

A model-based approach to estimating forest area

Ronald E. McRoberts

St. Paul, MN, USA

Received 23 July 2004; received in revised form 24 March 2006; accepted 25 March 2006

Abstract

A logistic regression model based on forest inventory plot data and transformations of Landsat Thematic Mapper satellite imagery was used to predict the probability of forest for 15 study areas in Indiana, USA, and 15 in Minnesota, USA. Within each study area, model-based estimates of forest area were obtained for circular areas with radii of 5 km, 10 km, and 15 km and were compared to design-based estimates based on inventory plot data. Precision estimates for the circular areas were also obtained using variance formulae developed for this application that incorporated spatial correlation among model predictions for individual pixels. The model-based estimates were generally comparable to the design-based estimates. The advantages of the model-based approach are that maps and small areas estimates may be obtained and the necessity of releasing exact plot locations for user-specific applications is alleviated.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Logistic regression model; Landsat Thematic Mapper; Indiana, USA; Minnesota, USA

1. Introduction

Traditionally, large-scale, natural resource inventory programs have used data collected from ground plots to respond to the user question “How much?” by reporting plot-based estimates of natural resource attributes for states or provinces, counties, or municipalities. Increasingly, users are also asking “Where?” and are requesting access to plot data for estimation for their own areas of interest (AOI). When data requests do not require exact plot locations, there are few constraints on data access. However, if exact locations are required, then several issues must be considered. First, revealing exact locations may entice users to visit the plots to obtain additional information, thus artificially disturbing the sampling location and contributing to bias in inventory estimates. Second, plots may be located on private land, and while landowners usually permit access by inventory field crews, they generally prohibit additional access. In these situations, user visits to plot locations may jeopardize future access by inventory field crews. Third, revealing the exact plot locations may violate constraints on the release of proprietary information. Thus, if exact plot locations are required for a user’s analysis, policy

constraints may prohibit the inventory program from accommodating the user’s data request.

In response to the “Where?” question, the Forest Inventory and Analysis (FIA) program of the USDA Forest Service has initiated local, regional, and national mapping efforts. Although the objectives of these efforts have been to map the spatial distributions of forest attributes, generally they have not included investigations of whether maps may be used to obtain unbiased and precise areal estimates of those attributes. For the latter objective, estimates obtained using only data for the plots located in the AOI have been necessary. If unbiased and sufficiently precise areal estimates of forest attributes could be obtained by aggregating mapping unit predictions, then several advantages would accrue. First, release of plot locations would be unnecessary for estimation for users’ AOIs. Second, because mapped values would be based on aggregated data from multiple plots, proprietary information would not be released. Third, estimation would be possible for small areas for which the number of plots is insufficient for plot-based estimation. Fourth, efficiencies would be gained by simultaneously addressing both the “How much?” and “Where?” questions.

Nelson et al. (2005) compared forest area estimates based on sample plot data from two national inventories of the USA to comparable estimates from four satellite image-derived maps.

E-mail address: mroberts@fs.fed.us.

The sample plot data were collected for non-federal lands by the National Resources Inventory (NRI), a program of the USDA Natural Resources Conservation Service, and for all forestland ownerships by the FIA program. The satellite image-derived maps included a nominal 1991 AVHRR forest cover type map of 1-km spatial resolution (Zhu & Evans, 1994), a nominal 1992–1993 AVHRR land cover map of 1-km spatial resolution from the National Atlas (2003), the Vegetation Continuous Fields (VCF) percent tree canopy cover data obtained from MODIS imagery with a 500-m spatial resolution (Hansen et al., 2003), and the National Land Cover Dataset (NLCD) obtained from nominal 1992 Landsat TM imagery with 30-m spatial resolution (Vogelmann et al., 2001). They found that the forest cover type map produced non-federal forest area estimates that were most similar to NRI and FIA estimates and that for the VCF data, a minimum canopy cover threshold of 25% produced national estimates for all ownerships that were most similar to FIA estimates. However, they found that the 25% threshold produced large deviations between state-level VCF and FIA estimates. They concluded that it would be inappropriate to use forest/non-forest maps created from VCF products to estimate forest area for states or smaller geographic areas.

Nelson et al. (2005) also reviewed and reported European comparisons of forest inventory and satellite imaged-derived estimates of forest area (Häme et al., 2001; Kennedy & Bertolo, 2002; Päivinen et al., 2001; Schuck et al., 2003). Their conclusion was similar to that for comparisons in the USA: While satellite image-derived estimates may be acceptable for large geographic areas, they have limited utility for smaller geographic areas. Although the European forest map was calibrated to match countrywide inventory estimates of proportion forest area, and the forest cover types for the USA were mapped on pixels with forest density estimates exceeding per-state thresholds based on inventory estimates, neither map was specifically designed to produce estimates of forest area, particularly for small areas.

Many investigators have mapped forest cover using classification techniques and coarse spatial resolution AVHRR data. Mayaux and Lambin (1995) note that the advantages of the coarser resolution maps are greater data availability and a spatial resolution that more closely matches large AOIs, while the disadvantage is a loss of spatial detail for smaller areas. Mayaux and Lambin (1995), Czaplewski and Catts (1992), and Walsh and Burk (1993) describe methods for correcting for misclassification bias.

Kennedy and Bertolo (2002) used AVHRR data with a maximum likelihood approach to calculate the probability of forest. They used unsupervised classification to select clusters of homogeneous 2×2 AVHRR pixel blocks and the Coordination of Information on the Environment (CORINE) data set, a Landsat image-based interpretation of land cover for computing forest area, to train the 2×2 pixel blocks. Inventory-based estimates of forest area were used to calibrate pixel values so that resulting map-based estimates matched inventory-based estimates for counties, regions, and provinces. They predicted forest area proportion for each image pixel as a weighted average of observed proportion forest for the classes where the weights were probabilities of class membership obtained from maximum likelihood analyses.

Alternative approaches to area estimation have included both mixture modeling and regression analyses (Häme et al., 2001). DeFries et al. (2000) used a linear mixture model approach to predict continuous fields of land cover categories from end member classes derived from Landsat Multispectral Scanner system data. Thomas et al. (1993) used a goodness of fit approach to investigate the number of ground cover components and spatial averaging for estimating woodland area by spectral mixing. Cross et al. (1991) identified spectral signatures of pure forest and non-forest cells for coarse resolution imagery and then used a linear mixture model to predict proportions of forest and non-forest by decomposing the spectral values of mixed resolution cells into signature components.

Both Magnussen et al. (2000) and Moisen and Frescino (2002) compared multiple approaches for predicting forest attributes. Magnussen et al. (2000) evaluated predictions of forest cover type proportions obtained from a maximum likelihood classifier and three models using predictors based on proportions of TM clusters obtained from unsupervised classification. Among the models, one was based on neural networks and two were variations of linear models. The neural networks approach yielded lowest mean absolute deviations, while the maximum likelihood approach was better for predicting non-vegetated cover types. Moisen and Frescino (2002) evaluated predictions of two discrete and four continuous forest attributes obtained from linear models, general additive models, classification and regression trees, multivariate adaptive regression splines, and artificial neural networks. They concluded that the multivariate adaptive regression splines and the artificial neural networks approaches were marginally superior.

Regression modeling has been a popular international approach for use with satellite imagery to map a variety of forest attributes: large area volume and above ground biomass in Finland (Tomppo et al., 2002); hardwood and conifer cover in Oregon, USA (Mayersberger et al., 2001); height and basal area in Scotland (Pühr & Donoghue, 2000); biomass in Brazil (Steininger, 2000); volume in British Columbia, Canada (Gemmell, 1995); age and structure in the Pacific Northwest of the USA (Cohen et al., 1995); age in Estonia (Nilson & Peterson, 1994); age in Colorado, USA (Nel et al., 1994); biomass in England (Danson & Curran, 1993); volume in New Brunswick, Canada (Ahern et al., 1991); and suites of forest inventory variables in Finland (Tomppo, 1987, 1988).

Numerous investigators have used regression techniques to estimate or enhance estimates of forest area. For example, Deppe (1998) used a regression approach to enhance estimation of forest area in Brazil and Bolivia. The combined use of regression techniques, double sampling, and fine and coarse resolution imagery has received considerable attention. Iverson et al. (1989) classified Landsat TM imagery with respect to forest and non-forest using forest inventory data and aerial photographs and then used linear regression to estimate the relationship between the amount of forest landscape in the classified TM imagery and AVHRR spectral values. Nelson (1989) used linear regression, ratio of means, and mean of ratios estimators with AVHRR-GAC 4-km data and Landsat MSS 80-m data. Zhu and Evans (1994) used ground sample data to classify finer resolution Landsat TM imagery with respect to a forest attribute

of interest and then used multiple regression to estimate the relationship between the Landsat TM classifications and the spectral characteristics of AVHRR data. [Mayaux and Lambin \(1995\)](#) used linear regressions to estimate relationships between proportions of fine resolution pixels and coarse resolution pixels classified as forest.

Finally, some studies have reported uncertainties associated with individual pixel predictions. However, very few that obtain areal estimates of forest attributes as means or sums of individual pixel predictions address methods for calculating the variances of these areal estimates. [Taskinen and Heikkinen \(unpublished manuscript\)](#) used a non-parametric Bayesian approach featuring partition modeling, and [de Bruin \(2000\)](#) used a sequential simulation indicator approach featuring collocated indicator cokriging. However, methods for obtaining variances of areal estimates obtained as means or sums of regression model predictions for individual pixels are not readily available.

The cited regression studies may be characterized generally by three limiting features. First, they usually depend on coarse resolution satellite imagery to produce the final map product which limits the ability of the map to produce accurate estimates for small areas. Second, they tend to feature linear models, despite the lack of a theoretical or empirical basis for assuming linearity. The two comparative studies ([Magnussen et al., 2000](#); [Moisen & Frescino, 2002](#)) concluded that approaches such as regression splines and neural networks which incorporate nonlinear features produced better results than did the linear model approaches. Third, very few of the studies that obtain estimates as means or sums of individual pixel predictions have reported methods for estimating variances of areal estimates. Estimation of variances in the latter situation is not trivial because of the necessity of accommodating spatial correlation among training data observations and pixel prediction residuals.

The objective of this study was to construct forest probability maps from FIA plot data and medium resolution Landsat Thematic Mapper (TM) or Enhanced Thematic Mapper (ETM+) satellite imagery using a logistic regression model. The motivation was fourfold: (1) to investigate the utility of logistic regression models using medium resolution satellite imagery for mapping the distribution of forestland, (2) to compare plot-based and logistic regression model-based estimates of forest area, (3) to develop a methodology that accommodates spatial correlation when estimating model parameter and areal precision estimates, and (4) to develop an approach to estimation, particularly for small areas, that circumvents the constraint on release of FIA plot locations. The study focused on predicting pixel-level probabilities of forest ground cover and comparing map- and plot-based estimates of forest area. A major component of the study was accommodating spatial correlation in the variance estimates for individual pixel predictions and for total forest area for multiple pixel AOIs.

2. Data

2.1. Forest inventory plot data

The FIA program has established field plot centers in permanent locations using a sampling design that is assumed to

produce a random, equal probability sample ([Bechtold & Patterson, 2005](#); [McRoberts et al., 2005](#)). The plot array has been divided into five non-overlapping, interpenetrating panels, and measurement of all plots in one panel is completed before measurement of plots in the next panel is initiated. Panels are selected on a 5-, 7-, or 10-year rotating basis, depending on the region of the country. Over a complete measurement cycle, the sampling intensity is approximately one plot per 2400 ha. Some states provide additional funding to double or triple the sample size in which case the sampling intensity is approximately one plot per 1200 ha or 800 ha, respectively. In general, locations of forested or previously forested plots are determined using global positioning system (GPS) receivers, while locations of non-forested plots are determined using aerial imagery and digitization methods. Each field plot consists of four 7.31-m radius circular subplots. The subplots are configured as a central subplot and three peripheral subplots with centers located at 36.58 m and azimuths of 0°, 120°, and 240° from the center of the central subplot. Among the observations field crews obtain are the proportions of subplot areas that satisfy specific ground land use conditions. Subplot estimates of proportion forest area are obtained by collapsing ground land use conditions into forest and non-forest classes consistent with the FIA definition of forestland (primarily, area ≥ 0.4 ha and width ≥ 36.6 m). For this study, all plots were observed between the beginning of 1999 and the end of 2003. FIA plot data were used for two large analytical areas, one corresponding approximately to a Landsat Thematic Mapper (TM) scene in south central Indiana, USA, and one corresponding approximately to a scene in northeastern Minnesota, USA ([Fig. 1](#)). Land use for the Indiana analytical area consists of agriculture and hardwood forest, mostly oak and hickory. Land use for the Minnesota analytical area consists of forestland dominated by aspen–birch and spruce–fir associations, wetlands, lakes, and agriculture.

2.2. Satellite imagery

Landsat imagery for the Indiana scene, row 33 of path 21, and the Minnesota scene, row 27 of path 27, was obtained from the Multi-Resolution Characterization 2001 (MRLC 2001) land cover mapping project ([Homer et al., 2004](#)) of the U.S. Geological Survey ([Fig. 1](#)). The imagery was characterized by several salient features: (1) a combination of Landsat 5 TM and Landsat 7 ETM+ data, (2) geometrically and radiometrically corrected, (3) cubic convolution resampling to 30 m $\times$ 30 m spatial resolution, (4) visible and infrared bands (1–5.7), and (5) conversion to at-satellite reflectance. Imagery for three dates corresponding to early, peak, and late vegetation green-up ([Yang et al., 2001](#)) were obtained for each scene: April 2001, July 2000, and October 2001 for the Indiana scene and March 2000, July 1999, and October 1999 for the Minnesota scene. Preliminary analyses indicated that NDVI and the tassell cap (TC) transformations (brightness, greenness, and wetness) ([Crist & Cicone, 1984](#); [Kauth & Thomas, 1976](#)) were superior to both the spectral band data and principal component transformations with respect to predicting the probability of forest cover. Thus, 12 satellite image-based predictor variables

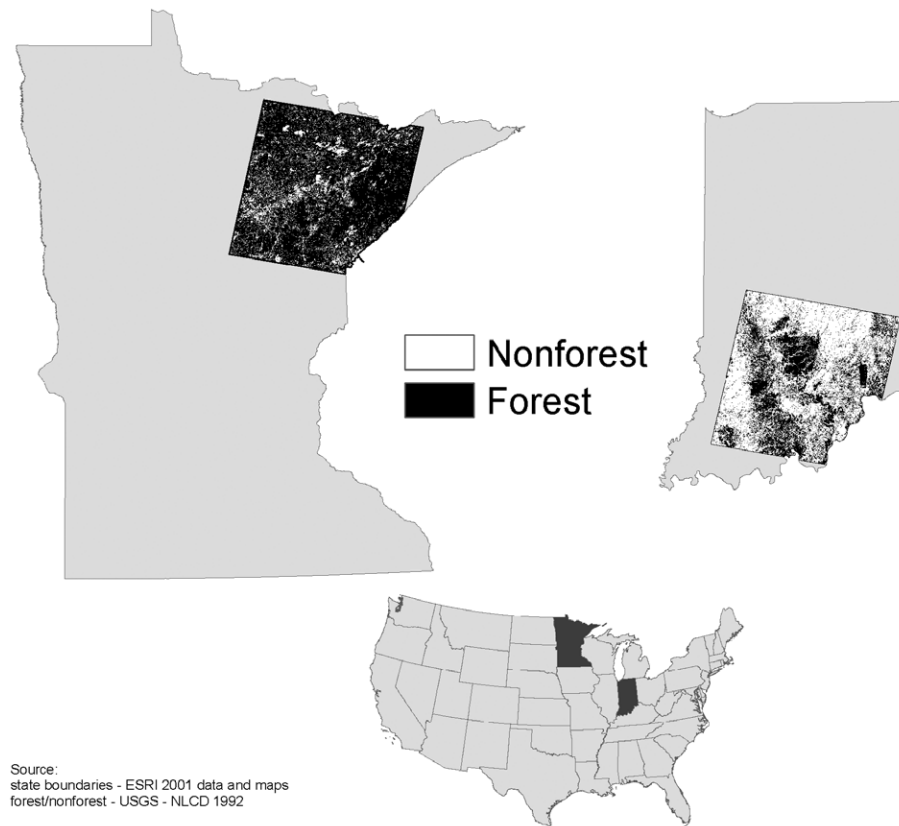


Fig. 1. Indiana and Minnesota analytical areas.

were used, NDVI and the three TC transformations for each of the three image dates.

2.3. Combining forest inventory data and satellite imagery

The spatial configuration of the FIA subplots with centers separated by 36.58 m and the 30-m $\times$ 30-m spatial resolution of the TM /ETM+ imagery permits individual subplots to be associated with individual image pixels. The subplot area of 167.87 m² is approximately 19% of the 900 m² pixel area. When describing relationships between forest attributes observed on FIA plots and TM satellite image spectral values, two phenomena must be considered. First, because a subplot is a single, 19%, contiguous sample of the pixel area, the proportion forest area for a partially forested subplot may not accurately represent the proportion forest area for the entire pixel. Second, GPS and image registration errors may cause a subplot to be associated with an incorrect pixel, resulting in the forest attribute of a subplot being erroneously associated with the spectral signature of a non-forested pixel, and vice versa. Both phenomena obscure relationships between observed forest attributes and spectral data, cause bias in estimates of parameters of models of the relationships, increase both model residual variability and the uncertainty in model parameter covariance estimates, and contribute to increasing the variance of model-based areal estimates of forest attributes. To reduce the effects of these phenomena, data for plots with mixtures of forest and non-forest cover were excluded when calibrating models. For the Indiana analytical area, 627 homogeneous plots or 2508

subplots were used of which 492 were completely forested and 2016 were completely non-forested. For the Minnesota analytical area, 1228 homogeneous plots or 4912 subplots were used of which 4368 were completely forested and 544 were completely non-forested. The assumption underlying exclusion of heterogeneous plots is that if their subplots were associated with the correct pixels, then the relationship between proportion forest area and the spectral data would be no different for the excluded subplots than for non-excluded subplots. Under this assumption, there should be no resulting effects on estimates of the relationships, although model parameter estimates may be less precise. However, if there were differences in the relationships, then differences should be observed when comparing forest area estimates based on predictions from models calibrated using only data for non-excluded plots to plot-based estimates for the same AOIs using data for all plots.

3. Methods

Traditionally, the FIA program has used methods that base areal estimates directly on sample observations. These methods assume a population for which there is a single fixed value for each population element and that the distribution of these fixed values may be characterized by a population distribution with mean, μ , and variance, σ^2 . This approach is characterized as design-based estimation, because the properties of the estimators depend on the features of the sampling design. The objective with design-based methods is estimation of the population mean, μ , and the precision

of the estimate of the mean, usually from a random sample drawn from the population according to a prescribed sampling design. An important property of the combination of the sampling design and design-based estimators used by FIA is that in expectation the sample mean is an unbiased estimate of the population mean.

A second approach to estimation, characterized as model-based estimation, assumes a superpopulation characterized by a distribution of possible observations for each population element, not a single fixed value as for design-based estimation (Kangas, 1994; Royall, 1970; Särndal, 1978). Thus, the population assumed for design-based estimation is only one of the possible populations that could be realized from the superpopulation. In this context, μ_i and σ_i^2 may be used to denote the mean and variance of the superpopulation distribution of observations corresponding to the i th set of values of a covariate vector, X . A superpopulation observation, y_i , may be expressed as,

$$y_i = \mu_i + \varepsilon_i,$$

where $E(\varepsilon_i)=0$, $\text{Var}(\varepsilon_i)=\sigma_i^2$, $\text{Cov}(\varepsilon_i, \varepsilon_j)=\sigma_i\sigma_j\rho_{ij}$, and ρ denotes spatial correlation. The objective with model-based methods is estimation of the superpopulation parameters, μ , σ^2 , and ρ for each superpopulation element. With model-based estimation, the parameters of a model relating the sample observations to ancillary data are estimated, and the model predictions form the basis for estimating superpopulation parameters. A distinction between design-based and model-based estimation is that model-based estimation focuses on parameters of the superpopulation, while design-based estimation focuses on parameters of the population that is realized from the superpopulation. Further, the unbiasedness of design-based estimators depends on the design, while the unbiasedness of model-based estimators depends on the model. However, an inadequate sampling design may contribute to the selection of model forms and estimation of model parameters that fail to represent adequately the relationship between the response and predictor variables.

Although estimates of population and superpopulation parameters are based on different conceptual assumptions, at least two reasons for comparing them are relevant. First, the forest inventory community has traditionally used design-based estimation, and the resulting estimates have become a standard for comparison. Second, concern regarding an inadequate model would be alleviated if the mean over all estimates of individual model-based superpopulation means were close to the design-based estimate of the population mean. Because the population assumed for design-based estimation is only one of the possible populations that could be realized from the superpopulation, the difference in the design-based estimate of the population mean and the model-based estimate of the mean of the individual superpopulation means could be interpreted with respect to the distribution of possible population means. If the model is correctly formulated and the sample used to estimate its parameters is adequate, then the design-based estimate of the population mean should be close in a relative sense to the model-based estimate of the mean of the individual superpopulation means. If they are not close, then one of three interpretations is possible: the model fails to represent adequately the relationship between the response and predictor

variables, the design-based sample is not representative of the realized population, or an unlikely population has been realized from the superpopulation. Attention is usually focused on the first two alternatives, although it may be difficult to distinguish between them. Methods for comparing design- and model-based estimates are described in a later section.

3.1. Design-based estimation

The FIA program uses a design-based approach to estimation that is based on an infinite population framework and a sampling design that is assumed to produce an equal probability sample (Bechtold & Patterson, 2005; McRoberts et al., 2005). In accordance with its infinite population framework, the program attributes the aggregation of data for the four subplots to the center point of the central subplot. The design-based estimator of total forest area, F , for an AOI is,

$$\hat{F} = A\bar{Y} \quad (1)$$

where A is the total area of the AOI, and

$$\bar{Y} = \frac{\sum_{i=1}^n y_i}{n}, \quad (2)$$

where y_i is the observation of proportion forest area for the i th plot, and n is the sample size. Although the FIA program uses post-sampling stratification and stratified estimation, design-based variance estimators for this study are based on an assumption of simple random sampling because sufficient numbers of plots per stratum are not available for stratified estimation for small AOIs. Because A is a constant, mean proportion forest area, $\bar{Y}$, is the quantity of interest. An estimator of the variance of $\bar{Y}$ is,

$$\text{Var}(\bar{Y}) = \frac{s_y^2}{n} \quad (3)$$

where

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n-1}.$$

When calculating design-based estimates, data were included for all four subplots of all plots with centers in the AOI, regardless of whether the plots had a mixture of forest and non-forest conditions.

3.2. Model-based estimation

3.2.1. Model calibration

An important result of the exclusion of data for plots with mixed forest and non-forest conditions is that observed proportion forest area, y , is a binary variable and may be quantified according to the convention that $y=1$ denotes a forested subplot and $y=0$ denotes a non-forested subplot. The relationship between the binary forest/non-forest-dependent

variable and the continuous spectral-independent variables, X , may be expressed as,

$$p_i = f(X_i; \beta) + \varepsilon_i, \quad (4a)$$

or

$$E(p_i) = f(X_i; \beta), \quad (4b)$$

where i indexes subplots, p_i is the probability that $y_i=1$, β is a vector of parameters to be estimated, $f(X_i; \beta)$ is a function expressing the relationship among the independent and dependent variables and the parameters, ε_i is unexplained residual uncertainty assumed to be distributed with zero mean, and $E(\cdot)$ denotes statistical expectation. For binary-dependent variables, $f(X_i; \beta)$ is often expressed as a logistic model of which one form is,

$$f(X_i; \beta) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im})}. \quad (5)$$

The parameters of (5) are often estimated by maximizing the likelihood, L , expressed as,

$$L = \prod_{i=1}^n f(X_i; \beta)^{y_i} [1 - f(X_i; \beta)]^{1-y_i}, \quad (6)$$

where n is the number of subplot observations. Maximum likelihood parameter estimates based on (6) may be obtained using any of several statistical packages, (e.g., SAS, 1988). However, because (6) cannot accommodate spatial correlation among subplot observations, another approach to parameter estimation must be considered if correct measures of uncertainty are required. Nevertheless, because (6) yields unbiased parameter estimates, it may be used to obtain parameter estimates if no measures of uncertainty are required or to obtain initial parameter estimates for iterative approaches that accommodate spatial correlation.

An iterative approach to logistic model parameter estimation that accommodates spatial correlation is based on generalized estimating equations (GEE) (Albert & McShane, 1995; Gumpertz et al., 2000). For the first iteration, ordinary logistic regression using maximum likelihood was used to fit (6) to the data. The GEE approach consisted of solving,

$$\sum_{i=1}^n Z_i' V_e^{-1} (y_i - p_i) = 0, \quad (7)$$

where the elements, z_{ij} , of Z_i are,

$$z_{ij} = \frac{\partial f(X_i; \hat{\beta})}{\partial \beta_j}, \quad (8)$$

and $f(X_i; \hat{\beta})$ is the logistic function, (5), evaluated using the parameter estimates, $\hat{\beta}$. After each iteration, the elements of the subplot residual covariance matrix, V_e , were recalculated as,

$$v_{ij} = s_i s_j \hat{\rho}_{ij} = \sqrt{\hat{p}_i(1-\hat{p}_i)} \sqrt{\hat{p}_j(1-\hat{p}_j)} \hat{\rho}_{ij}, \quad (9)$$

where $s_i^2 = \hat{p}_i(1-\hat{p}_i)$ is the variance of $\varepsilon_i = y_i - \hat{p}_i$. The estimated spatial correlation, $\hat{\rho}_{ij}$, was obtained using a fitted semi-variogram obtained as follows (Gumpertz et al., 2000):

- (1) calculate the standardized residuals,

$$\delta_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}; \quad (10)$$

- (2) construct an empirical semi-variogram,

$$\hat{\gamma}(d) = \frac{1}{2N(d)} \sum_{N(d)} (\delta_i - \delta_j)^2 \quad (11)$$

where $N(d)$ denotes a collection of pairs, (δ_i, δ_j) , whose Euclidean distances in geographic space place are within a given neighborhood of d .

- (3) using weighted nonlinear regression, fit an exponential semi-variogram (for example),

$$\hat{\gamma}(h) = \hat{\alpha}_0 + \hat{\alpha}_1 [1 - \exp(-\hat{\alpha}_2 h)], \quad (12)$$

to the empirical semi-variogram;

- (4) and estimate ρ_{ij} as,

$$\begin{aligned} \hat{\rho}_{ij} &= 1 - \frac{\hat{\gamma}(d_{ij})}{\hat{\gamma}_{\text{total}}} = 1 - \frac{\hat{\alpha}_0 + \hat{\alpha}_1 [1 - \exp(-\hat{\alpha}_2 d_{ij})]}{\hat{\alpha}_0 + \hat{\alpha}_1} \\ &= \frac{\hat{\alpha}_1 \exp(-\hat{\alpha}_2 d_{ij})}{\hat{\alpha}_0 + \hat{\alpha}_1}. \end{aligned} \quad (13)$$

For many applications, including this one, $\alpha_0=0$, in which case,

$$\hat{\rho}_{ij} = \exp(-\hat{\alpha}_2 d_{ij}),$$

and after substitution into (9) yields,

$$v_{ij} = s_i s_j \hat{\rho}_{ij} = \sqrt{\hat{p}_i(1-\hat{p}_i)} \sqrt{\hat{p}_j(1-\hat{p}_j)} \exp(-\hat{\alpha}_2 d_{ij}). \quad (14)$$

The $k+1$ st iterative updated parameter estimate, $\hat{\beta}^{k+1}$, was,

$$\hat{\beta}^{k+1} = \hat{\beta}^k + \left(\sum_{i=1}^n Z_i' V_e^{-1} Z_i \right)^{-1} \sum_{i=1}^n Z_i' V_e^{-1} (Y_i - \hat{p}_i). \quad (15)$$

The iterative procedure of updating $\hat{\beta}$ and recalculating V_e continues until convergence and is assumed to produce estimates with $E(\hat{\beta}) = \beta$. An estimator of the parameter covariance matrix is,

$$\hat{V}_{\hat{\beta}} \equiv \left(\sum_{i=1}^n Z_i' V_e^{-1} Z_i \right)^{-1}. \quad (16)$$

For the i th pixel, an estimator of the probability, p_i , that $y_i=1$ is,

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_m X_{im})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_m X_{im})} \quad (17)$$

where $\hat{\beta}$ is the vector of parameter estimates obtained from (15).

Estimators for the variance, $\text{Var}(\hat{p}_i)$, depend on whether $\hat{p}_i$ is considered to be an estimator of the mean, μ_i , of the i th superpopulation distribution or an estimator of an observation, p_i , from that distribution. In terms of generic regression problems, Draper and Smith (1981) characterize the first situation as predicting the mean of the response variable (superpopulation mean, μ_i) for the i th value of the predictor variable and the second situation as predicting an individual observation of the response variable, p_i , which varies about the superpopulation mean, μ_i , value with variance, σ_i^2 . Thus, when $\hat{p}_i$ is considered an estimator of μ_i , a first-order Taylor series approximation may be used to formulate an estimator of $\text{Var}(\hat{p}_i)$ as,

$$\hat{V}\hat{a}r(\hat{p}_i) \approx \hat{V}\hat{a}r(Z_i\hat{\beta}) = Z_i'\hat{V}_{\hat{\beta}}Z_i, \quad (18a)$$

where Z is obtained from (8), and $\hat{V}_{\hat{\beta}}$ is obtained from (16). When $\hat{p}_i$ is considered an estimator of an individual observation, p_i , an estimator of $\text{Var}(\hat{p}_i)$ may be formulated as,

$$\hat{V}\hat{a}r(\hat{p}_i) \approx \hat{V}\hat{a}r(Z_i\hat{\beta} + \varepsilon_i) = Z_i'\hat{V}_{\hat{\beta}}Z_i + s_i^2 \quad (18b)$$

where s_i^2 is as described following (9).

3.2.2. Estimation for multiple pixel AOIs

A model-based estimator of total forest area, F , for an AOI may be expressed as,

$$\hat{F} = A\bar{P} \quad (19)$$

where

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N \hat{p}_i, \quad (20)$$

$\hat{p}_i$ is the predicted probability of forest cover from (17) for the i th pixel, N is the number of pixels in the AOI, and A is total area of the AOI. Because A is a constant, the mean probability of forest, $\bar{P}$, is the quantity of interest. When $\hat{p}_i$ is considered an estimator of the superpopulation mean, μ_i , the variance of $\bar{P}$ may be expressed as,

$$\begin{aligned} \text{Var}(\bar{P}) &= \frac{1}{N^2} \text{Var} \left[\sum_{i=1}^N (\hat{p}_i) \right] \approx \frac{1}{N^2} \text{Var} \left[\sum_{i=1}^N (Z_i\hat{\beta}) \right] \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{Cov}(Z_i\hat{\beta}, Z_j\hat{\beta}). \end{aligned} \quad (21a)$$

When $\hat{p}_i$ is considered an estimator of an individual observation, p_i , the variance of $\bar{P}$ may be expressed as,

$$\begin{aligned} \text{Var}(\bar{P}) &= \frac{1}{N^2} \text{Var} \left[\sum_{i=1}^N (\hat{p}_i) \right] \approx \frac{1}{N^2} \text{Var} \left[\sum_{i=1}^N (Z_i\hat{\beta} + \varepsilon_i) \right] \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left[\text{Cov}(Z_i\hat{\beta}, Z_j\hat{\beta}) + \text{Cov}(\varepsilon_i, \varepsilon_j) \right] \end{aligned} \quad (21b)$$

The component $\text{Cov}(Z_i\hat{\beta}, Z_j\hat{\beta})$ in (21a) and (21b) quantifies the covariance between pairs of predictions resulting from their

calculation using a model with parameters estimated from the same sample. An estimator of $\text{Cov}(Z_i\hat{\beta}, Z_j\hat{\beta})$ is,

$$\hat{C}\hat{o}v(Z_i\hat{\beta}, Z_j\hat{\beta}) = Z_j'\hat{V}_{\hat{\beta}}Z_i. \quad (22)$$

The second component within the brackets of (21b), $\text{Cov}(\varepsilon_i, \varepsilon_j)$, quantifies the covariance between pairs of residuals. An estimator for $\text{Cov}(\varepsilon_i, \varepsilon_j)$ is

$$\hat{C}\hat{o}v(\varepsilon_i, \varepsilon_j) = s_i s_j \hat{\rho}_{ij} = \sqrt{\hat{p}_i(1-\hat{p}_i)} \sqrt{\hat{p}_j(1-\hat{p}_j)} \exp(\hat{\alpha}_2 d_{ij}), \quad (23)$$

where $\hat{\rho}_{ij}$ is calculated using the fitted semi-variogram from (14). Substituting from (22) and (23) into (21a) and (21b), an estimator, $\hat{V}\hat{a}r_1(\bar{P})$, for $\text{Var}(\bar{P})$ when $\hat{p}_i$ is considered an estimator of the superpopulation mean, μ_i , is,

$$\hat{V}\hat{a}r_1(\bar{P}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N Z_i'\hat{V}_{\hat{\beta}}Z_j, \quad (24a)$$

and when $\hat{p}_i$ is considered an estimator of an observation, p_i , an estimator, $\hat{V}\hat{a}r_2(\bar{P})$, for $\text{Var}(\bar{P})$ is,

$$\begin{aligned} \hat{V}\hat{a}r_2(\bar{P}) &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left[Z_i'\hat{V}_{\hat{\beta}}Z_j + \sqrt{\hat{p}_i(1-\hat{p}_i)} \right. \\ &\quad \left. \times \sqrt{\hat{p}_j(1-\hat{p}_j)} \exp(\alpha_2 d_{ij}) \right]. \end{aligned} \quad (24b)$$

In (24a) and (24b), the subscripts for the estimators, $\hat{V}\hat{a}r_1(\bar{P})$ and $\hat{V}\hat{a}r_2(\bar{P})$, distinguish the two situations: (1) $\hat{p}_i$ is considered an estimator of the superpopulation mean, μ_i , and (2) $\hat{p}_i$ is considered an estimator of an individual observation, p_i .

3.3. Analyses

Because the relationship between the probability of forest and the spectral data was expected to be approximately constant within each analytical area, the logistic model was calibrated separately for each area using data for FIA plots as previously described with centers in the area. Thus, two sets of logistic model parameters were estimated, one for the Indiana analytical area and one for the Minnesota analytical area. In addition, spatial variability was evaluated for each analytical area separately under the assumptions that spatial correlation is stationary, i.e., it does not change within the scene, and that spatial correlation is isotropic, i.e., it is the same in all directions. Thus, two fitted semi-variograms were used, one for the Indiana analytical area and one for the Minnesota analytical area.

For each analytical area, 15 study area centers were selected, and the probability of forest was predicted for each pixel with center within 15 km of the study area center. For 5-km, 10-km, and 15-km radius circular AOIs centered at each study area center, the design-based estimates, $\bar{Y}$ from (2) and $\hat{V}\hat{a}r(\bar{Y})$ from (3), and the model-based estimates, $\bar{P}$ from (20), $\hat{V}\hat{a}r_1(\bar{P})$ from (24a), and $\hat{V}\hat{a}r_2(\bar{P})$ from (24b), were calculated.

Although $\bar{Y}$ and $\bar{P}$ are different quantities and their estimators are based on different sets of assumptions, as

Table 1
Estimates for the 15-km radius Indiana AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	28	0.1596	0.0500	0.1304	0.0052
2	27	0.1027	0.0357	0.1259	0.0043
3	30	0.0985	0.0467	0.0690	0.0036
4	26	0.2759	0.0672	0.3346	0.0066
5	27	0.5708	0.0667	0.4502	0.0116
6	28	0.8407	0.0492	0.7131	0.0133
7	31	0.2648	0.0597	0.1648	0.0064
8	21	0.4887	0.0917	0.4117	0.0112
9	30	0.1786	0.0547	0.1318	0.0071
10	32	0.4466	0.0707	0.3497	0.0101
11	29	0.3758	0.0723	0.3520	0.0088
12	29	0.3105	0.0649	0.2992	0.0114
13	27	0.1452	0.0404	0.1447	0.0080
14	29	0.6944	0.0610	0.4457	0.0156
15	28	0.2698	0.0578	0.2732	0.0099

shown by (1) and (19), they are similar with respect to their roles in estimating total forest area for an AOI. For the 15 AOIs associated with each combination of analytical area and AOI radius, the design-based estimates, $\bar{Y}$, and the model-based estimates, $\bar{P}$, were compared using three measures: mean difference, mean absolute difference, and root mean squared difference. In addition, the three measures were also calculated for AOIs of the same radius without regard to analytical area.

4. Results and discussion

4.1. Spatial correlation

For each analytical area, the distribution of standardized residuals was approximately symmetrical and centered at zero. An empirical semi-variogram was constructed for each analytical area with no interval having fewer than 200 data points, and the exponential semi-variogram model (12) with nugget of zero ($\alpha_0=0$) was fit to each empirical semi-variogram. The estimated

Table 3
Estimates for the 5-km radius Indiana AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	5	0.3985	0.2025	0.0915	0.0044
2	3	0.2500	0.1443	0.0841	0.0030
3	3	0.0000	–	0.0831	0.0043
4	1	0.0000	–	0.4187	0.0088
5	2	0.6250	0.1250	0.4435	0.0118
6	1	1.0000	–	0.7285	0.0140
7	3	0.1467	0.0754	0.1056	0.0057
8	1	1.0000	–	0.6830	0.0207
9	3	0.2075	0.2075	0.0789	0.0060
10	4	0.3750	0.1614	0.2957	0.0096
11	4	0.2869	0.1627	0.2096	0.0067
12	3	0.7100	0.0400	0.3094	0.0123
13	2	0.1250	0.1250	0.1151	0.0073
14	3	0.6667	0.2205	0.4639	0.0167
15	3	0.0000	–	0.3187	0.0109

ranges of spatial correlation for the standardized residuals, calculated as the distances to 95% of the estimated sills of the semi-variograms, were approximately 120 m for the Indiana analytical area and approximately 100 m for the Minnesota analytical area.

4.2. Bias assessments

The bias assessments focused on comparing the design-based estimate, $\bar{Y}$, and the model-based estimate, $\bar{P}$ (Tables 1–6). The first approach to bias assessment focused on the differences, $\bar{Y}-\bar{P}$, for each combination of analytical area and AOI radius using the three measures: mean difference, mean absolute difference, and root mean squared difference (Table 7). Mean differences were between -0.06 and 0.09 for the 5-km radius AOIs and between -0.02 and 0.06 for both the 10-km and 15-km radius AOIs. Mean absolute differences were less than 0.20 for the 5-km radius AOIs and less than 0.08 for both the 10-km and 15-km radius AOIs. Root mean squared

Table 2
Estimates for the 10-km radius Indiana AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	13	0.2681	0.0956	0.1155	0.0049
2	12	0.1867	0.0621	0.1201	0.0040
3	12	0.0833	0.0641	0.0665	0.0037
4	13	0.2308	0.0871	0.3608	0.0074
5	14	0.5179	0.0962	0.4845	0.0120
6	15	0.8970	0.0531	0.7693	0.0139
7	12	0.2638	0.1031	0.1468	0.0063
8	9	0.5989	0.1409	0.4772	0.0131
9	12	0.1629	0.0624	0.0919	0.0072
10	15	0.3717	0.0946	0.3023	0.0095
11	11	0.3770	0.1101	0.3256	0.0082
12	12	0.3233	0.0955	0.3134	0.0127
13	14	0.1257	0.0570	0.1186	0.0072
14	15	0.6203	0.0945	0.4465	0.0160
15	12	0.1921	0.0782	0.2578	0.0097

Table 4
Estimates for the 15-km radius Minnesota AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	53	0.8803	0.0333	0.8763	0.0030
2	56	0.7517	0.0462	0.8035	0.0041
3	51	0.7655	0.0488	0.7774	0.0039
4	49	0.8411	0.0383	0.8344	0.0047
5	48	0.9341	0.0241	0.9010	0.0040
6	52	0.7297	0.0477	0.6938	0.0054
7	55	0.6872	0.0485	0.7830	0.0044
8	53	0.8294	0.0335	0.8246	0.0044
9	52	0.7189	0.0465	0.6944	0.0046
10	53	0.6149	0.0462	0.7003	0.0057
11	49	0.6708	0.0513	0.6771	0.0059
12	51	0.7119	0.0509	0.7309	0.0054
13	53	0.6506	0.0491	0.6371	0.0053
14	54	0.8169	0.0325	0.8799	0.0035
15	50	0.7094	0.0471	0.7773	0.0044

Table 5
Estimates for the 10-km radius Minnesota AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	24	0.8607	0.0560	0.8671	0.0029
2	25	0.7866	0.0642	0.8294	0.0039
3	23	0.7935	0.0697	0.7463	0.0036
4	21	0.8929	0.0407	0.8338	0.0045
5	20	0.9625	0.0205	0.9029	0.0044
6	25	0.6684	0.0772	0.6542	0.0058
7	22	0.7169	0.0659	0.8364	0.0043
8	24	0.8120	0.0501	0.8187	0.0040
9	24	0.6797	0.0751	0.6660	0.0046
10	19	0.6556	0.0746	0.6913	0.0061
11	22	0.6538	0.0727	0.6813	0.0058
12	23	0.7717	0.0701	0.7787	0.0055
13	24	0.6146	0.0752	0.6447	0.0053
14	19	0.8491	0.0546	0.8819	0.0034
15	24	0.6792	0.0730	0.7961	0.0042

differences were less than 0.24 for the 5-km radius AOIs and less than 0.10 for the 10-km and 15-km radius AOIs. For a given AOI radius, both mean absolute and root mean squared differences were smaller for the Minnesota AOIs than for the Indiana AOIs. This result is partly attributed to better design-based estimates for the Minnesota AOIs as a result of the availability of more plots; for some 5-km Indiana AOIs, only three plots were available. Finally, the graph of the model-based versus design-based estimates for the 10-km radius AOIs indicates no discernible bias relative to the design-based estimates (Fig. 2). The graphs for the 15-km radius AOIs and 5-km radius AOIs were similar to the graph for the 10-km radius AOIs, although there was greater scatter for the former.

4.3. Variance comparisons

Because estimates of means are more precise than estimates of observations, the result that $\text{Var}_1(\bar{P}) < \text{Var}_2(\bar{P})$ was as expected. For the Indiana AOIs, $0.0030 \leq \sqrt{\text{Var}_1(\bar{P})} \leq 0.0207$, and for the Minne-

Table 7
Difference statistics

Scene	AOI radius (km)	Difference		
		Mean	Mean absolute	Root mean squared
Indiana	5	0.09	0.20	0.24
	10	0.05	0.08	0.10
	15	0.06	0.07	0.09
Minnesota	5	−0.06	0.11	0.15
	10	−0.02	0.04	0.05
	15	−0.02	0.03	0.05
Combined	5	0.02	0.15	0.20
	10	0.02	0.06	0.08
	15	0.02	0.05	0.07

sota AOIs, $0.0029 \leq \sqrt{\text{Var}_1(\bar{P})} \leq 0.0086$. The differences between $\sqrt{\text{Var}_1(\bar{P})}$ and $\sqrt{\text{Var}_2(\bar{P})}$ were small, only being evident in the fifth or sixth digit to the right of the decimal point. Because estimates for $\sqrt{\text{Var}_1(\bar{P})}$ and $\sqrt{\text{Var}_2(\bar{P})}$ were so similar, only estimates for $\sqrt{\text{Var}_1(\bar{P})}$ are reported (Tables 1–6). These small differences are attributed to the relatively small range of spatial correlation among the standardized residuals. Semi-variogram ranges on the order of 100 m means that the standardized residual for a 30-m×30-m Landsat TM pixel may be expected to be correlated with the residuals of only 35–50 other pixels and, for many of these, the expected correlation would be small.

5. Conclusions and discussion

Several conclusions may be drawn from this study. First, the model-based estimates of the mean probability of forestland were comparable to the design-based estimates of mean proportion forestland. The forest conditions surrounding the 30 study area centers represented a broad range of species combinations, mostly hardwoods in Indiana and a mixture of conifers and hardwoods in Minnesota, and they represented a broad range of forest conditions including large blocks of contiguous forest in Minnesota and highly fragmented forest and sparse riparian forest in Indiana. In

Table 6
Estimates for the 5-km radius Minnesota AOIs

Circle	Design-based			Model-based	
	No. plots	$\bar{Y}$	$\sqrt{\text{Var}(\bar{Y})}$	$\bar{P}$	$\sqrt{\text{Var}_1(\bar{P})}$
1	5	1.0000	—	0.9152	0.0031
2	5	0.4832	0.2048	0.8274	0.0035
3	5	0.6500	0.1696	0.6446	0.0038
4	5	0.9000	0.0612	0.8837	0.0045
5	5	1.0000	—	0.8970	0.0055
6	4	0.4375	0.1875	0.5158	0.0066
7	5	0.9000	0.0612	0.8526	0.0043
8	6	0.5498	0.1190	0.7313	0.0036
9	6	0.3588	0.1160	0.6892	0.0051
10	7	0.5000	0.1220	0.6594	0.0063
11	8	0.6730	0.1228	0.6559	0.0086
12	8	0.7188	0.1451	0.7942	0.0068
13	4	0.6875	0.1573	0.5917	0.0055
14	5	0.9000	0.0612	0.8790	0.0036
15	5	0.7500	0.1936	0.8142	0.0046

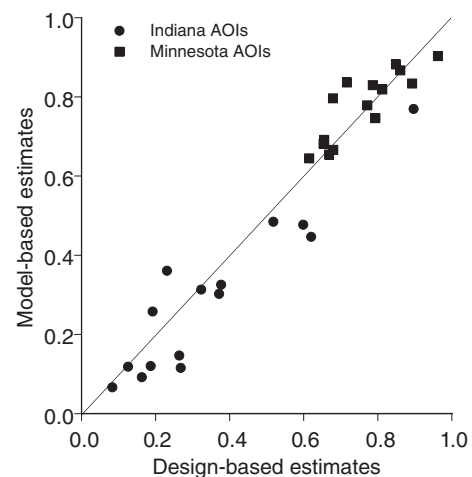


Fig. 2. Model-based vs. design-based estimates for 10-km AOIs.

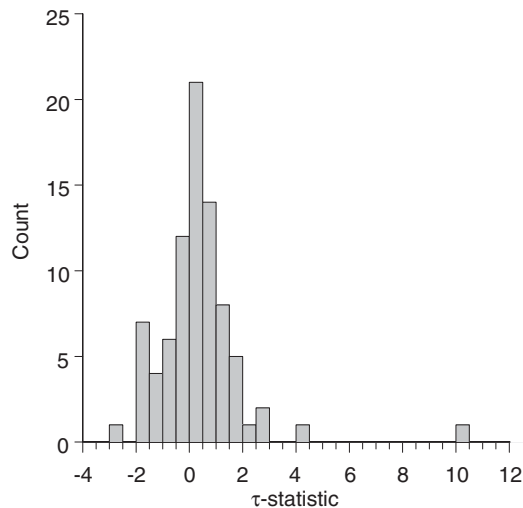


Fig. 3. Distribution of values of $\tau = \frac{\bar{Y} - \bar{P}}{\sqrt{\text{Var}(\bar{Y})}}$ for all AOIs where $\bar{Y}$ is the design-based mean, $\bar{P}$ is the model-based mean, and $\sqrt{\text{Var}(\bar{Y})} > 0$.

addition, the non-forestland, which had to be discriminated from the forestland, represented a range of conditions including intensive agriculture, abandoned agricultural fields, wetlands, and lakes.

Second, the comparability of the design-based estimates, which used data for all plots, regardless of the homogeneity of the ground cover, and the model-based estimates, which were based on model parameter estimates obtained using only completely forested or completely non-forested plots, indicates that exclusion of mixed forest/non-forest plots from the model calibration data produced no discernible effects.

Third, due to the superpopulation assumptions underlying the model-based approach to estimation, special procedures were necessary to accommodate spatial correlation. This accommodation required estimation of the spatial structure of residual variation via semi-variograms that were used in both estimation of the logistic regression model parameters and estimation of the model-based precision for multiple pixel AOIs. The drawbacks of accommodating the spatial correlation are that estimation of the model parameters is more difficult and that estimation of the model-based precision is computer intensive. In addition, for the data sets considered for this study, precision estimates were sensitive to the parameters of the fitted semi-variograms.

Fourth, maps derived from model-based predictions may at least partially circumvent the constraints on release of exact FIA plot locations. Via online access to digital maps, users may obtain estimates of forest area for their own AOIs without access to FIA plots data or locations. In addition, estimates derived from the maps will be possible for smaller AOIs than will be possible using plot data only.

Finally, a few comments related to decisions to change from design-based to model-based estimation are appropriate. In the decision-making process, the tradeoffs between design-based and model-based approaches must be considered from both theoretical and practical perspectives. Although the traditional FIA design-based approach is more familiar and more straightforward, it has drawbacks related to releasing exact plot

locations, failure to address the “Where?” question, and failure to accommodate small area estimation. The logistic regression model-based approach satisfactorily addresses these drawbacks. However, FIA users must be convinced that the cost of overcoming these drawbacks is not too great. Two potential costs have already been identified, greater estimation complexity and greater computational intensity. For online applications, the estimation complexity will be transparent for users. A possible effect of the computational intensity is that online applications will tax the user’s wait-time tolerance. This issue is beyond the scope of this manuscript, but possible solutions include faster hardware and faster computational algorithms. Two additional potential costs are that the model-based estimates may be biased and less precise than the design-based estimates. The issue of bias has been satisfactorily addressed by the comparisons of the design-based and model-based estimates for the circular study areas. In addition, with only a few exceptions, the model-based estimates of AOI means were within two design-based standard errors of the design-based estimates of the same means (Fig. 3).

Comparing the precision of the design-based and model-based estimates is more difficult. From a conceptual perspective, the two sets of precision estimates are not directly comparable, because the assumptions underlying the approaches and the means by which they generate variability are not comparable. In this sense, neither set of precision estimators is inherently superior or preferable; the choice of approach and estimators is arbitrary, and either choice can be justified. However, from a purely practical perspective, users will want to know if less precision in model-based estimates relative to design-based estimates is the cost of obtaining maps and small area estimates and circumventing the plot location release problem. Based on the precision estimates reported in Tables 1–6, the practical answer is that the precision of the logistic regression model-based estimates is greater than the precision of the traditional FIA design-based estimates.

Acknowledgement

The author gratefully acknowledges the GIS and data compilation assistance provided by Geoffrey R. Holden, Forest Inventory and Analysis, North Central Research Station, USDA Forest Service, St. Paul, MN, USA. The author also gratefully acknowledges the careful reviews and constructive suggestions of two reviewers. One reviewer noted a crucial typographical error, and a second reviewer provided particularly relevant comments regarding the consequences of the differences in the assumptions underlying design-based and model-based approaches to estimation.

References

- Ahern, F. J., Erdle, T., MacLean, D. A., & Kneppbeck, I. D. (1991). A quantitative relationship between forest growth rates and Thematic Mapper reflectance measurements. *International Journal of Remote Sensing*, 12, 387–400.
- Albert, P., & McShane, L. (1995). A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. *Biometrics*, 51, 627–638.
- Bechtold, W. A., & Patterson, P. L. (Eds.). (2005). *The enhanced Forest Inventory and Analysis program – National sampling design and estimation procedures*.

- Gen. Tech. Rep. SRS-80 Southern Res. Station, Asheville, NC: USDA Forest Service 85 pp.
- Cohen, W. B., Spies, T. A., & Fiorella, M. (1995). Estimating the age and structure of forests in a multi-ownership landscape of western Oregon, U.S.A. *International Journal of Remote Sensing*, 16, 721–746.
- Crist, E. P., & Ciccone, R. C. (1984). Application of the tasseled cap concept to simulated Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, 50, 343–352.
- Cross, A. M., Settle, J. J., Drake, N. A., & Päivinen, R. (1991). Subpixel measurement of tropical forest cover using AVHRR data. *International Journal of Remote Sensing*, 12, 1119–1130.
- Czaplewski, R. L., & Catts, G. P. (1992). Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sensing of Environment*, 39, 29–43.
- Danson, F. M., & Curran, P. J. (1993). Factors affecting the remotely sensed response of coniferous forest plantations. *Remote Sensing of Environment*, 43, 55–65.
- de Bruin, S. (2000). Predicting the areal extent of land cover types using classified satellite imagery and geostatistics. *Remote Sensing of Environment*, 74, 387–396.
- DeFries, R. S., Hansen, M. C., & Townshend, J. R. G. (2000). Global continuous fields of vegetation characteristic: A linear mixture model applied to multi-year 8 km AVHRR data. *International Journal of Remote Sensing*, 21, 1389–1414.
- Deppe, F. (1998). Forest area estimation using sample surveys and Landsat MSS data. *Photogrammetric Engineering and Remote Sensing*, 64, 285–292.
- Draper, N., & Smith, H. (1981). *Applied regression analyses* (2nd edition). New York: Wiley.
- Gemmell, F. M. (1995). Effects of forest cover, terrain, and scale on timber volume estimation with Thematic Mapper data in a Rocky Mountain site. *Remote Sensing of Environment*, 51, 291–305.
- Gumpertz, M. L., Wu, C., & Pye, J. M. (2000). Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science*, 46(1), 95–107.
- Häme, T., Stenberg, P., Andersson, K., Rauste, Y., Kennedy, P., Folving, S., et al. (2001). AVHRR-based forest proportion map of the Pan-European area. *Remote Sensing of Environment*, 77, 76–91.
- Hansen, M., DeFries, R., Townshend, J. R., Carroll, M., Dimiceli, C., & Sohlberg, R. (2003). *500 m MODIS vegetation continuous fields*. College Park, Maryland: The Global Land Cover Facility Version 1.0.
- Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 national landcover database for the United States. *Photogrammetric Engineering and Remote Sensing*, 70(7), 829–840.
- Iverson, L. R., Cook, E. A., & Graham, R. L. (1989). A technique for extrapolating and validating forest cover across large regions: Calibrating AVHRR data with TM data. *International Journal of Remote Sensing*, 10, 1805–1812.
- Kangas, A. (1994). Classical and model based estimators for forest inventory. *Silva Fennica*, 28(1), 3–14.
- Kauth, R. J., & Thomas, G. S. (1976). The Tasseled Cap – A graphic description of the spectral–temporal development of agricultural crops as seen by Landsat. *Proceedings of the symposium on machine processing of remotely sensed data* (pp. 41–51). West Lafayette, IN: Purdue University.
- Kennedy, P., & Bertolo, F. (2002). Mapping sub-pixel forest cover in Europe using AVHRR data and national and regional statistics. *Canadian Journal of Remote Sensing*, 28, 302–331.
- Maiersberger, T. K., Cohen, W. B., & Gano, L. M. (2001). A TM-based hardwood–conifer mixture index for closed canopy forest in the Oregon Coast Range. *International Journal of Remote Sensing*, 22, 1053–1066.
- Magnussen, S., Boudewyn, P., Wulder, M., & Seemann, D. (2000). Predictions of forest inventory cover type proportions using Landsat TM. *Silva Fennica*, 34, 351–370.
- Mayaux, P., & Lambin, E. F. (1995). Estimation of tropical forest area from coarse spatial resolution data: A two-step correction function for proportional errors due to spatial aggregation. *Remote Sensing of Environment*, 53, 1–15.
- McRoberts, R. E., Bechtold, W. A., Patterson, P. L., Scott, C. T., & Reams, G. A. (2005). The enhanced Forest Inventory and Analysis program of the USDA Forest Service: Historical perspective and announcement of statistical documentation. *Journal of Forestry*, 103(6), 304–308.
- Moisen, G. G., & Frescino, T. S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, 157, 209–225.
- National Atlas of the United States. 2003. <http://nationalatlas.gov>.
- Nel, E. M., Wessman, C. A., & Veblen, T. T. (1994). Digital and visual analysis of Thematic Mapper imagery for differentiating old growth from younger spruce–fir stands. *Remote Sensing of Environment*, 48, 291–301.
- Nelson, R. (1989). Regression and ratio estimators to integrate AVHRR and MSS data. *Remote Sensing of Environment*, 30, 201–216.
- Nelson, M. D., McRoberts, R. E., & Lessard, V. C. (2005). Comparison of forestland area estimates from FIA, NRI, and three satellite image-based land cover datasets. In R. E. McRoberts, G. A. Reams, P. C. Van Deusen, & W. H. McWilliams (Eds.), *Proceedings of the Fifth Annual Forest Inventory and Analysis Symposium. General Technical Report WO-69* (pp. 129–138). Washington, DC: U.S. Department of Agriculture Forest Service.
- Nilson, T., & Peterson, U. (1994). Age dependence of forest reflectance: Analysis of main driving factors. *Remote Sensing of Environment*, 48, 319–331.
- Päivinen, R., Lehtikainen, M., Schuck, A., Häme, T., Väättäinen, S., Kennedy, P., et al., 2001. Combining Earth observation data and forest statistics. Research Report 14, European Forest Institute, Joint Research Center–European Commission, Joensuu, Finland.
- Puhr, C. B., & Donoghue, D. N. M. (2000). Remote sensing of upland conifer plantations using Landsat TM data: A case study from Galloway, southwest Scotland. *International Journal of Remote Sensing*, 21, 633–646.
- Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, 57, 377–387.
- Särndal, C. E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5, 27–52.
- SAS Institute Inc. 1988. SAS/STAT user's guide, Release 6.03. Cary, NC: SAS Institute., 1028 pp.
- Schuck, A., Päivinen, R., Häme, T., Van Brusselen, J., Kennedy, P., & Folving, S. (2003). Compilation of a European forest map from Portugal to the Ural mountains based on earth observation data and forest statistics. *Forest Policy and Economics*, 5, 187–202.
- Steininger, M. K. (2000). Satellite estimation of tropical secondary forest above-ground biomass: Data from Brazil and Bolivia. *International Journal of Remote Sensing*, 21, 1139–1157.
- Taskinen, I. & Heikkinen, J. (unpublished manuscript). A nonparametric Bayesian method for assessing uncertainty in thematic maps of forest variables. (Available at <http://www.metla.fi/pp/JuHe/pub/JABES03083R.pdf>).
- Thomas, G., Hobbs, S. E., & Dufour, M. (1993). Woodland area estimation by spectral mixing: Applying a goodness of fit solution. *International Journal of Remote Sensing*, 17, 291–301.
- Tomppo, E. (1987). *Stand delineation and estimation of stand variables by means of satellite images* (pp. 61–76). Helsinki, Finland: Department of Forest Mensuration and Management, University of Helsinki Res. Note 19.
- Tomppo, E. (1988). *Standwise forest variate estimation by means of satellite images* (pp. 103–111). Helsinki, Finland: Department of Forest Mensuration and Management, University of Helsinki Res. Note 21.
- Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large tree stem volume and aboveground biomass. *Remote Sensing of Environment*, 82, 156–171.
- Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., & 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. *Photogrammetric Engineering and Remote Sensing*, 67, 650–662.
- Walsh, T. A., & Burk, T. E. (1993). Calibration of satellite classifications of land area. *Remote Sensing of Environment*, 46, 281–290.
- Yang, L., Homer, C. G., Hegge, K., Huang, C., Wylie, B., & Reed, B. (2001). A Landsat 7 scene selection strategy for a National Land Cover Database. *Proceedings of the IEEE 2001 International Geoscience and Remote Sensing Symposium, Sydney, Australia, CD-ROM, 1 disc*.
- Zhu, Z., & Evans, D. L. (1994). U.S. forest types and predicted percent forest cover from AVHRR data. *Photogrammetric Engineering and Remote Sensing*, 60, 525–531.