34	0.26	0.22	0.67	0.39	0.14	0.49	0.31	0.18	0.09	0.20	0.35	0.67	0.38	0.15	0.53	0.25	

Note: Overall accuracy = 0.49. OW, open water; LIR, low-intensity residential; HIR, high-intensity residential; CIT, commercial or industrial; RSC, rock, sand, or clay; QMG, quarries, mines, or gravel pits; TB, transitional barren; DF, deciduous forest; EF, evergreen forest; MF, mixed forest; SL, shrubland; GH, grassland or herbaceous; PH, pasture or hay; RC, row crops; SG, small grains; URG, urban or recreational grasses; WW, woody wetlands; EHW, emergent herbaceous wetlands.

*Insufficient cell count to ensure confidentiality if reported.

Table 2. Confusion matrix for 18 land cover categories using swapped Forest Inventory and Analysis (FIA) plot locations (*n* = 491).

	NLCD category for true plot locations																		
NLCD category for swapped plot locations	OW	LIR	HIR	CIT	RSC	QMG	TB	DF	EF	MF	SL	GH	PH	RC	SG	URG	WW	EHW	User's accuracy
OW																		*	0.00
LIR								*		*									0.00
HIR								*						*					0.00
CIT																			na
RSC														0.38					0.00
QMG														*					0.00
TB																	*		0.00
DF				*				0.44	0.48	0.46	*	*	0.21		*	*	0.24	*	0.57
EF								0.03	*	*							0.08		0.05
MF							*	0.05	*	*			*				*		0.17
SL								*									*		0.00
GH								*	*			*	*						0.13
PH								0.14		*			0.28	0.16			0.13	*	0.12
RC								0.15	*	0.14	*	*	0.31	0.32		*	0.13	*	0.19
SG								*											0.00
URG								*											0.00
WW								0.11	*	*		*	*	*			0.30	*	0.38
EHW	*							0.04	*	*			*	*			*	*	0.04
Producer's accuracy	0.00	na	na	0.00	na	na	0.00	0.44	0.04	0.11	0.00	0.14	0.28	0.32	0.00	0.00	0.30	0.11	

Note: Overall accuracy = 0.33. See Table 1 for abbreviations. na, not available.
*Insufficient cell count to ensure confidentiality if reported.

Table 3. Confusion matrix for eight aggregated land cover categories using perturbed Forest Inventory and Analysis (FIA) plot locations (n = 21 498).

	NLCD category for true plot locations								
NLCD category for perturbed plot locations	Water	Developed	Barren	Forested	Shrubland	Herbaceous natural	Herbaceous planted	Wetland	User's accuracy
Water	0.06	*	*	0.01			<0.01	0.01	0.21
Developed	0.04	0.53	*	0.01		*	0.01	0.01	0.52
Barren		*	0.23	0.01	*	*	<0.01	<0.01	0.24
Forested	0.46	0.14	0.45	0.63	0.54	0.49	0.13	0.26	0.59
Shrubland	*			<0.01	*		*	<0.01	0.08
Herbaceous natural	*	*	*	0.02		0.20	<0.01	0.01	0.14
Herbaceous planted	0.23	0.21	*	0.18	*	0.19	0.79	0.16	0.81
Wetland	0.20	0.10	0.21	0.15	0.26	0.09	0.06	0.54	0.53
Producer's accuracy	0.06	0.53	0.24	0.63	0.09	0.20	0.80	0.54	

Note: Overall accuracy = 0.67.

*Insufficient cell count to ensure confidentiality if reported.

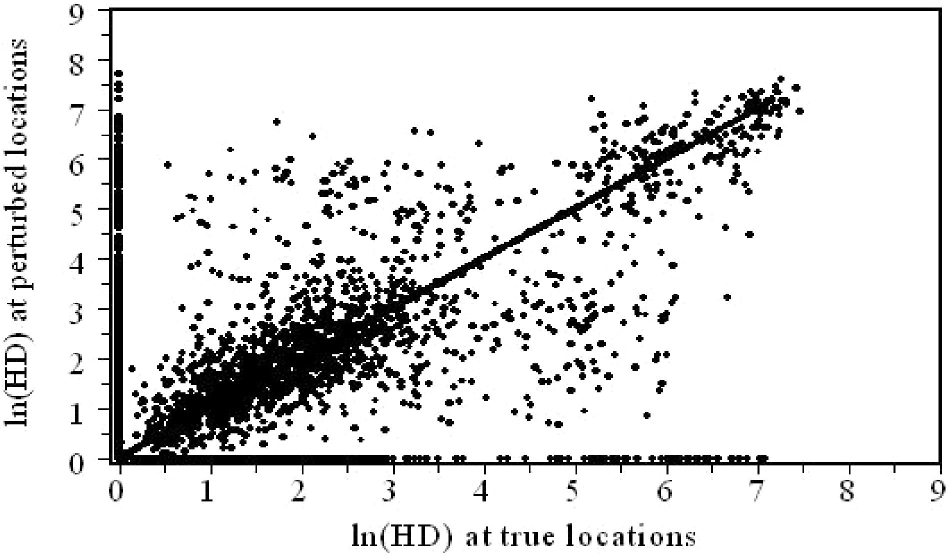
Table 4. Confusion matrix for eight aggregated land cover categories using swapped Forest Inventory and Analysis (FIA) plot locations (n = 491).

	NLCD category for true plot locations								
NLCD category for swapped plot locations	Water	Developed	Barren	Forested	Shrubland	Herbaceous natural	Herbaceous planted	Wetland	User's accuracy
Water								*	0.00
Developed				*			*		0.00
Barren								*	0.00
Forested		*	*	0.54	*	*	0.36	0.37	0.71
Shrubland				*				*	0.00
Herbaceous natural				0.02		*	*	*	0.13
Herbaceous planted				0.28	*	*	0.51	0.27	0.47
Wetland	*			0.15		*	0.11	0.32	0.33
Producer's accuracy	0.00	0.00	0.00	0.54	0.00	0.14	0.51	0.32	

Note: Overall accuracy = 0.48.

*Insufficient data to ensure confidentiality restrictions if reported.

Fig. 1. Housing density for true versus perturbed Forest Inventory and Analysis (FIA) plot locations (log–log scale, n = 21 498).



the study area or by state. Mean housing densities for counties differed significantly at approximately the same rate (3%) as when perturbed data were analyzed alone; however, mean housing density was significantly different from that derived using true plot locations approximately 13% of the time.

Fig. 2. Housing density for true versus swapped Forest Inventory and Analysis (FIA) plot locations (log–log scale, *n* = 491).

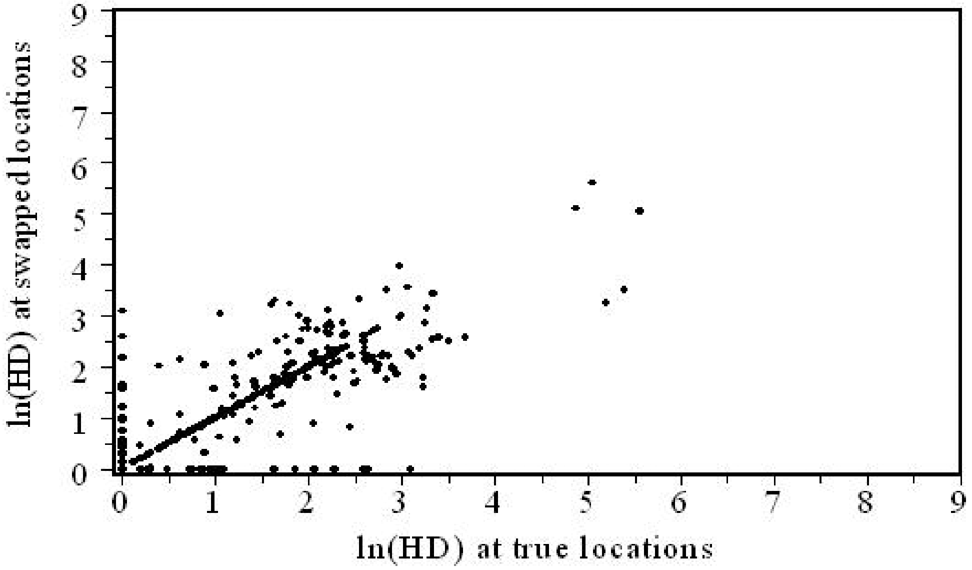
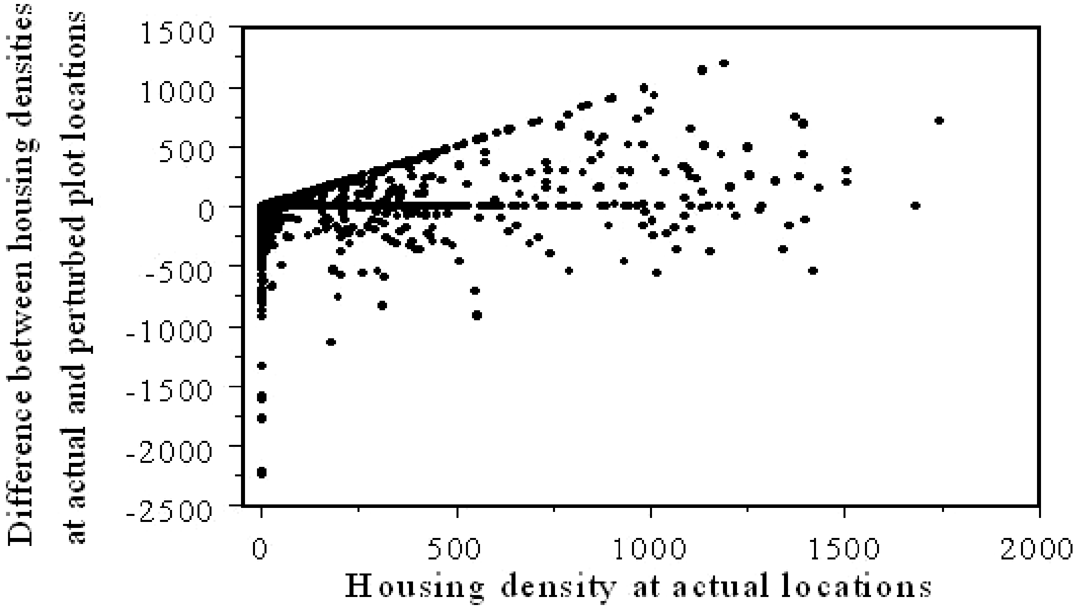


Fig. 3. Residual plot for housing density using true versus perturbed Forest Inventory and Analysis (FIA) plot coordinates (*n* = 21 498).



Ecological subsections

When we used perturbed coordinates, FIA plots changed ecological subsection only 0.5% (*n* = 107) of the time when compared with true plot locations. Among those plots that changed subsection, 27% (*n* = 29) were moved to a different section within the same province and about 6% (*n* = 6) were moved to a different province. Changes in ecological subsection also occurred relatively infrequently when swapped coordinates were used, affecting only 10.3% (*n* = 43) of plots in this sample. Of this 10.3% of affected plots, 14% (*n* = 6) changed sections within the same province, and 21% (*n* = 9) switched to a different province. Simple kappa statistics also indicated strong levels of agreement between ecological subsections derived using true, perturbed, and swapped plot locations, with κ = 0.99 for perturbed data and κ = 0.91 for swapped data. Analyses combining the perturbed and swapped data provided results similar to those obtained using only perturbed data, with ecological subsection changes occurring at a rate of less than 1%.

Discussion

FIA perturbs and swaps plot location data to comply with the law and maintain the ecological integrity of their sample plots while still providing useful data to outside users. Our results suggest that perturbed and swapped FIA plot locations may be used in correlative studies with other spatially explicit data layers having a wide range of polygon shapes and sizes without seriously compromising the quality of the information conveyed in the results. However, the misclassification rates associated with the use of altered plot locations exhibits a strong inverse relationship to the mean map unit size of the other geospatial data sets used in the analyses. Thus, we sug-

Fig. 4. Residual plot for housing density using true versus swapped Forest Inventory and Analysis (FIA) plot coordinates (n = 491).

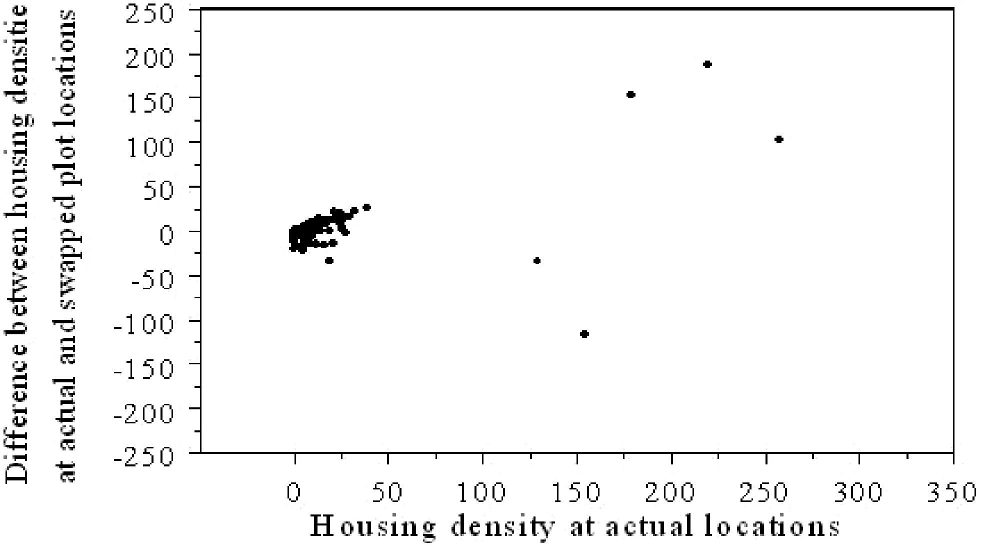
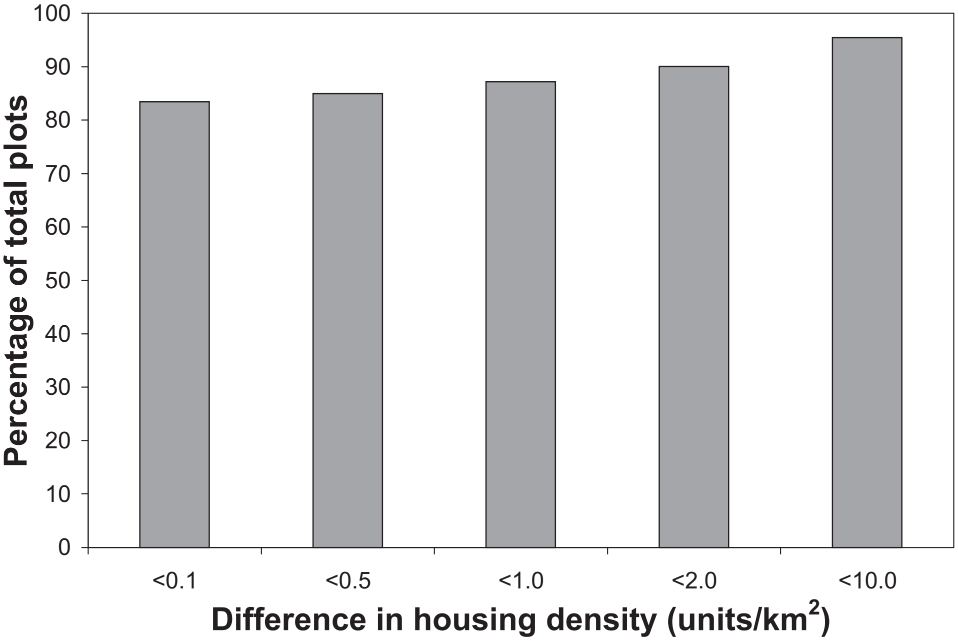


Fig. 5. Cumulative distribution of differences in housing density corresponding to true versus perturbed Forest Inventory and Analysis (FIA) plot coordinates (n = 21 498).



gest that users carefully evaluate the appropriateness of using perturbed and swapped plot locations for any other geospatial data set they wish to use in combination with FIA data.

When using coarse-scale ecological subsection data, we found that relatively few data points were misclassified because of the inclusion of uncertainty in FIA plot location data. Of those plots that were assigned an incorrect ecological subsection, the majority (65%–67%) were assigned to a different subsection within the same ecological section as the correct subsection and usually immediately adjacent to it. Because ecological subsections represent very broad ecotones rather than discrete ecological boundaries (Rowe 1996), the analytical consequences of occasionally misclassifying the subsections in which FIA plots occur are probably negligible in most cases.

Conversely, combining 30 m × 30 m pixel NLCD data resulted in frequent misclassification of land cover type when perturbed or swapped data were used. Even when we combined land cover data into broader categories, many points were still misclassified using perturbed and swapped data. Although the mean patch size increased to 0.41 km^2 when we aggregated contiguous cells with the same broad land cover classification, a size more than 450 times greater in area than a 30 m × 30 m pixel, these data still represent a much finer spatial scale than the other data layers used in this analysis. In addition, even when pixels were aggregated,

Fig. 6. Cumulative distribution of differences in housing density corresponding to true versus swapped Forest Inventory and Analysis (FIA) plot coordinates (n = 491).

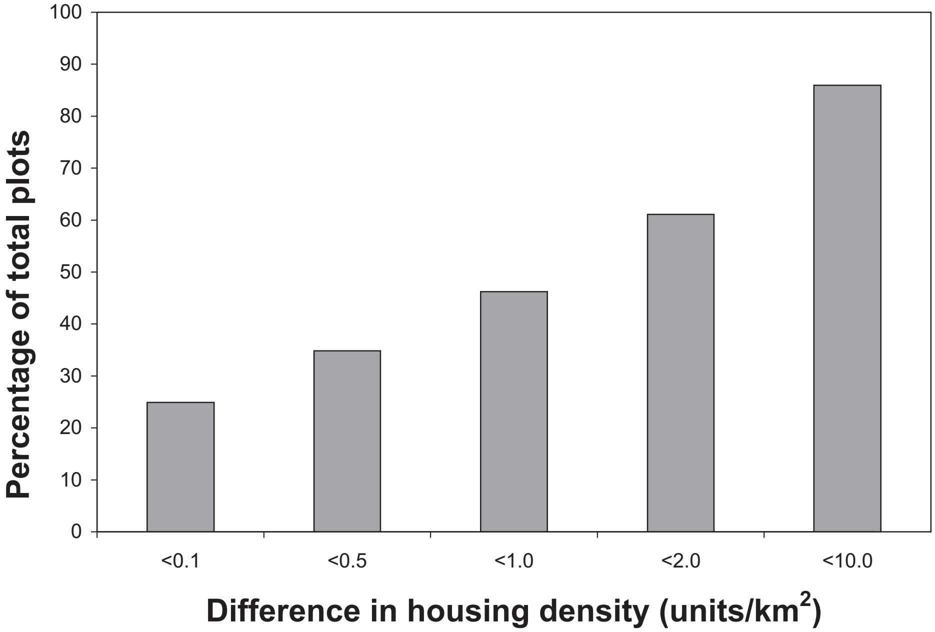
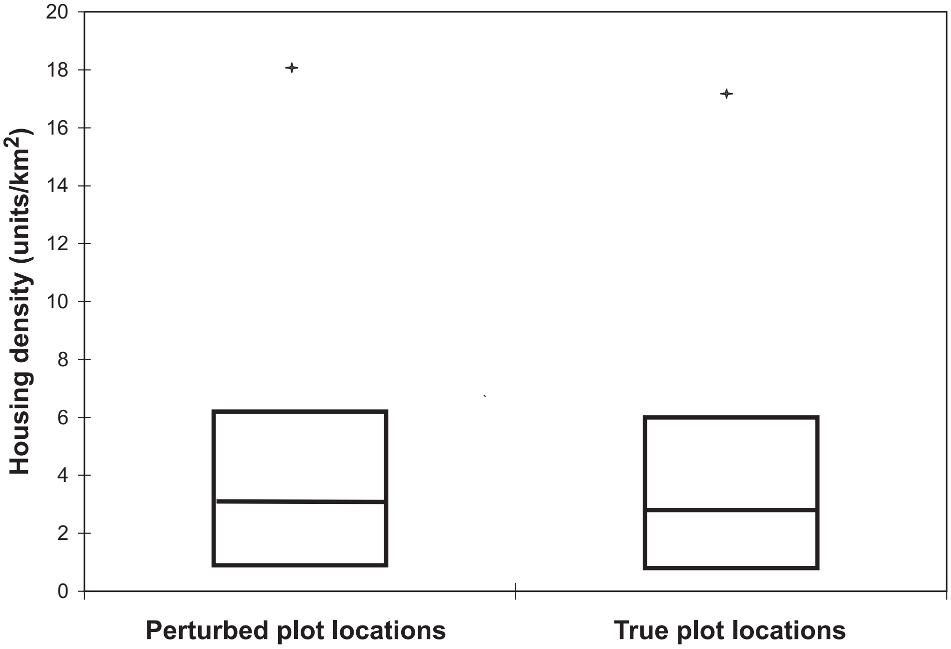


Fig. 7. Box plots of housing density distributions derived using true and perturbed Forest Inventory and Analysis (FIA) plot coordinates (n = 21 498). The box shows the 25th percentile (lower hortizontal line), median (middle hortizontal line), and 75th percentile (upper hortizontal line); +, sample mean.



the incidence of single pixels averaged 44.2% across all land cover categories. Although smoothing these data by merging isolated pixels with their neighbors or by specifying a minimum patch size would most likely result in fewer differences between true and perturbed or swapped locations, the extent to which these practices would affect analytical outcomes is beyond the scope of the present study. Based on our results and the likelihood of relatively high levels of misclassification in fine-resolution data such as NLCD, particularly in highly heterogeneous landscapes (Smith et al. 2003), we do not advise using data with such a small mean map unit size in conjunction with perturbed FIA plot locations.

Housing density represents an intermediate situation, in which a small to moderate number of data points are affected by the inclusion of uncertainty in FIA plot location data. The majority of all FIA plots showed differences of ≤0.5 housing units/km^2 regardless of whether true or altered plot locations were used. These differences were not statistically significant for the overall study region or indi-

Fig. 8. Box plots of housing density distributions derived using true and swapped Forest Inventory and Analysis (FIA) plot coordinates (*n* = 491). The box shows the 25th percentile (lower horizontal line), median (middle hortizontal line), and 75th percentile (upper horizontal line); +, sample mean.

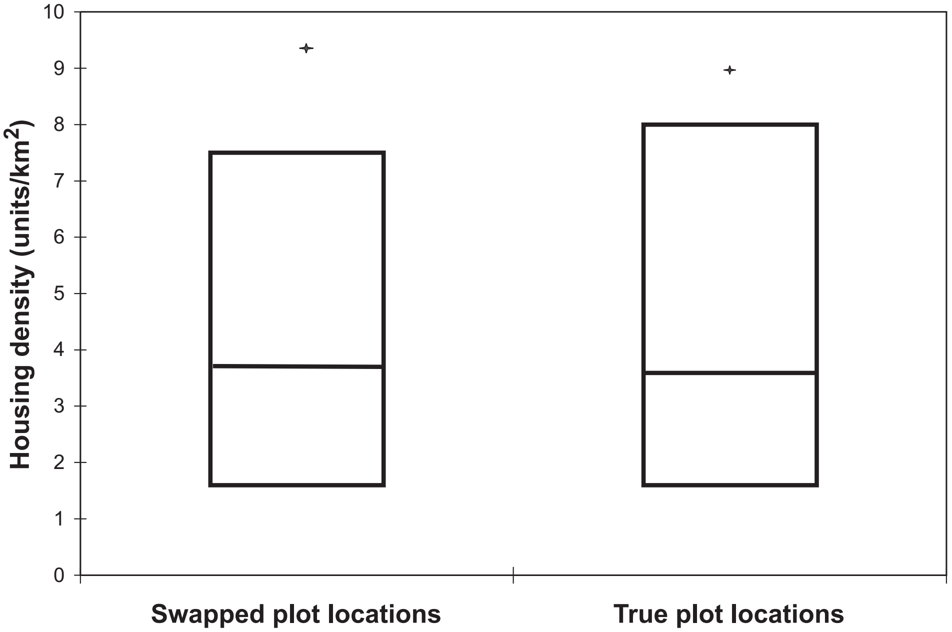
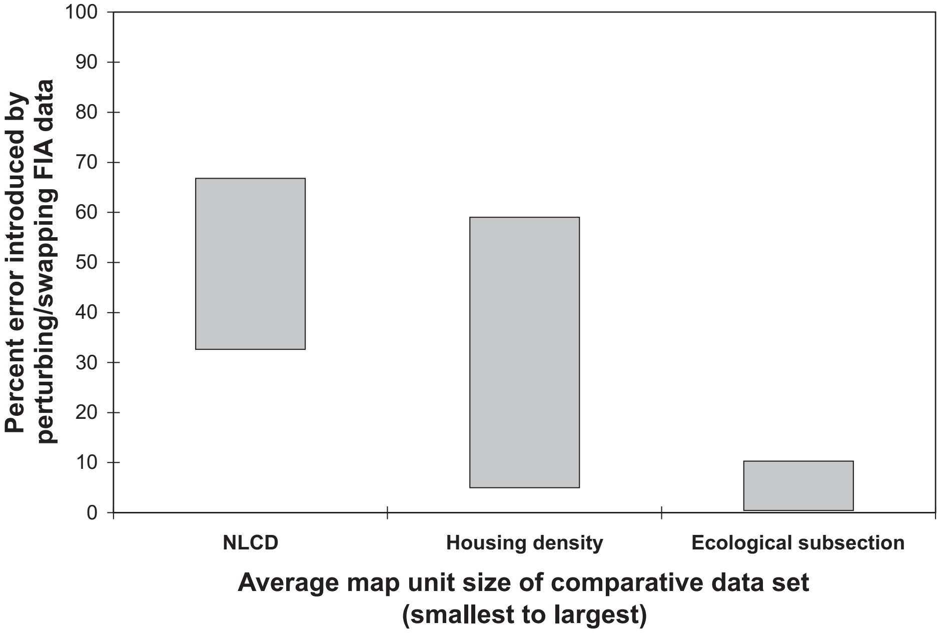


Fig. 9. Summary of error distributions when comparing true and perturbed or swapped Forest Inventory and Analysis (FIA) plot locations to data layers at multiple spatial scales.



vidual states and only infrequently (3%–13% of the time) significant when plots were aggregated at the ecological subsection or county level. Nonetheless, there are a small number (5%–17%) of instances in which using altered locations resulted in differences of >10 housing units·km^{-2}·plot^{-1}. These differences are most likely to occur in regions with high levels of heterogeneity in housing densities, and thus, researchers should use particular caution when associating FIA and PBG data in such areas.

Our results both concur with and differ from those of prior studies of the impacts of altering plot locations on data analyses. Overall, our findings agree with those of Lister et al. (2005), McRoberts et al. (2005), and Coulston et al. (2006) in concluding that analyses using fine-scale data are more likely to reflect the effects of perturbation and swapping in FIA plot location data (Fig. 9). Coulston et al. (2006) indicated that perturbed plot locations should only be used with fine-resolution (30–500 m) raster data with a high degree of contagion (i.e., exhibiting generally clumped patterns of landscape categories; cf. Li and Reynolds 1993; Riitters et al. 1996). Our results reinforce their conclusions, because we found extremely high rates of misclassification when using altered plot locations in conjunction with NLCD data. The small mean patch size and high incidence of single cells even after we aggregated pixels using eight categories indicate that these data do not exhibit high degrees of contagion, thus explaining the minimal improvement in misclassification rates that we observed. Conversely, there were few statistically significant differences between housing densities calculated using altered and true plot locations despite the relatively small mean size (5.67 km^2) of housing PBGs. Because the mean PBG size is much smaller than the minimum circular AOI size that previous studies (Lister et al. 2005; McRoberts et al. 2005) have found necessary to mitigate the effects of altering plot locations, our results suggest that it may be possible to use perturbed and swapped FIA data in conjunction with finer scale data than was previously supposed.

Although perturbed and swapped data exhibit similar overall patterns of error, the magnitude of the error generated using swapped data is generally greater than that from perturbed data. Thus, to obtain a particularly conservative error estimate it would be desirable to treat all privately owned plots as if they had been swapped. On the other hand, analyses in which we combined both perturbed and swapped data generally produced results most comparable to those derived using perturbed data alone. We surmise this is the case because only a small percentage of plots are swapped, and therefore perturbed plots are more important in influencing the outcomes of data analyses. Consequently, we suggest that, in most instances. it will not be necessary for users to treat privately owned plots as a separate case. We also note that, although our threshold distance value ensured accurate categorization of all plots in the subset of swapped data, it is likely that the subset of perturbed data include some plots that are swapped because FIA has no criterion specifying a minimum distance beyond which plots must be swapped. Given that only a small proportion of plots are ever swapped, however, we feel confident that the number of such plots analyzed as perturbed is small and the effects of their inclusion are negligible given the large sample size of perturbed data.

Our results, along with those of Coulston et al. (2006), highlight the importance of landscape configuration and contagion as well as map unit size. Although not a factor explicitly addressed in our study, we found that analyses using NLCD data were not much improved when like pixels were aggregated and that housing density was most likely to be assigned incorrectly in areas where this attribute was very heterogeneous. Landscape pattern is widely recognized as an important component of ecological study (Turner et al. 2001), and thus we suggest users pay particular attention not only to the size of their map units but to the configuration of those units when deciding whether using altered FIA data is appropriate for their purposes.

Our results can provide researchers with an analytic framework to evaluate the sensitivity of their own geospatial data to errors introduced by altering FIA plot location data. For those who wish to use land cover, census, or ecoregion data or other geospatial data sets with similar mean map unit sizes, our results may be directly applicable. In other cases, users can perform simple analyses to determine how much their results would change because of perturbation of plot locations. For example, users could reperturb Web-available FIA plot locations randomly within a 0.8 km radius circle, perform analyses such as we have done, and use these results as a proxy for the difference between true and perturbed plot locations. Alternatively, users could evaluate the likelihood that altered plot locations would affect their analyses by placing a 1.6 km buffer around the plot and calculating the percentage of the buffer area that falls in a different polygon. Plots with a high probability of changing to a dissimilar polygon could then be removed from the data set provided this did not bias the sample.

Finally, we suggest that further research along the lines of the present study and those conducted by Lister et al. (2005), McRoberts et al. (2005), and Coulston et al. (2006) are needed to help maintain confidentiality while continuing to making this valuable database more broadly available to those who require spatially explicit information to answer questions about forest ecology, management, and policy.

Acknowledgements

We thank Mark Hatfield, Geoff Holden, and Mark Nelson for facilitating our data analyses at the USDA Forest Service Spatial Data Services Center in St. Paul, Minnesota. Roger Hammer supplied the housing density data layer. Sherry Holcomb and Jason McKeefry provided technical assistance with the NLCD and housing density analyses. This manuscript greatly benefited from comments by Lucy Burde, John Coulston, Ray Guries, Roger Hammer, Keith Moser, and two anonymous reviewers. We gratefully acknowledge financial support for this study from the USDA Forest Service North Central Research Station, St. Paul, Minnesota.

References

Bechtold, W.A., and Patterson, P.L. 2005. The enhanced Forest Inventory and Analysis program — National sampling design and estimation procedures. USDA For. Serv. Gen. Tech. Rep. SRS-80.

Brand, R. 1997. Microdata protection through noise addition. *In* Inference control in statistical databases: from theory to practice. *Edited by* J. Domingo-Ferrer. Springer-Verlag, Berlin. Lect. Notes. Comput. Sci. 2316. pp. 97–116.

Chen, K.C., and Rea, A.I. 2004. Protecting personal information online: a survey of user privacy concerns and control. J. Comput. Inf. Syst. **44**: 85–92.

Cleland, D.T., Avers, P.E., McNab, W.H., Jensen, M.E., Bailey, R.G., King, T., and Russell, W.E. 1997. National hierarchical framework of ecological units. *In* Ecosystem management applications for sustainable forest and wildlife resources. *Edited by* M.S. Boyce and A. Haney. Yale University Press, New Haven, Conn. pp. 181–200.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Measure. **20**: 37–46.

Coulston, J.W., Riitters, K.H., McRoberts, R.E., Reams, G.A., and Smith, W.D. 2006. True versus perturbed forest inventory plot locations for modeling: a simulation study. Can. J. For. Res. **36**: 801–807. doi:10.1139/x05-265.

Domingo-Ferrer, J., Sebe, F., and Castella-Roca, J. 2004. On the security of noise addition for privacy in statistical databases. *In* Privacy in Statistical Databases. Proceedings of the Computational Aspects of Statistical Confidentiality Project Final Conference, 9–11 June 2004, Barcelona, Spain. *Edited by* J. Domingo-

Ferrer and V. Torra. Springer-Verlag, Berlin. Lect. Notes. Comput. Sci. 3050. pp. 149–161.

Faust, T.D., Fuller, M.M., McAlister, R.H., and Zarnoch, S.J. 1994. Assessing internal hurricane damage to standing pine poletimber. Wood Fiber Sci. **26**: 536–545.

Great Lakes Ecological Assessment. 2004. Ecological classification systems [online]. Available from http://www.ncrs.fs.fed.us/gla [accessed 18 December 2005].

Hammer, R.B., Stewart, S.I., Winkler, R.L., Radeloff, V.C., and Voss, P.R. 2004. Characterizing dynamic spatial and temporal residential density patterns from 1940–1990 across the north central United States. Landsc. Urban Plann. **69**: 183–199.

Iverson, L.R., and Prasad, A.M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecol. Monogr. **68**: 465–485. doi:10.1890/0012-9615(1998)068[0465:PAOTSF]2.0.CO:2.

Lechner, S., and Pohlmeier, W. 2004. To blank or not to blank? A comparison of the effects of disclosure limitation methods on nonlinear regression estimates. *In* Privacy in Statistical Databases. Proceedings of the Computational Aspects of Statistical Confidentiality Project Final Conference, 9–11 June 2004, Barcelona Spain. *Edited by* J. Domingo-Ferrer and V. Torra. Springer-Verlag, Berlin. Lect. Notes. Comput. Sci. 3050. pp. 187–200.

Li, H., and Reynolds, J.F. 1993. A new contagion index to quantify spatial patterns of landscapes. Landsc. Ecol. **8**: 155–162. doi:10.1007/BF00125347.

Lister, A., Scott, C., King, S., Hoppus, M., Butler, B., and Griffin, 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 4th Annual Forest Inventory and Analysis Symposium, 19–21 November 2002, New Orleans, La. *Edited by* R.E. McRoberts, A. Reams, P.C. VanDeusen, W.H. McWilliams, and C.J. Cieszewski. USDA For. Serv. Gen. Tech. Rep. NC-252. pp. 163–166.

McNab, W.H., and Avers, P.E. 1994. Ecological subregions of the United States [online]. Available from http://www.fs.fed.us/land/pubs/ecoregions/ [accessed 18 December 2005].

McRoberts, R.E. 2006. A model-based approach to estimating forest area. Remote Sens. Env. **103**: 56–66.

McRoberts, R.E., Holden, G.H., Lister, A.J., King, S.L., Coulston, J.W., and Smith, W.B. 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. J. For. **103**: 275–279.

Miles, P.D., Brand, G.J., Alerich, C.L., Bednar, L.F., Woudenberg, S.W., Glover, J.F., and Ezzell, E.N. 2001. The Forest Inventory and Analysis database: database description and users manual. Version 1.0. USDA For. Serv. Gen. Tech. Rep. NC-218.

Munn, I.A., Barlow, S.A., Evans, D.L., and Cleaves, D. 2002. Urbanization's impact on timber harvesting in the south central United States. J. Environ. Manage. **64**: 65–76. doi:10.1006/jema.2001.0504. PMID:11876075.

Muralidhar, K., and Sarathy, R. 2005. An enhanced data perturbation approach for small data sets. Decis. Sci. **36**: 513–529. doi:10.1111/j.1540-5414.2005.00082.x.

O'Herrin, J.K., Fost, N., and Kudsk, K.A. 2004. Health insurance accountability act (HIPAA) regulations: effect on medical record research. Ann. Surg. **239**: 772–778. doi:10.1097/01.sla.0000128307.98274.dc. PMID:15166956.

Riitters, K.H., O'Neill, R.V., Wickham, J.D., and Jones, K.B. 1996. A note on contagion indices for landscape analysis. Landsc. Ecol. **11**: 197–202. doi:10.1007/BF02071810.

Rowe, J.S. 1996. Land classification and ecosystem classification. Environ. Monit. Assess. **39**: 11–20. doi:10.1007/BF00396131.

Schwartz, M.W., Iverson, L.R., and Prasad, A.M. 2001. Predicting the potential future distribution of four tree species in Ohio using current habitat availability and climatic forcing. Ecosystems (N.Y., Print), **4**: 568–581. doi:10.1007/s10021-001-0030-3.

Shifley, S.R., and Sullivan, N.H. 2002. The status of timber resources in the north central United States. USDA For. Serv. Gen. Tech. Rep. NC-228.

Smith, J.H., Stehman, S.V., Wickham, J.D., and Yang, L. 2003. Effects of landscape characteristics on land-cover class accuracy. Remote Sens. Environ. **84**: 342–349. doi:10.1016/S0034-4257(02)00126-8.

Stearns, F.W. 1997. Physical environment supporting Lake States forests. *In* Lake States regional forest resources assessment: technical papers. *Edited by* J.M. Vasievich and H.H. Webster. USDA Forest Service, Washington, D.C. pp. 1–7.

Stolte, K.W. 2001. Forest health monitoring and forest inventory analysis programs monitor climate change effects in forest ecosystems. Hum. Ecol. Risk Assess. **7**: 1297–1321. doi:10.1080/20018091095014.

Turner, M.G., Gardner, R.H., and O'Neill, R.V. 2001. Landscape ecology in theory and practice. Springer-Verlag, New York.

USDA Forest Service. 2004. National Forest Inventory and Analysis spatial data services [online]. Available from http://www.fs.fed.us/ne/fia/spatial/index.html [accessed 12 January 2007].

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Van Driel, N. 2001. Completion of the 1990s national land cover data set for the conterminous United States from Landsat thematic mapper data and ancillary data sources. Photogramm. Eng. Remote Sens. **67**: 650–652.