A Proposal for Phase 4 of the Forest Inventory and Analysis Program

Ronald E. McRoberts[1]

Abstract.—Maps of forest cover were constructed using observations from forest inventory plots, Landsat Thematic Mapper satellite imagery, and a logistic regression model. Estimates of mean proportion forest area and the variance of the mean were calculated for circular study areas with radii ranging from 1 km to 15 km. The spatial correlation among pixel predictions was incorporated into the variance calculations. The map-based estimates were compared to estimates obtained using only plot data for the same circular areas. For three circular study areas in Minnesota, the map-based estimates were similar to the plot-based estimates and more precise.

The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture (USDA) Forest Service surveys the Nation's forested lands using a combination of field plot measurements and remotely sensed data. Traditionally, the FIA program has used data from these inventories to respond to the "How Much?" question by reporting plot-based estimates of forest attributes by county or combinations of counties. The estimates are obtained using a three-phase approach: Phase 1 consists of using remotely sensed data, often satellite imagery, to stratify the land area of interest for increasing the precision of estimates, and, optionally, to determine if a plot has accessible forest land; Phase 2 consists of measuring a suite of mensurational variables for plots with accessible forest land; and Phase 3 consists of measuring a suite of variables related to forest health for a 1:16 subset of the Phase 2 plots.

Increasingly, users are also asking "Where?" and are requesting access to plot data for estimating for their own areas of interest (AOI), frequently for use as training or accuracy assessment data for spatial products. Many of these applications require the coordinates of plot locations to determine if a plot is located in the AOI. Although forest inventory programs generally release plot information to the public, many resist releasing plot locations. First, release of plot locations may entice users to visit plots for additional data, which could artificially disturb the ecology of the sites and contribute to bias in inventory estimates. Second, many forest inventory programs rely on the goodwill of private forest landowners for permission to observe plots on their land. Landowners generally do not welcome unwarranted or frequent intrusions and often only permit visits by inventory crews contingent on assurances that plot locations and proprietary information will not be released.

In response to the "Where?" question, the FIA program has initiated local, regional, and national mapping efforts. Although the objectives of these efforts have been to map the spatial distribution of forest attributes, generally they have not included investigations of whether the maps may be used to produce unbiased and precise estimates of forest attributes. For the latter objective, plot-based estimation using data for plots located in the AOI has been necessary. However, if unbiased and sufficiently precise estimates of forest attributes could be obtained from maps, then several advantages would accrue. First, release of plot locations would be unnecessary for estimation for user AOIs. Second, because mapped values for individual mapping units 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.

The objective of the study was to investigate estimation of forest attributes from maps constructed using inventory plot data and satellite imagery. In this context, FIA Phase 4 is defined as the construction of forest attribute maps that satisfy two criteria: (1) estimates at all spatial scales are within tolerance limits relative to plot-based estimates, and (2) estimates of the precision of the map-based estimates are comparable to the precision of plot-based estimates. For this study, the terms plot-based and map-based estimates are considered equivalent to the terms design-based and model-based estimates, respectively. Although the second set of terms may be regarded as more correct from a

[1] Mathematical Statistician, U.S. Department of Agriculture, Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108.

statistical perspective, the terms of the first set are generally used for this study because they are considered more descriptive. The study focused on constructing maps depicting pixel-level probabilities of forest ground cover using inventory plot data, Landsat Thematic Mapper (TM) satellite imagery, and a logistic regression model, and then comparing the map-based and plot-based estimates of forest land area.

Data

FIA Plot Data

Under the FIA program's annual inventory system (McRoberts 1999), field plot centers are established in permanent locations using a systematic sampling design that is assumed to produce a random, equal probability sample. In each State, a fixed proportion of plots is measured annually. For the FIA program of the North Central Research Station , plots measured in a single federal fiscal year (e.g., FY 2004: 1 October 2003 to 30 September 2004) comprise a single panel of plots with panels selected for annual measurement on a 5-year rotating basis. In aggregate, over a complete 5-year measurement cycle, a plot represents approximately 2,400 ha (slightly less than 6,000 ac). In general, locations of forested or previously forested plots are determined using global position system receivers, while locations of non-forested plots are determined using digitization methods.

Each field plot consists of four 7.31-m (24-foot) radius circular subplots. The subplots are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0^{o}, 120^{o}, and 240^{o} 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 forest land proportion are obtained by aggregating ground land use conditions into forest and non-forest classes consistent with the FIA definition of forest land, and plot-level estimates are obtained as means over the four subplots. For this study, all plots were observed or measured between the beginning of FY1999 and the end of FY2002. Although the locations of the central subplots are considered a random, equal probability sample, the fixed spatial configuration of the peripheral subplots with respect to the central subplots requires accommodation of spatial correlation among subplot observations for subplot-level analyses.

Satellite Imagery

Landsat TM imagery for scenes Row 27 and Row 28 of Path 27 was obtained from the Multi-Resolution Characterization 2001 land cover mapping project (Homer *et al*., in press) of the U.S. Geological Survey. The imagery was characterized by several salient features: (1) a combination of Landsat 5 and Landsat 7 Enhanced TM+ data, (2) three dates including early, peak, and late vegetation green-up (Yang *et al*., 2001), (3) geometrically and radiometrically corrected, (4) cubic convolution resampling to 30 m x 30 m spatial resolution, (5) visible and infrared bands (1-5, 7), and (6) conversion to at-satellite reflectance.

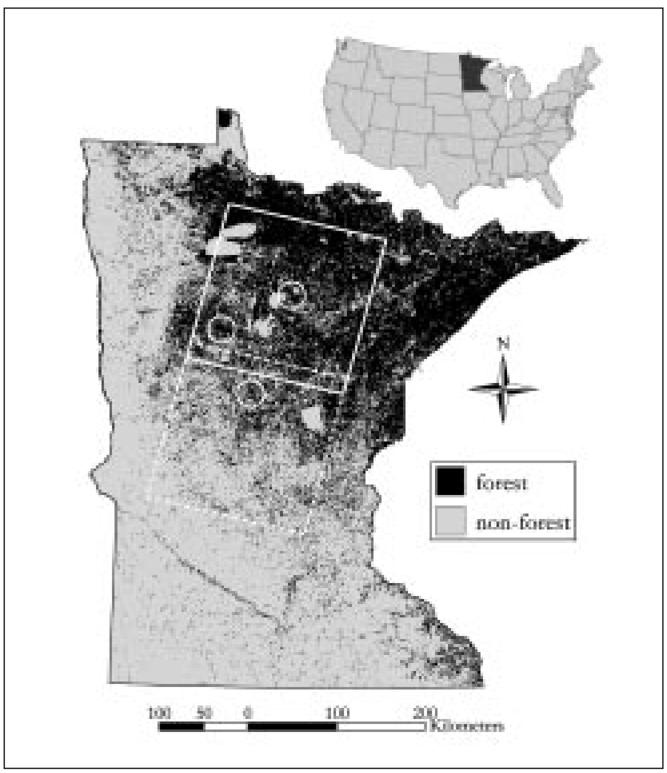
For predicting the probability of forest cover, the satellite image spectral data were used in two forms: the 18 raw spectral band data and 12 transformations including normalized difference vegetation index (NDVI) and the tree tassel cap transformations (brightness, greenness, and wetness) for each image date (Kauth and Thomas 1976). The raw band data and the transformations were evaluated separately.

Combining FIA Data and Satellite Imagery

The spatial configuration of the FIA subplots with centers separated by 36.58 m and the 30 m x 30 m spatial resolution of the TM imagery permits individual subplots to be associated with individual image pixels. Further, the subplot area of 167.87 m^2 is approximately 19 percent of the 900 m^2 pixel area and is generally deemed an adequate sample of the ground characteristics of the pixel area. However, when describing relationships between forest attributes and satellite image spectral values, two phenomena must be considered. First, because a subplot is a single, 19-percent, contiguous sample of the pixel area, the proportion forest for a partially forested subplot may not accurately represent the proportion forest for the entire pixel. Second, Global Positioning System 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 values, 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 map-based estimates of forest attributes. To avoid these phenomena, when fitting models data were excluded for plots with a mixture of forest and non-forest cover. An important result of the exclusion of these data was that proportion forest area, Y, was a binary variable. This variable was quantified according to the convention that Y = 1 denoted a forested subplot and Y = 0 denoted a non-forested subplot.

Within the two TM scenes, three 15-km radius circular study areas were arbitrarily selected to represent a spectrum of forest/non-forest ground cover conditions (fig. 1). Because the relationship between spectral values and the forest/non-forest composition of ground cover was expected to be approximately constant over the scenes, data for calibrating models were combined for the three study areas. In addition, spatial variability was evaluated for the three study areas collectively 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).

Figure 1.—*Study areas (white circles represent study areas; solid white rectangle denotes TM Path 28 Row 27; dashed white rectangle denotes TM Path 28 Row 28).*



Methods

Model Calibration

The relationship between the binary forest/non-forest dependent variable, Y, and the continuous spectral value independent variables, X, may be expressed as,

$$p_i = f(X_i;\beta) + \varepsilon_i, \tag{1a}$$

or

$$E(p_i) = f(X_i;\beta), \tag{1b}$$

where i denotes subplot, 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 variables and the parameters, ε_i is unexplained residual uncertainty, and E(.) denotes statistical expectation. For binary data, $f(X_i,\beta)$ is often expressed as a logistic model of which one form is,

$$f(X_i;\beta) = \frac{1}{1+\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im})}, \tag{2a}$$

and another is,

$$f(X_i;\beta) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im})}{1+\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im})}. \tag{2b}$$

Each of the two forms, (2a) and (2b), may be expressed as one minus the other, and parameter values for the two forms are the same, except that the signs are reversed.

The parameters of (2) are often estimated by maximizing the likelihood, L, expressed as,

$$L = \prod_{i=1}^{n} (p_i)^{Y_i} (1-p_i)^{1-Y_i}, \tag{3}$$

where n is the number of subplot observations. Maximum likelihood parameter estimates using (3) may be obtained using SAS PROCs LOGISTIC, CATMOD, or GLIM (SAS 1988), although care must be exercised to assure the model form, (2a) or (2b), to which the parameter estimates apply. However, because (3) cannot accommodate spatial correlation among subplot observations, another approach to parameter estimation must be considered if correct measures of uncertainty are required. Nevertheless, because (3) 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 based on generalized estimating equations (GEEs) as described by Albert and McShane (1995) and Gompertz *et al*. (2000) was used to accommodate spatial correlation. For the first iteration, ordinary logistic regression using maximum likelihood was used to fit (2) to the data. The GEE approach consisted of solving,

$$\sum_{i=1}^{n} Z_i' V_\varepsilon^{-1}(Y_i - p_i) = 0, \tag{4}$$

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

$$z_{ij} = \frac{\partial f(X_i;\hat{\beta})}{\partial \beta_j}, \tag{5}$$

and $f(X_i;\hat{\beta})$ is the logistic function, (2), evaluated using the parameter estimates, $\hat{\beta}$. The subplot residual covariance matrix, V_e, was recalculated following each iteration. The standardized residuals,

$$\varepsilon_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}, \tag{6}$$

were calculated and used to define an empirical semivariogram,

$$\hat{\gamma}(h) = \frac{1}{2n_h}\sum_{i,j}^{n_h}(\varepsilon_i - \varepsilon_j)^2, \tag{7}$$

where h is the distance between subplots, n_h is number of subplots h km apart, and the subscripts i and j indicate the i^{th} and j^{th} subplots, respectively. The exponential semivariogram,

$$\hat{\gamma}(h) = \alpha_0 + \alpha_1[1 - \exp(\alpha_2 h)], \tag{8}$$

was fit to the empirical semivariogram using weighted nonlinear least squares techniques. The elements, v_{ij}, of the subplot residual covariance matrix, V_e, were estimated as,

$$v_{ij} = \sqrt{\hat{p}_i(1-\hat{p}_i)}\sqrt{\hat{p}_j(1-\hat{p}_j)}\frac{\hat{\alpha}_1}{\hat{\alpha}_0 + \hat{\alpha}_1}\exp(\hat{\alpha}_2 h). \tag{9}$$

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

$$\hat{\beta}^{k+1} = \hat{\beta}^k + \left(\sum_{i=1}^{n} Z_i' V_\varepsilon^{-1} Z_i\right)^{-1} \sum_{i=1}^{n} Z_i' V_\varepsilon^{-1}(Y_i - \hat{p}_i). \tag{10}$$

The iterative procedure of updating $\hat{\beta}$ and recalculating V_e continued until convergence. The covariance matrix, $\mathbf{V}_\beta$, for the parameter estimates was approximated as,

$$\hat{V}_{\hat{\beta}} \cong \left(\sum_{i=1}^{n} Z_i' V_\varepsilon^{-1} Z_i\right)^{-1} \tag{11}$$

Individual variables from among the 18 raw spectral bands and the 12 transformations were selected for inclusion in the model by repeatedly fitting the model using the GEE technique after eliminating the least significant variable among those that did not contribute significantly to improving the quality of fit.

Map-Based Estimation

Because the estimate of forest area for an AOI may be expressed as the product of the AOI's total area and an estimate of its expected proportion forest land, the remaining discussion focuses on the estimation of the expected proportion forest land and the precision of the estimate. For the i^{th} pixel in the AOI, the probability, p_i, that $Y_i = 1$ was estimated as,

$$\hat{p}_i = \frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{ij} + \ldots + \hat{\beta}_m X_{im})} \tag{12}$$

where $\hat{\beta}$ is the vector of parameter estimates obtained from (10). The estimate, $\hat{P}$, of the expected proportion forest, P, for the entire AOI is the mean of the probability estimates over all pixels,

$$\hat{P} = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_i, \tag{13}$$

where N is the number of image pixels in the AOI. The variance of $\hat{P}$ was estimated as,

$$\mathrm{Var}(\hat{P}) = \frac{1}{N^2}\left\{\left[\sum_{i=1}^{N}\sum_{j=1}^{N} Z_i V_{\hat{\beta}} Z_j\right] + \left[\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(\varepsilon_i, \varepsilon_j)\right]\right\} \tag{14}$$

where $V_{\hat{\beta}}$ is the covariance matrix for the parameter estimates from (11) and ε is residual uncertainty. The first component within the brackets of (14) quantifies the uncertainty in the estimate because predictions for each pixel are obtained from a model with parameter estimates obtained from the same sample. The second component within the brackets of (14) quantifies the effects on uncertainty of residual variation around the model predictions. The derivation of (14) is provided in the appendix.

An estimate of correlation among residuals at a distance h, $\hat{\rho}$ (h), was obtained from the relationship,

$$\hat{\rho}_h = 1 - \frac{\hat{\gamma}(h)}{\hat{\gamma}_{tot}},$$

where $\hat{\gamma}(h)$ is obtained from (8) and $\hat{\gamma}_{tot} = \hat{\alpha}_0 + \hat{\alpha}_1$. Thus, in application, $Cov(\varepsilon_i,\varepsilon_j)$ was estimated as,

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = \sqrt{\hat{p}_i\left(1-\hat{p}_i\right)}\sqrt{\hat{p}_j\left(1-\hat{p}_j\right)}\frac{\hat{\alpha}_1}{\hat{\alpha}_o + \hat{\alpha}_1}\exp\left(\hat{\alpha}_2 h\right) \quad (15)$$

where h is the distance between the centers of the i^{th} and j^{th} pixels.

Plot-Based Estimation

Plot-based estimates of the mean and variance of proportion forest land were calculated as a standard of comparison for map-based estimates. When calculating these estimates, data were included for all four subplots of each plot having any portion of a subplot in the AOI. In addition, all plots in the AOI were selected for calculation of the mean and variance, regardless of the homogeneity of the ground cover. Inclusion of subplots with a mixture of forest and non-forest ground cover meant that proportion forest area could not be considered a binomial variable when calculating the variance of the estimate of the mean. Because the sampling design was considered to produce an equal probability, random sample, central subplots of all plots were considered to be randomly located, and spatial correlation among observations from different plots was not considered. However, the fixed spatial configuration of peripheral subplots with respect to their corresponding central subplots meant that spatial correlation, ρ, among subplot observations of the same plot could not be ignored. Large area analyses indicated that the correlation between central and peripheral subplot observations for the same plot was approximately 0.9, while the correlation between peripheral subplot observations of the same plot was approximately 0.8. Estimates of the variances of the plot-based mean proportion forest land area estimates were calculated as,

$$\begin{aligned} Var\left(\bar{Y}\right) &= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var\left(Y_i\right) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov\left(Y_i, Y_j\right)\right] \\ &= \frac{1}{n^2}\left[nVar\left(Y\right) + 2Var\left(Y\right)\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho_{ij}\right] \\ &= \frac{Var\left(Y\right)}{n^2}\left\{n + 2\left[\frac{n}{4}3\left(0.9\right) + \frac{3n}{4}\left(0.8\right)\right]\right\} \\ &= \frac{Var\left(Y\right)}{n}\left(3.55\right) \end{aligned} \quad (16)$$

where i denotes subplots, n is the number of subplots in the AOI, $\frac{n}{4}$ is the number of central subplots, and $\frac{3n}{4}$ is the number of peripheral subplots.

Analyses

The inventory plot data and the satellite image spectral data were pooled for the three study areas, and a single set of parameters for model (2a) was estimated for all three study areas. For each circular area, the prediction of the probability of forest land was calculated for each pixel, and a map of the predicted probabilities was created. Within each 15-km radius circular study area, the map-based estimates $\hat{P}$ and Var ($\hat{P}$)were calculated using (13) and (14), respectively, for circular areas with radii ranging from 1 km to 15 km centered at the center of the study area. In addition, the plot-based estimates $\bar{Y}$ and $Var\left(\bar{Y}\right)$ were calculated using (15) and (16), respectively, for the same circular areas.

Results and Discussion

The NDVI and tassel cap transformations were better predictors of the probability of forest ground cover than were the raw spectral band data. However, these results may vary for different ground covers and for satellite imagery of different dates. Because of the relatively small numbers of plots in each study area and the common set of parameter estimates for all three study areas, a single semivariogram was fit to residual data collectively for all three study areas. The fitted semivariogram indicated that spatial correlation did not extend beyond 350 m.

The inventory data, satellite imagery, and logistic model produced considerable detail in the maps depicting the probability of forest land for each pixel (fig. 2). Comparisons of the map-based and plot-based estimates of mean proportion forest land and estimates of the standard errors of the estimates yielded three primary results (table 1). First, the map-based estimates of mean proportion forest land were very similar to the plot-based estimates. Of the 37 circular areas for which the plot-based standard error was greater than zero, all except four map-based estimates of proportion forest land were within two plot-based standard errors of the plot-based estimates. For these exceptional four cases, t-tests using plot-based variances in the denominator of the t-statistic yielded P = 0.04, P=0.04, P=0.02, and P<0.001.

Table 1.—*Mean proportion forest estimates by study area.*

Radius (km)	Plot-based estimates			Map-based estimates	
	No. plots	Mean	SE	Mean	SE
Study Area 1					
1	0	-----	-----	0.4803	0.1038
2	0	-----	-----	0.6585	0.0563
3	1	1.0000	0.0000	0.7370	0.0468
4	2	0.5000	0.3561	0.7578	0.0449
5	3	0.6667	0.2678	0.7184	0.0418
6	6	0.8333	0.1464	0.7002	0.0384
7	6	0.8333	0.1464	0.6691	0.0359
8	7	0.8571	0.1269	0.6467	0.0299
9	10	0.7000	0.1383	0.5965	0.0281
10	13	0.6154	0.1284	0.5965	0.0281
11	16	0.6094	0.1158	0.6002	0.0309
12	18	0.6300	0.1072	0.6111	0.0314
13	23	0.5490	0.0975	0.6223	0.0317
14	26	0.5914	0.0906	0.6271	0.0321
15	29	0.5992	0.0855	0.6279	0.0332
Study Area 2					
1	0	-----	-----	0.8390	0.0069
2	1	1.0000	0.0000	0.8362	0.0400
3	1	1.0000	0.0000	0.8199	0.0378
4	1	1.0000	0.0000	0.7587	0.0402
5	3	0.6667	0.2678	0.7684	0.0381
6	7	0.8571	0.1269	0.7768	0.0359
7	7	0.8571	0.1269	0.7756	0.0398
8	7	0.8571	0.1269	0.7758	0.0359
9	12	0.9167	0.0760	0.7599	0.0390
10	12	0.9167	0.0760	0.7372	0.0361
11	14	0.9286	0.0654	0.7309	0.0351
12	17	0.8676	0.0780	0.7216	0.0351
13	22	0.8010	0.0796	0.7133	0.0344
14	24	0.8170	0.0737	0.7099	0.0333
15	32	0.7065	0.0755	0.7032	0.0332
Study Area 3					
1	0	-----	-----	0.7563	0.0388
2	1	0.7500	0.4710	0.7506	0.0570
3	1	0.7500	0.4710	0.7430	0.0714
4	2	0.8750	0.2355	0.7016	0.0608
5	4	0.9375	0.1178	0.6846	0.0675
6	5	0.9500	0.0942	0.6563	0.0705
7	6	0.7917	0.1596	0.6378	0.0719
8	8	0.8438	0.1229	0.6341	0.0726
9	10	0.6750	0.1413	0.6358	0.0717
10	13	0.7500	0.1142	0.6347	0.0694
11	16	0.6692	0.1111	0.6316	0.0713
12	18	0.5949	0.1093	0.6245	0.0726
13	23	0.5525	0.0979	0.6119	0.0734
14	27	0.5447	0.0905	0.5971	0.0742
15	30	0.5236	0.0861	0.5760	0.0735

Figure 2a.—*Probability of forest for Study Area 1 (white=non-forest, black=forest).*

Figure 2b.—*Probability of forest for Study Area 2 (white=non-forest, black=forest).*

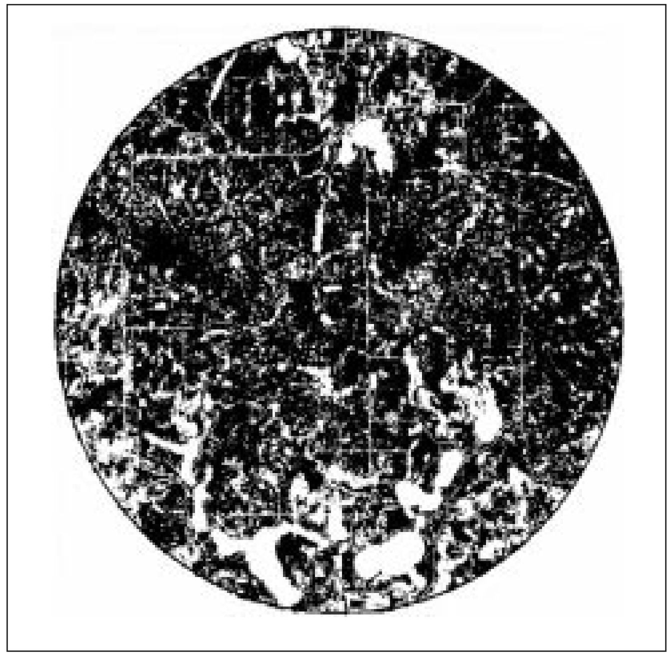
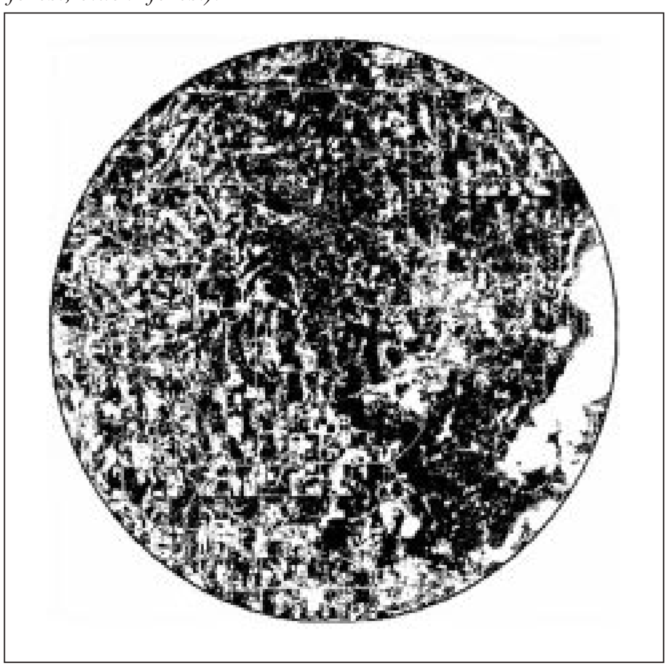


Figure 2c.—*Probability of forest for Study Area 3 (white=non-forest, black=forest).*



However, it must be noted that, on the one hand, these P-values are conservative because the denominator of the t-statistic did not account for uncertainty in the map-based estimates. On the other hand, P-values for circular areas of different radii in the same study area are not independent, because the plot data used to obtain an estimate for any particular circular area uses plot data for all circular areas of smaller radii in the same study area. Nevertheless, these results indicate that the map-based and plot-based estimates are quite similar. Second, the map-based estimates revealed a smooth transition as the radii of the circular areas increased from 1 km to 15 km. This result suggests that even when the number of plots in a small circular area was insufficient to obtain a reliable plot-based estimate, a reliable map-based estimate was possible. Third, the map-based standard errors were consistently smaller than the plot-based standard errors.

Two conclusions are warranted from this study. First, for forest land area, map-based estimation is not only feasible but also may produce estimates that are comparable, if not superior, to plot-based estimates. Second, although map-based estimation for other forest attributes is expected to be more difficult than for forest land area, the results of this study are sufficient to encourage the FIA program to initiate a Phase 4 focusing on construction of maps for estimation.

Literature Cited

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.

Gompertz, 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.

Homer, C.; Huang, C.; Yang, L.; Wylie, B.; Coan, M. [In press]. Development of a 2001 national landcover database for the United States. Photogrammetric Engineering and Remote Sensing.

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. West Lafayette, IN: Purdue University: 41–51.

McRoberts, R.E. 1999. Joint annual forest inventory and monitoring system: the North Central perspective. Journal of Forestry. 97: 21–26.

SAS Institute Inc. 1988. SAS/STAT user's guide, release 6.03. Cary, NC: SAS Institute. 1028 p.

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.

Appendix

The estimate, $\hat{P}$, of the expected proportion forest, P, for an entire area of interest is simply the mean of the probability estimates, $\hat{p}_i$, over all pixels,

$$\hat{P} = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_i ,$$

where N is the number of image pixels in the AOI. The objective of the appendix is to provide a derivation of the estimate for the variance of $\hat{P}$. First, however, two intermediate results are demonstrated.

The first intermediate result is that for $Z_i = (z_{i1},z_{i2},\ldots,z_{in})$, a vector of constants, and $\Delta' = (\delta_1,\delta_2,\ldots,\delta_n)$, a random vector distributed $N(0,V_\Delta)$,

$$E\left[(Z_i\Delta)(Z_j\Delta)\right] = Z_i V_\Delta Z_j' . \qquad \text{(A1)}$$

Although (A1) is not proven for the general case, a demonstration for the case of n=2 is provided.

$$\begin{aligned}
E\left[(Z_i\Delta)(Z_j\Delta)\right] &= E\left[(z_{i1}\delta_1 + z_{i2}\delta_2)(z_{j1}\delta_1 + z_{j2}\delta_2)\right] \\
&= E\left(z_{i1}\delta_1 z_{j1}\delta_1 + z_{i1}\delta_1 z_{j2}\delta_2 + z_{i2}\delta_2 z_{j1}\delta_1 + z_{i2}\delta_2 z_{j2}\delta_2\right) \\
&= E\left(z_{i1}\delta_1^2 z_{j1} + z_{i1}\delta_1\delta_2 z_{j2} + z_{i2}\delta_2\delta_1 z_{j1} + z_{i2}\delta_2^2 z_{j2}\right) \\
&= z_{i1}E(\delta_1^2)z_{j1} + z_{i1}E(\delta_1\delta_2)z_{j2} + z_{i2}E(\delta_2\delta_1)z_{j1} + z_{i2}E(\delta_2^2)z_{j2} \\
&= z_{i1}\mathrm{Var}(\delta_1)z_{j1} + z_{i1}\mathrm{Cov}(\delta_1\delta_2)z_{j2} + z_{i2}\mathrm{Cov}(\delta_1\delta_2)z_{j1} + z_{i2}\mathrm{Var}(\delta_2)z_{j2} \\
&= Z_i V_\Delta Z_j' .
\end{aligned}$$

The general case follows by analogy.

The second intermediate result is that for $Z_i = (z_{i1},z_{i2},\ldots,z_{in})$, a vector of constants, and $\Delta' = (\delta_1,\delta_2,\ldots,\delta_n)$, a random vector distributed $N(0,V_\Delta)$,

$$E\left[(Z_1\Delta + Z_2\Delta + \ldots + Z_n\Delta)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} Z_i V_\Delta Z_j' \qquad \text{(A2)}$$

Although (A2) is not proven for the general case, a demonstration for n=2 is provided.

$$\begin{aligned}
E\left[(Z_1\Delta + Z_2\Delta)^2\right] &= E\left[(Z_1\Delta)(Z_1\Delta) + (Z_1\Delta)(Z_2\Delta) + (Z_2\Delta)(Z_1\Delta) + (Z_2\Delta)(Z_2\Delta)\right] \\
&= E\left[(Z_1\Delta)(Z_1\Delta)\right] + E\left[(Z_1\Delta)(Z_2\Delta)\right] + E\left[(Z_2\Delta)(Z_1\Delta)\right] + E\left[(Z_2\Delta)(Z_2\Delta)\right] \\
&= Z_1 V_\Delta Z_1' + Z_1 V_\Delta Z_2' + Z_2 V_\Delta Z_1' + Z_2 V_\Delta Z_2' \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} Z_i V_\Delta Z_j' . \qquad \text{by (A1)}
\end{aligned}$$

The extension of this result to the general case also follows by analogy.

The primary result is that for p_i equal to the probability that $Y_i = 1$ (i.e., that the i^{th} pixel has forested ground cover), $\hat{p}_i = f(X_i;\hat{\beta})$, and $\hat{P} = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_i$, the variance of $\hat{P}$ may be estimated as,

$$\mathrm{Var}(\hat{P}) = \frac{1}{N^2}\left\{\left[\sum_{i=1}^{N}\sum_{k=1}^{N} Z_i V_{\hat{\beta}} Z_k'\right] + \left[\sum_{i=1}^{N}\sum_{k=1}^{N}\mathrm{Cov}(\varepsilon_i,\varepsilon_k)\right]\right\} \quad \text{(A3)}$$

where $V_{\hat{\beta}}$ is the estimated covariance matrix for the estimates of $\boldsymbol{\beta}$, Z_i is a vector with elements,

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

and ε_i is a residual. The proof follows.

$$\begin{aligned}
\mathrm{Var}(\hat{P}) &= E(\hat{P} - P)^2 \\
&= E\left\{\left[\frac{1}{N}\sum_{i=1}^{N}\hat{p}_i\right] - \left[\frac{1}{N}\sum_{i=1}^{N}Y_i\right]\right\}^2 \\
&\cong \frac{1}{N^2}E\left\{\left[\sum_{i=1}^{N}Z_i\hat{\beta}\right] - \left[\sum_{i=1}^{N}(Z_i\beta + \varepsilon_i)\right]\right\}^2 \\
&= \frac{1}{N^2}E\left\{\left[\sum_{i=1}^{N}Z_i(\hat{\beta} - \beta)\right] + \left(\sum_{i=1}^{N}\varepsilon_i\right)\right\}^2 \\
&= \frac{1}{N^2}E\left\{\left[\sum_{i=1}^{N}Z_i(\hat{\beta} - \beta)\right]^2 + 2\left[\sum_{i=1}^{N}Z_i(\hat{\beta} - \beta)\right]\left(\sum_{i=1}^{N}\varepsilon_i\right) + \left(\sum_{i=1}^{N}\varepsilon_i\right)^2\right\} \\
&= \frac{1}{N^2}E\left\{\left[\sum_{i=1}^{N}Z_i(\hat{\beta} - \beta)\right]^2 + \left(\sum_{i=1}^{N}\varepsilon_i\right)^2\right\} \\
&= \frac{1}{N^2}E\left\{\left[\sum_{i=1}^{N}\sum_{k=1}^{N}Z_i(\hat{\beta} - \beta)Z_k(\hat{\beta} - \beta)\right] + \left(\sum_{i=1}^{N}\sum_{k=1}^{N}\varepsilon_i\varepsilon_k\right)\right\} \\
&= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\sum_{k=1}^{N}E\left[Z_i(\hat{\beta} - \beta)Z_k(\hat{\beta} - \beta)\right] + \sum_{i=1}^{N}\sum_{k=1}^{N}E(\varepsilon_i\varepsilon_k)\right\} \\
&= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\sum_{k=1}^{N}E\left[Z_i(\hat{\beta} - \beta)Z_k(\hat{\beta} - \beta)\right] + \sum_{i=1}^{N}\sum_{k=1}^{N}\mathrm{Cov}(\varepsilon_i\varepsilon_k)\right\} \\
&= \frac{1}{N^2}\left\{\left[\sum_{i=1}^{N}\sum_{k=1}^{N}Z_i V_{\hat{\beta}} Z_k'\right] + \left[\sum_{i=1}^{N}\sum_{k=1}^{N}\mathrm{Cov}(\varepsilon_i,\varepsilon_k)\right]\right\} \quad \text{by (A2).}
\end{aligned}$$