

STRATIFIED ESTIMATES OF FOREST AREA USING THE k-NEAREST NEIGHBORS TECHNIQUE AND SATELLITE IMAGERY

Ronald E. McRoberts, Mark D. Nelson, and Daniel G. Wendt¹

ABSTRACT.—For two study areas in Minnesota, stratified estimation using Landsat Thematic Mapper satellite imagery as the basis for stratification was used to estimate forest area. Measurements of forest inventory plots obtained for a 12-month period in 1998 and 1999 were used as the source of data for within-strata estimates. These measurements further served as calibration data for a k-Nearest Neighbors technique that was used to predict forest land proportion for image pixels. The continuum of forest land proportion predictions were separated into strata to facilitate stratified estimation. The variances of the stratified forest area estimates were smaller than variances based on simple random estimates by factors as great as 5, and when including all plots over a 5-year plot measurement cycle, the forest area precision estimates may be expected to satisfy national standards.

The five regional Forest Inventory and Analysis (FIA) programs of the Forest Service, U.S. Department of Agriculture, report estimates of forest land area for their respective regions every 5 years. Each estimate is obtained as the product of total area inventoried and the mean over a systematic array of field plots of the proportion of each plot in FIA-defined forest land. The FIA definition of forest land includes commercial timberland, some pastured land with trees, forest plantations, unproductive forested land, and reserved, non-commercial forested land. In addition, forest land must satisfy minimum stocking levels, a 0.405-ha (1-acre) minimum area, and a minimum continuous canopy width of 36.58 m (120 ft). It therefore excludes lands such as wooded strips, idle farmland with trees, and narrow wind-breaks. A combination of budgetary constraints and natural variability among plots prohibits sample sizes sufficient to satisfy national FIA precision standards for forest land estimates unless the estimation process is enhanced using ancillary data.

One approach to enhancing the estimation process is to use stratified estimation with classified satellite imagery as the

basis for the stratification. With this approach, image pixels for the area of interest are classified with respect to predictions of land cover attributes into homogeneous classes, and the classes are then used as strata in the stratified analyses. Strata weights are the proportions of pixels in strata, and plots are assigned to strata on the basis of the strata assignment of their associated pixels. If the stratification is accomplished prior to sampling and the within-strata variances of the inventory variables are well-estimated, then maximum precision may be achieved by designing within-strata sampling intensities to be proportional to within-strata variances. However, even when the within-strata sampling intensities are independent of the stratification, stratified estimation may still yield increases in precision.

The timeliness of FIA estimates is enhanced when cycles for obtaining and classifying imagery are comparable to the 5-, 7-, or 10-year plot measurement cycles, depending on region. On average, a regional FIA program on a 5-year plot measurement cycle will need to classify approximately 125 TM images over the cycle. In addition, sufficient training data to guide the classifications must be obtained in close temporal proximity to the imagery dates. These are not insignificant tasks, and investigation of efficient means of obtaining training data and processing images are worthwhile FIA endeavors. The objective of this study is to investigate the utility of the k-Nearest Neighbors technique for processing TM imagery to be used as the basis for enhancing forest area estimates through stratification.

¹ Mathematical Statistician, and Computer Specialist, respectively, North Central Research Station, U.S. Department of Agriculture, Forest Service, St. Paul, MN 55108. Phone: (651) 649-5174; fax: (651) 649-5285; e-mail: rmcroberts@fs.fed.us; and Computer Specialist, R9, U.S. Department of Agriculture, Milwaukee, WI.

DATA

Study Areas

The study was conducted in two areas, designated St. Louis and St. Cloud (fig. 1). The St. Louis study area encompasses most of St. Louis County, Minnesota; includes approximately 2.1 million hectares of which approximately 75 percent is forest land; and is dominated by Aspen-Birch and Spruce-Fir associations. The St. Cloud study area contains the St. Cloud, Minnesota, urban area; includes approximately 3.3 million hectares of which approximately 20 percent is forest land; and is characterized by prairie agriculture and a diverse mixture of forest lands including both coniferous and deciduous species.

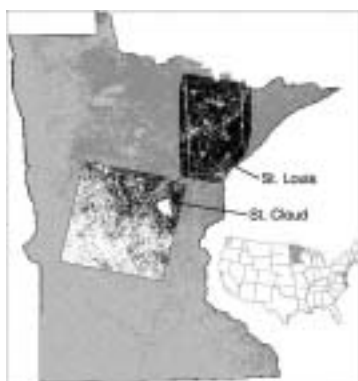


Figure 1.—Minnesota study areas.

Satellite Imagery

The St. Louis study area is covered by the Landsat TM Path 27, Row 27 scene and includes all of St. Louis County except the northern portion. For this scene, Landsat-7 ETM+ images were obtained for two seasons: autumn (5 November 1999) and summer (31 May 2000). The St. Cloud study area is covered by the Landsat TM Path 28, Row 28 scene. For this scene, Landsat-7 ETM+ images were obtained for three seasons: summer (23 July 1999), autumn (27 October 1999), and spring (3 March 2000). The following attributes pertain to all five images: (1) 30 x 30 m pixels from bands 1-5 and band 7; (2) absolute radiance units scaled to 8 bits; (3) processing to level 1G (processing level 08; radiometrically and geometrically corrected using satellite model and platform/ephemeris information); and (4) geo-referencing to Albers Equal Area projection, NAD83. In addition, for the St. Louis study area, the November image was rectified using 40 ground control points with resulting root mean square error of 12.1 m. The

May image was registered to the November image using 26 ground control points and resampled using first-order polynomial and nearest neighbor techniques with resulting root mean square error of 31.9 m. For the St. Cloud study area, all three images were rectified using ground control points and digital elevation model terrain correction (processing level 10) and resampled using cubic convolution with resulting root mean square error less than 8.5 m. Finally, bands are distinguished using an alphanumeric character representing the first letter of the month of the image and a numeric character designating the band. The context of band references indicates whether they refer to St. Louis or St. Cloud images.

FIA PLOT DATA

Under the FIA program's annual inventory system (McRoberts 1999), field plots are established in permanent locations using a systematic sampling design. In each State, a fixed proportion of plots are measured annually; plots measured in a single Federal fiscal year (e.g., FY-1999: 1 October 1998 to 30 September 1999) make up a single panel of plots with panels selected for annual measurement on a rotating basis. In aggregate, over a complete measurement cycle, a plot represents 2,403 ha (5,937 acres). 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-ft) 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°, 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-level estimates of forest land proportion are obtained by aggregating these ground land use conditions consistent with the FIA definition of forest land, and plot-level estimates are obtained as means over the four subplots.

For both study areas, measurements for the FY-1999 panel of inventory plots were available. For the St. Louis study area, measurements for 133 plots or 532 subplots were used of which 387 subplots were completely forested, 7 subplots were partially forested, and 138 subplots were non-forested. For the St. Cloud study area, measurements for 268 plots or 1,072 subplots were used of which 226 subplots were completely forested, 13

subplots were partially forested, and 833 subplots were non-forested.

K-NEAREST NEIGHBORS TECHNIQUE

The k-Nearest Neighbors (k-NN) technique is a non-parametric approach to predicting values of point variables on the basis of similarity in a covariate space between the point and other points with observed values of the variables. For this application, consider a TM image pixel to be a point, let Y denote a ground attribute (e.g., forest land proportion, cumulative volume of individual trees, tree density) of the pixel, and let X denote its vector of TM spectral values. Let U denote the finite set of image pixels, and let U be indexed by $i=1, \dots, N$; let S denote the subset of pixels of U that correspond to FIA subplots (i.e., the subset of pixels of U for which surrogates for the pixel-level observations of the ground attribute are available), and let S be indexed by $j=1, \dots, n$. The objective is to obtain a prediction of the ground attribute for each pixel. With the k-NN technique, predictions, $\hat{Y}_i$, for Y_i are obtained in two steps:

1. for each $(Y_i, X_i) \in U$ re-order $\{(Y_j, X_j) | Y_j \in S\}$ with respect to increasing distance, d_{ij} , between X_i and each X_j , excluding Y_i from the ordering if $(Y_i, X_i) \in S$, also; denote the resulting ordering $\{(Y_{ij}, X_{ij})\}$;
2. for each $Y_i \in U$,

$$\hat{Y}_i = \left(\sum_{j=1}^k w_{ij} \right)^{-1} \left(\sum_{j=1}^k w_{ij} Y_{ij} \right) \quad (1)$$

where k is a predetermined constant, $1 \leq k < n$, and $\{w_{ij}\}$ are point weights to be selected.

The quality of predictions may be assessed using $\{Y_j | (Y_j, X_j) \in S\}$, the root mean square error (RMSE) criterion,

$$RMS_e = \sqrt{\left(Y_j - \hat{Y}_j \right)^2} \quad (2)$$

and the leaving-one-out method. With the leaving-one-out method, a k-NN prediction, $\hat{Y}_j$, is sequentially obtained for each Y_j , but with the provision that Y_j itself cannot be included in the mean forming its own k-NN prediction. In addition, to avoid

issues related to the expected high correlation expected among attributes for subplots of the same plot, for this study the prediction for a subplot was constrained against including an observation for any of the other three subplots of the same plot. By comparing the observations, $\{Y_j | (Y_j, X_j) \in S\}$, and corresponding predictions, $\{\hat{Y}_j | (Y_j, X_j) \in S\}$, with respect to the RMS_e criterion, the quality of predictions may be evaluated.

Before implementation, the k-NN technique must be calibrated. First, the particular spectral bands used to calculate the distances, d_{ij} , between X_i and each element of the set, $\{X_j | (Y_j, X_j) \in S\}$, must be selected. Let Z denote the subset of X consisting of the selected bands, and let the elements of Z be indexed by $m=1, \dots, M$. Second, a distance metric, d , must be selected; among the alternatives are weighted Euclidean distance,

$$d_{ij} = \left[\sum_{m=1}^M v_m (Z_{im} - Z_{jm})^2 \right]^{-1} \quad (3)$$

where $\{v_m\}$ are variable weights, and Mahalanobis distance,

$$d_{ij} = (Z'_i - Z'_j)' V^{-1} (Z'_i - Z'_j) \quad (4)$$

where V is the covariance matrix for $Z_j \subset X_j$. If weighted Euclidean distance is selected, then the variable weights $\{v_m\}$ for (3) must also be selected. Third, the value of k , the number of nearest neighbors to be included in the calculation of predictions (1), must be selected. Finally, the point weights, $\{w_{ij}\}$, for (1) must be selected; common alternatives include constant weighting for which $w_{ij}=1$, inverse distance weighting for which $w_{ij}=d_{ij}^{-1}$, and inverse distance squared weighting for which $w_{ij}=d_{ij}^{-2}$.

The k-NN analyses were conducted at the subplot-pixel level, because a plot-level approach would require calibration using means of inventory observations over the four subplots and either means of TM spectral values over the four pixels corresponding to the four subplots or means over a block of pixels covering the plot. Predictions for image pixels must likewise then be based on the mean over four pixels in the same configuration as the four pixels corresponding to the four subplots or the mean over a block of pixels of the same size and configuration as the block covering the plot. For this study, subplot-pixel level analyses entail a simpler approach without sacrificing statistical validity. Thus, each subplot was associated with the TM pixel with center closest to the subplot center.

ANALYSES

For this application, the selected k-value was $k=k_{opt}$, the k-value that minimizes RMS_e . For each study area, k_{opt} was determined for each combination of spectral bands by comparing values of RMS_e obtained with constant variable and constant point weighting. For each study area, the five spectral band combinations with the smallest RMS_e without regard to the number of bands, were selected for further evaluation.

Creating Strata

For each of the five best spectral band combinations for each study area, forest land proportion was predicted for each image pixel using the k-NN technique with the k_{opt} determined for that band combination. From the resulting continuum of predictions for each image, four optimal strata were selected by considering all possible divisions of the continuum into four classes under three constraints: first, the lower bound of the first stratum was always 0.00, and the upper bound of the fourth stratum was always 1.00; second, the minimum stratum width was 0.05; and third, at least five plots were required to be assigned to each stratum. Stratifications were limited to four strata, because the preponderance of observed forest land proportions were either 0.00 or 1.00. Each pixel was assigned to a stratum based on its forest land proportion prediction, and strata weights were calculated as the proportions of pixels assigned to strata. To avoid the mathematical complexity necessary to accommodate the spatial correlation among the four subplot observations, FIA assigns plots rather than subplots to strata for stratified analyses. Each plot was assigned to a stratum on the basis of the stratum assignment of the pixel corresponding to the center of the center subplot. Plots were stratified using predictions of forest land proportion for their corresponding pixels rather than observations so that the assignment of plots to strata would be consistent with the calculation of strata weights. Stratifications were evaluated with respect to relative efficiency, RE, the ratio of the variance of the mean obtained using simple random analyses and the variance obtained using stratified analyses.

Stratified Estimation

Stratified estimates of mean forest land proportion, $\bar{Y}$, and estimated variance, $\text{Var}(\bar{Y})$, were calculated using standard methods (Cochran 1977)

$$\bar{Y} = \sum_{j=1}^J w_j \bar{Y}_j \quad (5)$$

and

$$\text{Var}(\bar{Y}) = \sum_{j=1}^J w_j^2 \frac{\hat{\sigma}_j^2}{n_j} \quad (6)$$

where $j=1, \dots, J$ denotes stratum; w_j is the weight for the j^{th} stratum; $\bar{Y}_j$ denotes the mean forest land proportion for plots assigned to the j^{th} stratum; n_j is the number of plots assigned to the j^{th} stratum; and $\hat{\sigma}_j^2$ is the within-stratum variance for the j^{th} stratum calculated as,

$$\hat{\sigma}_j^2 = (n_j - 1)^{-1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 \quad (7)$$

where Y_{ij} is the forest land proportion observed by the field crew for the i^{th} plot in the j^{th} stratum. Variance estimates obtained using (6) ignore the slight effects due to finite population correction factors and to variable rather than fixed numbers of plots per stratum.

The FIA program reports precision estimates as coefficients of variation scaled to compensate for varying sample sizes using as a reference standard the sample size corresponding to 404,694 ha (1 million acres) (USDA FS 1970). For forest area estimate, $\hat{FA} = A \bar{Y}$, the scaled precision estimate, denoted PREC, is defined for this study as

$$\text{PREC} = \left[\frac{\hat{\text{Var}}(\hat{FA})}{\hat{FA}} \right]^{\frac{1}{2}} \left[\frac{\hat{FA}}{404,694} \right]^{\frac{1}{2}} = \left[\frac{\hat{\text{Var}}(\hat{FA})}{404,694(A\bar{Y})} \right]^{\frac{1}{2}} \quad (8)$$

where $\bar{Y}$ is again mean forest area proportion per plot, and A is total area inventoried in ha. Two values of PREC are reported, the value obtained from (8) which corresponds to the sample size resulting from a single panel of plot measurements and the value obtained from (8) divided by the square root of 5 which corresponds to the value expected with the sample size resulting from all five panels of plots. The national FIA precision standard is $\text{PREC} \leq 0.03$.

RESULTS

The general results are that the k-NN algorithm was very simple to implement, straightforward to calibrate, and required no user intervention after initiation. The k-NN predictions captured

much of the forest/nonforest detail and provided an excellent basis for stratifications. When compared to variances of forest area estimates obtained using simple random estimation, the variances obtained using stratified estimation were smaller by factors as great as 5. Specific results follow.

Both similarities and differences were noted among calibrations for the five best band combinations and the resulting stratified estimates (table 1). The similarities:

1. the means for the five best band combinations were comparable within study areas;
2. values of RMS_e , SE, RE, and PREC were generally of the same order of magnitude both within and between study areas;
3. the bands selected for the five best band combinations were similar within study areas with N3, N4, and M4 selected for all five combinations for the St. Louis study area, and J3, M3, and M4 selected for all five combinations for the St. Cloud study area; bands 3 and 4 were most commonly selected, while bands from the spring 2000 images were selected for all five best band combinations for each study area;
4. for both study areas, the stratifications based on the k-NN analyses produced expected five-panel precision for forest land area estimates that satisfied the national FIA precision standards;
5. for each best band combination, multiple sets of between-strata boundaries produced similar values of RE.

The differences:

1. the ordering of the band combinations with respect to RE, or equivalently PREC, was not the same as that with respect to RMS_e , suggesting that if the optimal band combination is desired, then evaluating the five best band combinations selected with respect to RMS_e , as was done for this study, is recommended;
2. optimal between-strata boundaries for the St. Louis study area differed considerably, although as noted previously multiple sets of between-strata boundary combinations for the same band combination produced similar values of RE.

CONCLUSIONS AND DISCUSSION

Four important conclusions may be drawn from this study: first, although the k-NN technique is conceptually easy to implement, careful attention must be paid to its calibration if optimal results are expected; second, stratifications derived from TM imagery reduced variances of forest area estimates by factors as great as 5 for both a heavily forested area and a sparsely forested area; third, the stratifications may be expected to produce forest land area estimates that satisfy national FIA precision standards for sample sizes corresponding to five panels of measurements; and fourth, the k-NN technique is a viable alternative for processing satellite imagery that is both faster and easier to implement than traditional image classification methods.

The implications of the latter three conclusions for the FIA program are considerable. First, in the absence of stratification, sample sizes would have to be increased by factors as least as great as 5 to achieve the same level of precision as was obtained with the stratifications. The magnitude of the resulting cost saving is substantial. For the State of Minnesota, with a sampling intensity of one plot for every 2,403 ha, approximately 825 plots are field-measured annually at an FY1999 cost of approximately \$1,000 (US) per plot. Thus, the annual cost savings obtained with such stratifications is approximately \$3,300,000 (US).

Second, the effectiveness of the k-NN algorithm frees the FIA program from more costly and less timely alternatives. The speed and automation of the k-NN technique make it vastly superior to FIAs time-consuming, labor-intensive, traditional approach based on interpreting aerial photographs. A crew of four photointerpreters, working full-time, could be expected to complete the photointerpretation and stratification task for the State of Minnesota in 2-3 years. Alternatively, processing of the approximately 15 TM images necessary for full coverage of the State of Minnesota could be expected to be accomplished using the k-NN technique in 2-3 weeks by a single computer technician.

Table 1.—*Mean forest land proportion estimates*

Rank ^a	RMS _e	Bands		k	Optimal between- strata boundaries	Mean	SE ^b	RE ^c	PREC ^d 1 panel	PREC ^d 5 panels			
St. Louis study area													
1	0.2652	N3	N4	M4	9	0.20	0.75	0.95	0.7547	0.0175	4.4836	0.0455	0.0203
2	0.2687	N1	N3	N4	M4	7	0.10	0.80	0.7493	0.0188	3.8864	0.0490	0.0219
3	0.2690	N2	N3	N4	M4	9	0.20	0.75	0.7593	0.0177	4.3943	0.0459	0.0205
4	0.2690	N3	N4	M1	M4	11	0.55	0.90	0.7845	0.0157	5.5912	0.0400	0.0179
5	0.2693	N2	N3	N4	M1	M4	13	0.50	0.7681	0.0182	4.1655	0.0469	0.0210
St. Cloud study area													
1	0.2392	J2	J3	M3	M4	21	0.10	0.75	0.2312	0.0107	4.9189	0.0635	0.0285
2	0.2399	J2	J3	M1	M3	29	0.15	0.70	0.2313	0.0107	4.9587	0.0635	0.0284
3	0.2406	J2	J3	M2	M3	33	0.20	0.75	0.2346	0.0109	4.7837	0.0643	0.0287
4	0.2420	J3	M3	M4		23	0.25	0.60	0.2367	0.0103	5.3847	0.0605	0.0270
5	0.2423	J3	M1	M3	M4	23	0.25	0.65	0.2398	0.0105	5.1238	0.0612	0.0274

^a Rank based on RMS_e.

^b Standard error of mean, calculated as the square root of the variance of the mean (6).

^c Relative efficiency of the stratification, calculated as the ratio of the variance of the mean assuming simple random sampling and the variance of the mean based on stratified analyses (6).

^d FIA precision calculated from (8).

Finally, better future results may be expected with the k-NN technique. Fine-tuning the calibration of the k-NN technique by including point- and variable-weighting will increase the accuracy of classifications. Also, five panels of plot measurements will increase the density of observations in spectral space, allow each k-NN prediction to be based on subplot-pixel observations in closer spectral proximity, and, therefore, increase the accuracy of individual pixel predictions.

In conclusion, the k-NN technique is a viable and efficient method for processing TM images to obtain predictions of forest area proportion, and stratifications derived from these predictions produce forest area estimates that may be expected to satisfy national FIA precision standards.

LITERATURE CITED

- Cochran, W.G. 1977. Sampling techniques, 3d ed. New York, NY: Wiley. 428 p.
- Hansen, M.H. 1990. A comprehensive sampling system for forest inventory based on an individual tree growth model, St. Paul, MN: University of Minnesota, College of Natural Resources. 256 p. Ph.D. dissertation.
- Loetsch, E. Haller, K.E. 1964. Forest inventory, volume I: statistics of forest inventory and information from aerial photographs. BLV Verlagsgesellschaft, Munchen. 436 p.
- McRoberts, R.E. 1999. Joint annual forest inventory and monitoring system: the North Central perspective. *Journal of Forestry*. 97(12): 21-26.
- USDA Forest Service. 1970. Operational procedures. For. Serv. Handb. 4809.11, Chapt. 10:11-1. Washington, DC: U.S. Department of Agriculture, Forest Service.