Estimating and Circumventing the Effects of Perturbing and Swapping Inventory Plot Locations

Ronald E. McRoberts, Geoffrey R. Holden, Mark D. Nelson, Greg C. Liknes, Warren K. Moser, Andrew J. Lister, Susan L. King, Elizabeth B. LaPoint, John W. Coulston, W. Brad Smith, and Gregory A. Reams

ABSTRACT

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service reports data and information about the Nation's forest resources. Increasingly, users request that FIA data and information be reported and distributed in a geospatial context, and they request access to exact plot locations for their own analyses. However, the FIA program is constrained by law from disclosing exact locations and, instead, releases only approximate locations. The results of the study indicate that the effects of using approximate locations are negligible for design-based estimates for circular areas of radii greater than approximately 20 mi (32 km). For model-based estimates, the results are less definitive and depend on the modeling technique, the spatial resolution of units on which variables are observed, spatial correlation in the variables, and the quality of fit of the model to the data. Several methods for circumventing the effects of the approximate locations are discussed.

Keywords: estimation, models, correlation, bias

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service conducts comprehensive forest inventories of the United States to estimate the area of forestland; the volume, growth, and removal of forest resources; and the health and condition of the resources. The program's sampling design has an intensity of one plot per approximately 6,000 ac (2,400 ha) and is assumed to produce a random, equal probability sample.

Traditionally, the FIA program has reported estimates of forest attributes for states and counties. Increasingly, however, users request plot data for estimation for their own areas of interest (AOI). If the users' requests include exact plot locations, then precautions must be observed to assure compli-ance with the public law[1] that prohibits disclosure of proprietary information. This law further provides a legal basis for protecting information that would otherwise be available under the Freedom of Information Act of 1966. Disclosure violations could result in criminal penalties including fines of up to $250,000, imprisonment of up to 5 years, or both.

The FIA program has additional concerns related to disclosing plot locations. First, knowledge of exact locations may entice users to visit the plots to obtain additional information, thus artificially disturbing the location by damaging trees, trampling vegetation, compacting soils, or vandalism, all of which contribute to sampling bias. Second, large proportions of plots are located on privately owned lands. The success of the FIA program depends on repeated access to these plots, which is possible only via the voluntary permission of landowners. User visits to plot locations may jeopardize this access. Therefore, to protect proprietary information, preserve the ecological integrity of sampling locations, gain repeated access to plot locations, and retain its credibility as a neutral provider of unbiased forest resource information, the FIA program has established a policy of not disclosing the owners or exact locations of plots.

To comply with the policy, FIA plot locations on private lands made available to the public are perturbed,[2] the locations of some plots are swapped with those of similar plots, and private ownership information is masked. All plot locations are perturbed within circular areas of radii of 1.0 mi (1.6

---

km) centered at the exact locations, although the radii for most perturbations are less than 0.5 mi (0.8 km). Swapping consists of exchanging the locations for a small proportion of close proximity, ecologically similar, privately owned plots. Similarity criteria vary regionally but often include forest type group, stand size, and geographic proximity. Except in unusual situations, perturbing/swapping occurs only within counties so that the county statistics based on exact and perturbed/swapped locations are identical. In addition, precautions are taken to assure that perturbed plot locations are not in inappropriate locations such as large bodies of water. Finally, if a data request does not include at least three unique owners in each ownership category represented, then the ownership categories are collapsed into a more generic category. Although perturbing/swapping still retains the general characteristics of exact plot locations, the FIA program recognizes that perturbing/swapping may have negative effects for some analyses. Therefore, the program has initiated investigations to estimate the effects of perturbing/swapping and to develop methods for circumventing unacceptable effects while still complying with the policy.

Framework for Investigations

Historically, estimation using FIA plot data has been design based, although model-based applications are becoming more extensive. The properties of design-based estimators derive from the sampling designs used to obtain the data. Design-based estimates often are based on plot-based means and variances of means for AOIs; e.g., a design-based estimate of total forestland for a county (or parish) may be obtained as the product of total county or parish area and the mean proportion forest area observed on FIA plots in the county. For design-based estimates, the primary effect of perturbing/swapping is that the set of plots determined to be in an AOI on the basis of perturbed/swapped plot locations will exclude some plots that are in the AOI and include some that are outside the AOI. The negative effects of perturbing/swapping decrease as the sizes of the circular AOIs increase and as the geographic distance and strength of spatial correlation among observations increases. Spatial correlation describes the relationship between observations of forest attributes in terms of the geographic distance between observations. Observations that are closer together will be more highly correlated, whereas observations that are greater distances apart will have smaller correlations. In addition, observations of some attributes such as proportion forest area are more highly correlated at the same distance than are observations of other attributes such as volume per unit area.

The properties of model-based estimators derive primarily from the mathematical forms of the model and the unexplained residual variability around model predictions. Model-based estimation typically entails formulating mathematical models of the relationships between the observations of the dependent variable and independent variables, predicting the value of the dependent variable for each sampling unit in the AOI, and calculating the mean of predictions over all sampling units in the AOI. Model-based estimation may be based on prediction approaches such as regression, kriging, and nearest neighbor techniques. For example, a regression model of the relationship between observations of proportion forest area from FIA plots as the dependent variable and the values of spectral bands of satellite imagery as independent variables may be used to predict the proportion forest area for every pixel in a county or parish. A model-based estimate of total forest area for the county then may be obtained as the product of the total county area and the mean model prediction of proportion forest area over all pixels with centers in the county.

When the dependent and independent variables are observed at the same geographic location and the coordinates themselves are not independent variables, then the model of the relationship between them is unaffected by perturbing/swapping. For many analyses, however, the dependent and independent variables are not observed coincidentally. As in the foregoing example, the dependent variable may be proportion forest area observed on FIA plots, and the independent variable may be values of spectral bands of satellite imagery for the pixels associated with perturbed plot locations rather than the exact locations.

The effects of associating the incorrect spectral values of the imagery with the proportion forest area observations may introduce bias into the model estimate of the relationship and may cause bias and decreased precision in model predictions. For these situations in which the independent and dependent variables are not observed at the same geographic location, the effects of perturbing/swapping on model-based estimates decrease as the size of the sampling unit increases because it is less likely that an incorrect value of the independent variable will be associated with the observation of the forest attribute. In addition, as distance at which observations of both the dependent and the independent variables are highly correlated increases, the effects of perturbing/swapping decrease because even though incorrect values of the independent variable are associated with the dependent variable, the incorrect value is more similar to the correct value. However, unlike the case of design-based estimation, the negative effects of perturbing/swapping do not necessarily decrease as the sizes of AOIs increase.

The Effects of Perturbing/Swapping on Design-Based Estimates

The purpose of swapping plot locations is to create uncertainty in plot ownership. Although the national maximum proportion of plots that requires swapping is 0.10, the actual proportion may be less depending on local conditions. For example, in regions in which landownership is fragmented into small parcels, the ownership of land containing perturbed plot locations are often different than land containing the exact plot location. Thus, when a side effect of perturbing plot locations is uncertainty in ownership, the proportion of plot locations requiring swapping may be smaller.

Because the proportion of plot locations that are swapped and the proportion that are perturbed more than 0.5 mi (0.8 km) are small, the effects of perturbing/swapping will be dominated by the component of the effects associated with plots whose locations are perturbed 0.5 mi (0.8 km) or less. For a circular AOI (Figure 1), the expected correlation ρ between design-based estimates using exact and perturbed plot locations may be expressed in terms of four quantities: (1) the radius R of the AOI (region A + region B), (2) the number of plots with exact locations in region B whose perturbed locations place them in region C, (3) the number of plots with exact locations in region C whose perturbed locations place them in region B; and (4) the spatial correlation of the attribute of interest.

An extreme case scenario resulting in the minimum correlation between estimates based on exact and perturbed plot locations occurs when all plots with exact locations in region B are replaced by plots with exact lo-
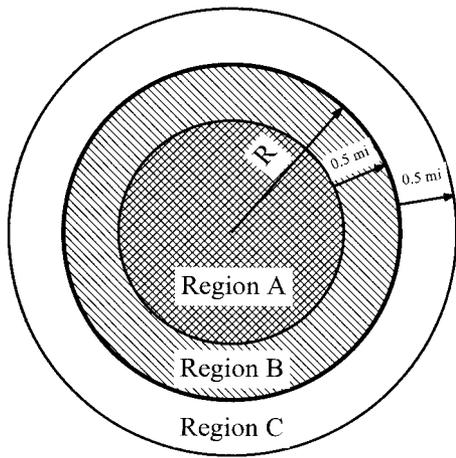
Figure 1. Circular AOI.

cations in region C and when the estimates for these two sets of plots are uncorrelated. Under this scenario and assuming a maximum perturbing distance of 0.5 mi (0.8 km), the expected correlation between means estimated using data from exact and perturbed locations may be expressed as

$$\rho = \frac{\text{area}_{\text{region A}}}{\text{area}_{\text{region A}} + \text{area}_{\text{region B}}}$$

$$= \frac{\pi(R - 0.50)^2}{\pi R^2}$$

$$= \left(1 - \frac{0.50}{R}\right)^2.$$

A graph of ρ versus R (Figure 2) indicates that for circular AOIs of radius $R > 10$ mi (16 km), the extreme case correlation between estimates based on exact and perturbed locations will be $\rho > 0.90$ and for $R > 20$ mi (32 km), the extreme case correlation will be $\rho > 0.95$.

Empirical studies of the effects of perturbing/swapping have focused on comparisons of design-based means using exact and perturbed/swapped plot locations. Lister et al. (2005) compared estimates of mean board foot volume using exact locations and locations that had been perturbed/swapped using procedures that mimic the current procedures as previously described. In Maine, for circular AOIs of radii 3.1 mi (5 km), 6.2 mi (10 km), and 12.5 mi (20 km), they found correlations of 0.78, 0.89, and 0.95, respectively (Figure 2), between estimates of mean board foot volume based on exact and perturbed/swapped locations. Guldin et al. (in press) selected a location in Virginia and calculated estimates of forestland and timberland, number of live and

growing stock trees, and volumes per unit area for circular AOIs of radii 50 mi (80 km), 75 mi (120 km), 100 mi (160 km), 125 mi (200 km), and 150 mi (240 km). Differences in estimates obtained using exact and perturbed/swapped plot locations were always less than 1% and usually were less than 0.5% for all attributes and for AOIs of all radii. These results would be expected, because the radii of all AOIs were greater than the extreme case 20-mi (32-km) radius beyond which correlations are nearly 1.

For this study, 20 circular study areas of radius 18.7 mi (30 km) were selected, 10 in heavily forested areas in Minnesota, Pennsylvania, and Maine, and 10 in more fragmented, sparsely forested areas in the same states. Within each study area, estimates of mean proportion forest area and mean volume per unit area were calculated for circular AOIs centered at the center of the study area and with radii of 3.1 mi (5 km), 6.2 mi (10 km), 12.5 mi (20 km), and 18.7 mi (30 km). Correlations between estimates using data based on exact and perturbed/swapped locations were all greater than 0.90 (Figure 2). The correlations were greater for proportion forest area than for volume per unit area because the distance at which observations of proportion forest area are highly correlated is greater than the distance for volume per unit area.

Figure 2 summarizes the extreme case scenario and the empirical studies and confirms expectations regarding correlations between estimates obtained using exact and perturbed/swapped plot locations. First, the correlations obtained from the empirical studies are always greater for the same radius of the circular AOI than for the extreme case scenario, which represents the minimum correlation that should be expected. Second, correlations are greater for the same radius for proportion forest area than for the volume variables because the distance at which observations of the proportion forest area are highly correlated is greater than the distance for the volume variables. Third, the 20- and 30-mi radii for which correlations should be expected to be greater than $\rho > 0.90$ or $\rho > 0.95$, respectively, are smaller than those used for most applications, suggesting that the negative effects of perturbing/swapping plot locations may be minimal for those applications.

In summary, the design-based investigations indicate that in the extreme case the effects of perturbing/swapping are negligible for circular AOIs of radii greater than ap-

proximately 20 mi (32 km), a conclusion confirmed by the three empirical studies. The primary factors are the size of the AOI and the distance at which observations of the variable of interest are highly correlated. These results are expected to hold for AOIs of regular polygonal shapes, although long, narrow AOIs and irregularly shaped AOIs may require special consideration.

The Effects of Perturbing/Swapping on Model-Based Estimates

Coulston et al. (in press) report a study based on 12,730 plots in Minnesota that simulated the effects of perturbing/swapping on the accuracy of predictions for an arbitrary, simulated dependent variable. Two approaches to prediction were considered: ordinary kriging and multiple linear regression using arbitrary, simulated independent variables. They concluded that perturbed/swapped plot locations did not contribute to bias in estimates based on kriging. However, for the regression approach perturbed/swapped plot locations contributed to bias in the selection of statistically significant independent variables and in model predictions. These negative effects were most pronounced when the independent variables were derived from data with fine spatial resolution (e.g., 30 m) and when the quality of fit of the model to the data was greater. In addition, the negative effects were worse when the distance at which observations of variables were correlated was small.

Guldin et al. (in press) reported estimates of the effects of perturbing/swapping on maps of cubic foot volume per acre for Connecticut, constructed using a multiple linear regression model. Volume per acre observations were obtained from FIA plot data, and independent variables included variables derived from 30-m Landsat Thematic Mapper (TM) imagery, road densities, and geographic plot locations. The quality of fit for the multiple linear regression model was $R^2 = 0.43$ when using data for both exact and perturbed/swapped plot locations. These results are consistent with the Coulston et al. (in press) study that found fewer discernible differences when the quality of fit of the model to the data was moderate or poor.

McRoberts and Holden (in press) estimated the effects of perturbing plot locations on model-based small area estimates of proportion of forest area. A logistic regres-
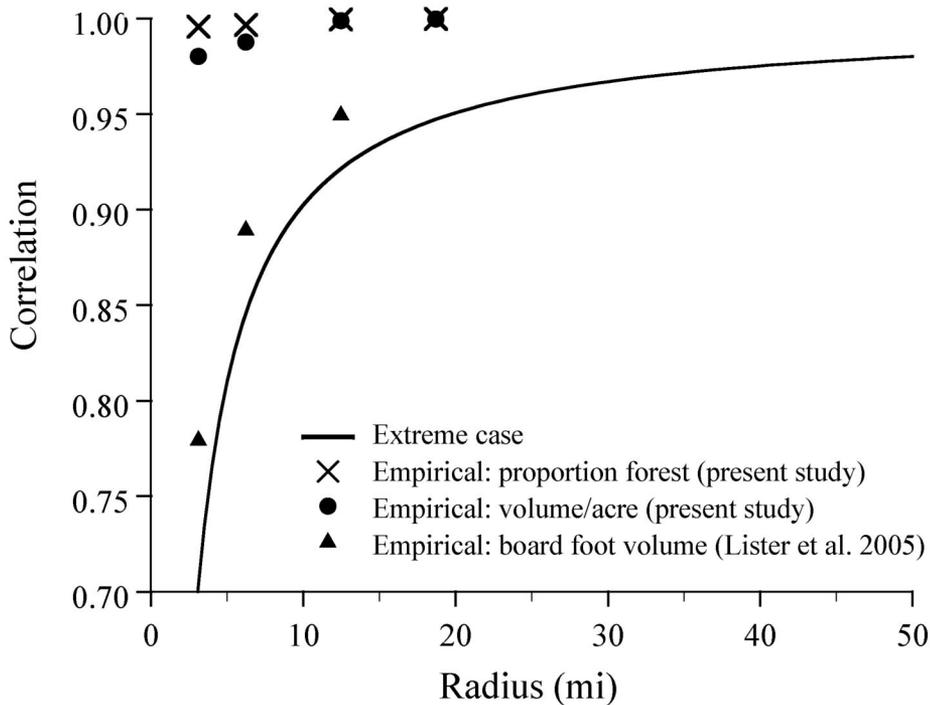
Figure 2. Correlations for extreme case scenario and empirical estimates.

sion model was developed using observations of proportion of forest area for the 24-ft (7.32-m) radius FIA subplots and spectral values of 30-m TM imagery. The model was used to predict the probability of forest for each pixel in three circular study areas of radius 9.3 mi (15 km) in Minnesota. Within each study area, design- and model-based estimates of mean proportion forest area were calculated for circular AOIs with radii ranging from 0.6 mi (1.0 km) to 9.3 mi (15.0 km) and centered at the study area centers. In nearly all instances, the model-based estimates using data for exact plot locations were within two design-based standard errors of the design-based estimates. For comparison purposes, the locations of all plots were randomly perturbed within a circular area of radius 0.5 mi (0.8 km). Each subplot was then associated with the pixel containing the perturbed subplot center, the model was recalibrated, pixel predictions were recalculated, and AOI estimates were recalculated. This procedure was repeated 10 times. Although the bias in model-based estimates based on perturbed plot locations was not severe, it was consistent and was not dependent on the size of the AOIs.

In summary, the effects of perturbing/swapping plot locations on model-based estimation are quite different than the effects on design-based estimation. For model-based estimation, the relevant factors are the

size of the unit from which the independent variables are obtained relative to the spatial correlation of both the dependent and the independent variables. The relative effects of the approach to prediction (e.g., regression, kriging, etc.) still are undetermined.

Circumventing the Negative Effects of Perturbing/Swapping

For situations in which the effects of perturbed/swapped plot locations are unacceptable, the FIA program is developing alternatives. First, users may seek assistance from the FIA Spatial Data Services (SDS) centers. These centers respond to user requests by integrating FIA plot data with geospatial data, checking for compliance with policy, and returning integrated data sets or summaries. Services include summarizing FIA plot data by classification category for geographic information systems layers provided by users, attaching spectral values of satellite imagery to FIA plot data, and hosting users whose plot location requirements can not otherwise be satisfied.

Increasingly, users seek FIA plot data for use in training or evaluating the accuracy of classifiers of satellite imagery. If the pixel resolution is substantially greater than the perturbing radius, then there are few difficulties. However, for moderate resolution imagery such as 250-m MODIS and 30-m TM the pixel resolution is much less than

the perturbing radius. If the combination of spectral values for the pixel associated with a plot is unique within the TM scene, then the plot may be located to within the pixel resolution by searching the satellite image for that particular combination of spectral values. The probability of locating the plot in the scene is greater when the values of more spectral bands are used. If imagery for a single date of TM imagery is used, values for six spectral bands are available; if imagery for two dates is used, values for 12 spectral bands are available; and if imagery for three dates is used, values for 18 spectral bands is available. Recent investigations indicate that when only one or two dates of imagery are used, the probability of locating plots is small, but with three dates, the probability is unacceptably large. Expanding a study by Liknes et al. (2005), McRoberts and Holden (2005) found that by adding a randomly selected integer between −5 and + 5 to each spectral value of the three dates of TM imagery, the probability of locating the plot was negligible, whereas the logistic regression model-based estimates of proportion forest area based on exact and perturbed spectral values were nearly indistinguishable.

Finally, a prototyping effort (McRoberts and Miles 2005) is underway to construct unbiased, moderate resolution, model-based maps of forest attributes and make them available via the Internet. The objective is for users to be able to submit polygons representing their AOIs and receive in return model-based means, totals, and respective variance estimates. The challenge will be to develop techniques that produce maps that are sufficiently unbiased at appropriate spatial resolutions and to develop Internet client-server interfaces that are fast enough to accommodate users.

Summary

The cumulative results of studies of the effects of perturbing/swapping on bias indicate that for design-based estimation, the relevant factors are the size of the AOI and the spatial correlation of the variable of interest. For most forest attributes, the effects of perturbing/swapping are negligible for circular AOIs of radii greater than approximately 20 mi (32 km). For model-based estimation, the relevant factors are the size of the sampling unit on which the independent variable is observed and the spatial correlation for both the dependent and the independent variables. The results of empirical

studies of the effects of perturbing/swapping on model-based estimates are mixed. One study using a kriging approach to estimation reported no negative effects. Two studies using regression approaches found unacceptable effects when independent variables were observed on sampling units of the approximate size of 30-m TM pixels, although the results of a third study that used TM imagery with other independent variables found no effects.

When the effects of perturbing/swapping are unacceptable, the FIA program provides users access to the SDS centers. In addition, the program is investigating methods for matching plot data with satellite imagery while still complying with policy.

Literature Cited

COULSTON, J.W., K.H. RIITERS, G.A. REAMS, R.E. MCROBERTS, AND W.D. SMITH. Practical considerations when using perturbed forest inventory plot locations to develop simple linear regression models and ordinary kriging: A simulation study. *Proc. of the 6th annual forest inventory and analysis symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and W.H. McWilliams (eds.). USDA For. Serv. Gen. Tech. Rep. WO-GTR-70 (in press).

GULDIN, R.W., S.L. KING, AND C.T. SCOTT. Vision for the future of FIA: Paean to progress, possibilities, and partners. *Proc. of the 6th annual forest inventory and analysis symposium*, McRoberts, R.E., G.A. Reams, P.C Van Deusen, and W.H. McWilliams (eds.). USDA For. Serv. Gen. Tech. Rep. WO-GTR-70 (in press).

LIKNES, G.C., G.R. HOLDEN, M.D. NELSON, AND R.E. MCROBERTS. 2005. Spatially locating FIA plots from pixel values. P. 99–104 in *Proc. of the 4th annual forest inventory and analysis symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and W.H. McWilliams (eds.). USDA For. Serv. Gen. Tech. Rep. NC-GTR-252. USDA Forest Service, North Central Research Station, St. Paul, MN.

LISTER, A., C.T. SCOTT, S.L. KING, M. HOPPUS, B. BUTLER, AND D. GRIFFITH. 2005. Strategies for preserving owner privacy in the national information management system of the USDA Forest Service's Forest Inventory and Analysis unit. P. 163–166 in *Proc. of the 4th annual forest inventory and analysis symposium*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and W.H. McWilliams (eds.). USDA For. Serv. Gen. Tech. Rep. NC-GTR-252. USDA Forest Service, North Central Research Station, St. Paul, MN.

MCROBERTS, R.E., AND G.H. HOLDEN. Evaluating the effects of uncertainty in inventory plot locations. *Sustainable forestry in theory and in practice, recent advances in inventory and monitoring, statistics and modelling, information and knowledge management, and policy science, Proc. of the IUFRO meeting, April 5–8, 2005, Edinburgh, Scotland*, Reynolds, K., K. Rennolls, A. Thomson, M. Shannon, and M. Köhl (eds.). Gen. Tech. Rep. USDA Forest Service, Pacific Northwest Research Station, Portland, OR. (in press).

MCROBERTS, R.E., AND P.D. MILES. 2005. On-line, map-based estimation of forest attributes. P. 701–705 in *Proc. of the 16th international workshop on database and expert systems applications (DEXA* 2005), August 22–26, 2005, Copenhagen, Denmark. IEEE Computer Society, Conference Publishing Services, Los Alamitos, CA, IEEE Computer Society Order P2424, ISBN 0-7695-2424-9, ISSN 1529-4188.

Ronald E. McRoberts (rmcroberts@fs.fed.us) is group leader for Research and Analysis, Geoffrey R. Holden (gholden@fs.fed.us) is forester, Mark D. Nelson (mdnelson@fs.fed.us) is research forester, Greg C. Liknes (gliknes@fs.fed.us) is research physical scientist, and Warren K. Moser (wkmoser@fs.fed.us) is research forester, North Central Research Station, USDA Forest Service, 1992 Folwell Avenue, St. Paul, MN 55108. Andrew J. Lister (alister@fs.fed.us) is research forester, Susan L. King (sking01@fs.fed.us) is operations research analyst, and Elizabeth J. LaPoint (elapoint@fs.fed.us) is forester, Northeastern Research Station, USDA Forest Service, 11 Campus Boulevard, Suite 200, Newtown Square, PA 19703. John W. Coulston (jcoulston@fs.fed.us) is research assistant professor, Department of Forestry and Environmental Resources, North Carolina State University, Box 8008, Raleigh, NC 27695. W. Brad Smith (bsmith12@fs.fed.us) is associate national program leader, and Gregory A. Reams (greams@fs.fed.us) is national program leader, Washington Office, USDA Forest Service, 1601 North Kent Street, Arlington, VA 22209.