K-Nearest Neighbor Estimation of Forest Attributes: Improving Mapping Efficiency

Andrew O. Finley, Alan R. Ek, Yun Bai, and Marvin E. Bauer[1]

Abstract.—This paper describes our efforts in refining k-nearest neighbor forest attributes classification using U.S. Department of Agriculture Forest Service Forest Inventory and Analysis plot data and Landsat 7 Enhanced Thematic Mapper Plus imagery. The analysis focuses on FIA-defined forest type classification across St. Louis County in northeastern Minnesota. We outline three steps in the classification process that highlight improvements in mapping efficiency: (1) using transformed divergence for spectral feature selection, (2) applying a mathematical rule for reducing the nearest neighbor search set, and (3) using a database to reduce redundant nearest neighbor searches. Our trials suggest that when combined, these approaches can reduce mapping time by half without significant loss of accuracy.

The k-nearest neighbor (kNN) multisource inventory has proved timely, cost-efficient, and accurate in the Nordic countries and initial U.S trials. (Franco-Lopez *et al*. 2001, Haapanen *et al*. 2004, McRoberts *et al*. 2002). This approach for extending field point inventories is ideally suited to the estimation and monitoring needs of Federal agencies, such as the U.S. Department of Agriculture (USDA) Forest Service, that conduct natural and agricultural resource inventories. It provides wall-to-wall maps of forest attributes, retains the natural data variation found in the field inventory (unlike many parametric algorithms), and provides precise and localized estimates in common metrics across large areas and various ownerships.

At a pixel-level classification, the kNN algorithm assigns each unknown (target) pixel the field attributes of the most similar reference pixels for which field data exists. Similarity is defined in terms of the feature space, typically measured as Euclidean or Mahalanobis distance between spectral features. The kNN algorithm is not mathematically complex; however, using multiple image dates and features from each date, along with several thousand field reference observations, makes kNN pixel-based mapping of large areas very inefficient. Specifically, the kNN classification approximates to $F \cdot N$ distance calculations, where F is the number of pixels to classify and N is the number of references. For example, standard kNN mapping of a 1.3×10^6 ha area, with a pixel resolution of 30 m^2, and approximately 1,500 FIA field reference observations requires about 22 billion distance calculations and around 16 hours to process on a Pentium 4, single-processor computer.

Our study examined using USDA Forest Service Forest Inventory and Analysis (FIA) plot data and Landsat 7 Enhanced Thematic Mapper Plus (ETM+) imagery in kNN classification of FIA-defined forest types. Specific emphasis is placed on improving mapping efficiency by reducing classification feature space, decreasing the number of distance calculations in the nearest neighbor search, and eliminating redundancy in redundant nearest neighbor searches by building a database of feature patterns associated with different forest type classes.

Study Area and Data

Study Area

St. Louis County, in northeastern Minnesota, is located in the FIA aspen-birch unit. For a detailed description of the study area, see Bauer *et al.* (1994).

FIA Plot Data

The FIA program began fieldwork for the sixth Minnesota forest inventory in 1999. This effort also initiated a new annual inventory or monitoring system. In this new system, approximately one-fifth of the field plots in the State are measured each year. The new inventory protocol collected field data on the four-subplot cluster plot configuration (USDA Forest Service 2000). This plot design consists of four 1/60-ha, fixed-radius, circular subplots linked as a cluster, with each of the three outer subplots located

[1] Department of Forest Resources, College of Natural Resources, University of Minnesota, 1530 Cleveland Ave. N., St. Paul, MN 55108. E-mail: afinley@gis.umn.edu

36.6 m from the center subplot. FIA assigns each subplot to the land use class recorded at the subplot center. The 2001 inventory sampled 1,853 forested subplots in our study area. We removed 83 subplots because they fell under cloud-covered areas in the Landsat imagery. We removed an additional 58 subplots because the FIA field crew and FIA algorithm disagreed in the subplot forest type. The subsequent analysis used the remaining 1,712 subplots.

Satellite Imagery

We used Landsat 7 ETM+ satellite images for the analysis. The study area fell within two Landsat image scenes—path 27, rows 26 and 27. Bands 1 to 5 and 7 of three year 2000 dates were used including a late winter scene from March 12, a spring scene from April 29 and a late spring scene from May 31.

The images were geo-referenced to the Universal Transverse Mercator coordinate system using the following parameters: spheroid GRS80, datum NAD83, and zone 15. The resampling method was nearest neighbor using a 30-m by 30-m pixel size. The geo-referencing reference map was road vectors from the Minnesota Department of Transportation. For image portions with few roads, we used the U.S. Geological Survey digital orthophoto quads from the years 1991 to 1992 with 3-m resolution. The number of control points used in geo-referencing was 38 to 46 per date in path 27, row 27 images and 20 to 22 in 27/26 images. A second order polynomial regression model was used to fit the image. The root mean square error for all six images was less than 8 m. The clouds were digitized by hand and a cloud mask was created.

A forest/nonforest mask was generated using a kNN classification described in Haapanen *et al*. (2004). This mask was used to define the area extent of our forest type classification.

Methods

k-Nearest Neighbor Algorithm

For estimating with Euclidean distances, consider the spectral distance $d_{p_i,p}$, which is computed in the feature space from the target pixel p to each reference pixel p_i for which the forest type class is known. For each pixel p, sort the k-nearest field plot pixels (in the feature space) by $d_{p_i,p} \le ... \le d_{p_k,p}$. The imputed value for the pixel p is then expressed as a function of the closest units, each such unit weighted according to this distance decomposition function:

$$w_{p_i,p} = \frac{1}{d^t_{p_i,p}} \Bigg/ \sum_{j=1}^{k} \frac{1}{d^t_{p_j,p}}, \quad (1)$$

where t is a distance decomposition factor set equal to 1 for all trials. To impute class variables such as forest type, the distance decomposition function calculates a weighted mode value.

For a class variable, the error rate (*Err)* indicates the disagreement between a predicted value $\hat{y}$ and the actual response y in a dichotomous situation such that y does or does not belong to class i (Efron and Tibshirani 1993). Thus, we used the overall accuracy (*OA*) (Stehman 1997, Congalton 1991) defined as follows:

$$OA = 1 - Err, \quad (2)$$

where

$$Err = \sum_{i=1}^{n} \left(y_i - \hat{y}\right) \Big/ n \quad (3)$$

This is a special case of the mean square error for an indicator variable. These estimators were preferred over the usual Kappa estimator for reasons given by Franco-Lopez *et al.* (2001).

Errors were estimated by leave-one-out cross-validation. This technique omits training sample units one by one and mimics the use of independent data (Gong 1986). For each omission, we applied the kNN prediction rule to the remaining sample. Subsequently, the errors from these predictions were summarized. In total, we applied the prediction rule n times and predicted the outcome for n units. Such estimates of prediction error are nearly unbiased (Efron and Tibshirani 1993).

Spectral Feature Selection

As described by McRoberts *et al.* (2002), it is useful to select a parsimonious set of image features to use in the nearest neighbor search. Specifically, McRoberts *et al.* caution that including features unrelated to the attribute being estimated can reduce classification accuracy. When using a non-weighted Euclidian measure for the minimum distance criterion, the inclusion of these unrelated features directly reduces the class discriminating power of the entire feature set.

Instead of testing all combinations of spectral features in our analysis set, we used the statistical separability measure of transformed divergence to find feature subsets that adequately

discriminate among forest type classes. As described by Swain and Davis (1978), measures of divergence can be used to select a subset of feature axes that maximally separate class density functions. The degree to which class density functions diverge, or are separated in a multidimensional space, determines the classification accuracy of parametric classifiers. This approach to feature subsetting should also be effective with non-parametric classifiers such as kNN.

We used the transformed divergence measure implemented in ERDAS, Inc. Imagine® geospatial imaging software to derive feature subsets. Then kNN classification accuracy statistics were generated for each subset and judged against the classification accuracy of the full set of 18 features (i.e., 6 bands from each of 3 Landsat images). The smallest feature subset that performed at least as well as the full feature set was then moved forward in the analysis.

Stratification

In a similar study using the kNN classifier to characterize forested landscape in northeastern Minnesota, Franco-Lopez *et al*. (2001) found that simply stratifying by upland and lowland significantly improved forest type classification accuracy. Based on these findings, we divided the study area into upland and lowland strata as delineated by the U.S. Wildlife and Fish Service (USWFS) National Wetland Inventory. These strata were then classified separately using their respective subplots.

Nearest Neighbor Search Reduction

As previously noted, using minimum Euclidian distance as a nearest neighbor criterion is not mathematically complex; however, mapping large areas computing $F \cdot N$ distance calculations can take significant computer processing time.

Ra and Kim (1993) proposed the mean-distance-ordered partial codebook search (MPS) algorithm to reduce the number of Euclidian distance calculations required in a nearest neighbor search. The first component of the algorithm is the minimum distance criterion, defined by Ra and Kim as the squared Euclidian distance (SED):

$$d_{E,i} = \sum_{j=i}^{m} \left(x_j - c_{ij} \right)^2 , \quad (4)$$

where x_j and c_{ij} are the j^{th} component of the target and reference vector respectively, and m is the dimension of the vector (i.e., number of features). The next element in the algorithm is the squared mean distance (SMD), defined as follows:

$$d_{M,i} = \left(\sum_{j=i}^{m} x_j - \sum_{j=i}^{m} c_{ij} \right)^2 . \quad (5)$$

The algorithm calculates and sorts the first k nearest neighbor distances in the reference set. Then, the SMD is calculated for the $k + 1$ vector in the reference set. This value is then tested with this inequality:

$$d_{M,i} \leq m d_{E,\max} , \quad (6)$$

where $d_{E,max}$ is the largest distance in the sorted set of k nearest neighbors. If this inequality is true, the SED is calculated for the $d_{M,i}$ and the set of $k + 1$ nearest neighbors is resorted and the maximum value is discarded. If the inequality is false, the $d_{M,i}$ reference vector is discarded. This procedure is repeated for every subsequent vector in the reference set until each is either included in the k nearest neighbor set or rejected.

Depending on the amount of dispersion in the reference set, the MPS algorithm can significantly reduce the number of Euclidian distance calculations required to classify a given target pixel. Specifically, a reference set that contains observations that are highly dispersed in feature space will require fewer SED calculations to find the k nearest neighbor set when compared to a reference set that contains observations that are underdispersed. Further, as k increases, the number of observations that pass the inequality will also increase and need to be considered using a full Euclidian comparison. Therefore, analyses with lower values of k will be more time efficient than analyses with higher values of k. We evaluate the usefulness of the MPS algorithm by its ability to reduce the average number of Euclidean distance calculations for different levels of k in the classification of our study area.

Database-Assisted Mapping

The use of the kNN classifier, or any classifier, relies on a correlation between characteristics of the target pixel (e.g., spectral features) and the characteristics of observations in a reference set for which additional information is known. It is this correlation

that allows for meaningful assignment of class-specific reference pixel information to target pixels. In forest type classification, for example, we hope for low variability of spectral features in a forest type class and high variability of spectral features among forest type classes. This desirable "within" relationship versus "among" class variability can also help to increase mapping efficiency.

When implementing a typical kNN classification, the algorithm discards the target/reference similarity information and imputed value after each pixel is processed. If within class variability is low, the typically discarded information could be saved and reused successfully to classify pixels elsewhere in the image. Both the storage and subsequent retrieval of this information would be more efficient than computing the kNN for a given image pixel. Using this premise, we tested the marginal efficiency of incorporating a database system into a kNN image classification.

Using the MySQL database system and the MySQL++ API, we designed a program that would insert, search, and retrieve records that hold pixel spectral features and the kNN imputed values associated with each feature pattern. To allow the database to efficiently search for a feature pattern, it was necessary to discretize the 0–256 range of the Landsat bands into a smaller number of units, referred to in this report as bins.

Many methods are available to discretize continuous data ranging from arbitrary bin assignments across the variable's distribution to using complex algorithms for deciding bin range and placement (Chmielewski and Grzymala-Busse 1996). For our study, we used a relatively simple approach based on the normal probability density function. Each band in the image was divided into the same number of bins. The range and placement of the bins was contingent on the band's mean and standard deviation. This approach holds the area under the bands' theoretical distribution equal for each bin. That is, the bins that occurred near the mean are narrow and the bins on the distribution's tails are proportionally wider. Bin counts of 6, 8, and 10 were tested for database search and retrieval efficiency.

Repeated discretization of a variable's distribution will ultimately result in information loss. A balance must be struck between the amount of reduction in classification accuracy and improved classification efficiency. For this reason, we compared the information loss and efficiency gain through reducing the bin counts.

Our database-assisted mapping program started by making a connection with a predefined MySQL database. The database contained one table with columns to hold the discretized image features and a column to hold the kNN estimated imputation value. For more efficient record insertion, search, and retrieval, the database existed as a hash table in the main memory.

For each image pixel, the feature pattern was extracted and compared to all records in the database. If a match was found, the pixel received the imputation from the matching database record; otherwise, the kNN algorithm was used to assign the value. Each time the kNN algorithm was implemented to assign a classification value, the associated feature pattern and resulting imputation value were inserted in the database.

As noted above, for this database-assisted mapping to be useful, the information insertion, search, and retrieval process must be more efficient than implementing a single kNN search of the reference data set. Further, the discretization process required to make pattern matching efficient must not significantly degrade classification accuracy. Therefore, we evaluated the utility of database-assisted mapping though a series of time tests and classification accuracy comparisons.

Results and Discussion

Spectral Feature Selection

Using the transformed divergence measure implemented in ERDAS Imagine, we derived optimal subsets of 16, 14, 12, 10, 8, and 6 spectral features. Subsets that contained fewer than 14 features produced suboptimal classification accuracy and degraded confusion matrices. Therefore, the remainder of our analysis was performed using the 14 spectral feature subset.

Stratification

Based on results from previous studies, we divided our study area by upland and lowland strata. These areas were delineated based on USWFS National Wetland Inventory classification maps. Approximately 10 percent of the forested landscape in the study area was designated as lowland and contained 149 FIA subplots. The upland portion of the study area contained the remaining 1,563 FIA subplots.

Table 1.—*Forest type group classification confusion matrix for combined upland and lowland strata, using* k = *7.*

		FIA Forest Type Groups					
		100	120	800	900	R. tot.	P. Acc.
Observed	100	37	67	1	71	176	43.53
	120	28	562	0	93	683	81.21
	800	0	0	0	28	28	0.00
	900	20	63	0	742	825	79.44
	C.tot.	85	692	1	934	1712	

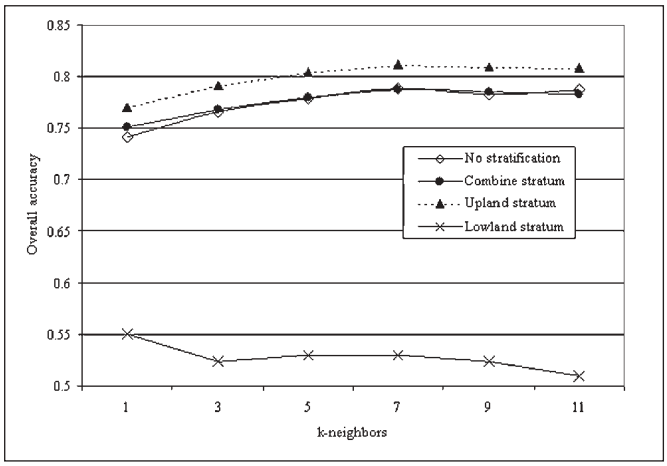
Recognizing that most classification errors resulted from within forest type group confusion, we collapsed the 12 forest types into their respective base groups. Figure 1 shows stratified and nonstratified classification accuracy for the four forest type groups sampled in our study area. Table 1 presents the combined classification confusion metrics for both strata. The spruce-fir group (forest type code 120) and aspen-birch group (forest type code 900) show satisfactory classification accuracy. Generalizing to the group level, however, does not address poor classification of the maple-beech-birch group (forest type code 800) or the aspen-birch group's overclassification.

Nearest Neighbor Search Reduction

Substituting the brute force nearest neighbor search, which considers the distance to all reference observations, with the MPS algorithm significantly reduced the number of distance calculations needed to classify each pixel. In leave-one-out cross-validation trials of 1,712 observations, each consisting of 14 features, we recorded the average number of observations in the $n - 1$ reference set that failed the inequality described in Equation 6. The trial results were $k = 1$ (74.4 percent failed), $k = 3$ (69.3 percent failed), $k = 5$ (66.4 percent failed), $k = 7$ (64.4 percent failed), $k = 9$ (62.6 percent failed), and $k = 11$ (61.2 percent failed).

As noted in the Methods section, as k increases, a higher probability exists that a reference observation will pass the inequality and require a full Euclidean distance comparison with the target. Our trials confirmed this relationship between increasing k and number of Euclidean distance measurements. Most importantly, our research shows that using the MPS algorithm can significantly improve mapping efficiency by reducing the number of calculations needed to classify each target pixel.

Figure 1.—*Overall classification accuracy of four forest type groups at increasing values of* k *for no stratification, upland stratum, lowland stratum, and combined upland/lowland strata.*



Database-Assisted Mapping

Time trials using our database-assisted mapping program were conducted on a Pentium 4, 2 GHz, Linux OS-based- computer with 1 GB of memory. Our *k*NN program was written in C++, using the MySQL++ API to interact with a local MySQL version 4.0.14 server. Our program was compiled with g++ (GCC) 3.2.2.

The program was tested on the 3-date, 14-feature image of St. Louis County, which contains 14.72 x 10^6 pixels. The kNN reference set contained 1,712 subplot observations. Across all bin counts, the average time required for our program to search

the database for a given feature pattern and return a null or imputed value was 372 milliseconds. If a null value was returned, the kNN algorithm was initiated and took an average of 3,947 milliseconds to run. When a kNN instance was complete, the program used an additional average of 196 milliseconds to insert the feature pattern and imputed record in the database.

For individual bin counts, the insert, search, and retrieval time was contingent on the number of records in the table. Figure 2 shows that the total interaction time with the database increases with the database size. After processing completed, the table held 10-bin = 10.6 x 10^6 records; 8-bin = 7.29 x 10^6 records; and 6-bin = 3.2 x 10^6 records.

Figure 3 shows the frequency at which imputed values were retrieved from the database as the image is processed. The fewer bins in the image, the greater redundancy there was in feature patterns. The greater the redundancy in feature patterns, the greater dependence on the database was required to provide imputed values. Figure 3 also shows a rough plateau starting at about 6.5 x 10^5 processed pixels. This leveling off point describes the percent redundancy in the image for the given bin count (e.g., approximately 80-percent redundancy in the 6-bin image).

The brute force kNN procedure took 16.1 hours to process the sample image. Incorporating the database with a bin count of 10 decreased mapping time by 23.3 percent. Reducing the bin count to 8 provided a 43.2 percent decrease in mapping time. Generalizing the image further to a bin count of 6 reduces mapping time by 67.1 percent.

This improvement in mapping efficiency must be balanced against accuracy loss from the discretization process. Figure 4 compares the forest type group classification overall accuracy of the binned images against the nonbinned image and shows that minor loss of information occurred in the 10- and 8-bin images. At the 6-bin image count of, accuracy declined more substantially.

The reduction in overall accuracy does not appear to be significant despite the severity of the image generalization. Deciding on the level of acceptable loss of classification accuracy in return for increased efficiency, however, is specific to the mapping project.

Our simple approach to imposing bin boundaries on each feature appears to maintain a significant portion of information; however, many other discretization procedures exist that may more effectively preserve class discriminating information contained in the images. Because of high spectral similarity, it is difficult to differentiate between forest types or forest type groups. Using our database-assisted mapping may enjoy experience success if classes were more spectrally distinct.

Figure 2.—*Total database interaction time versus number of pixels processed for image bin counts of 10, 8, and 6.*

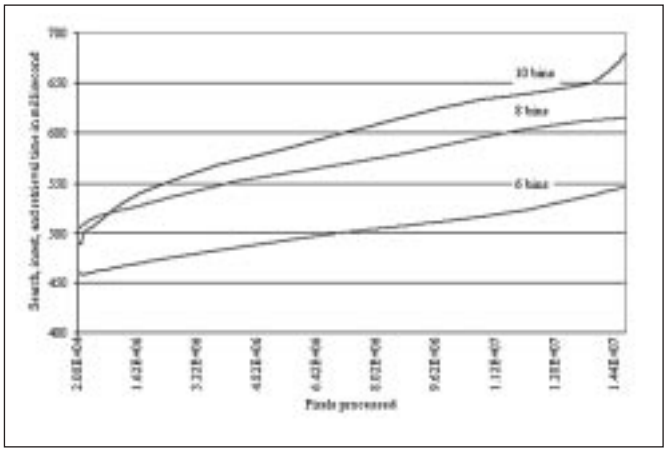


Figure 3.—*Percent of imputed values retrieved from the database versus number of pixels processed for image bin counts of 10, 8, and 6.*

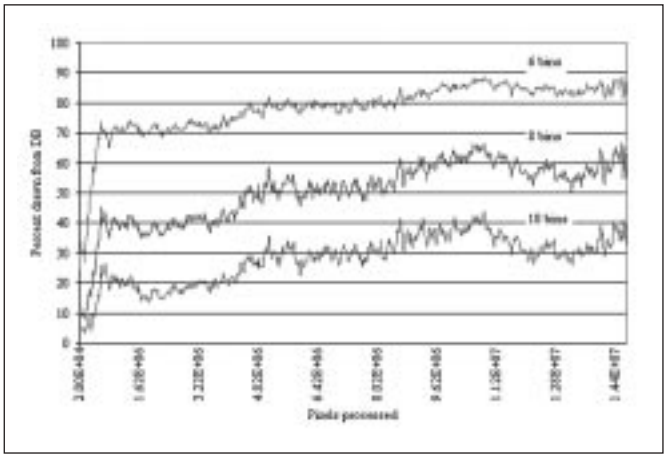
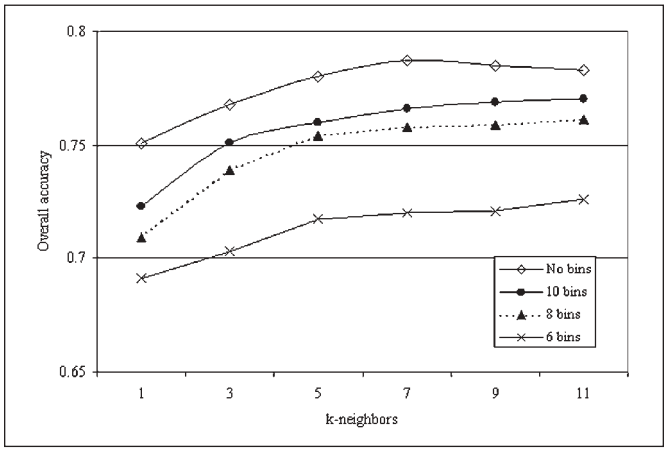


Figure 4.—*Overall classification accuracy of four forest type groups at increasing values of* k *for varying degrees of image generalization.*



Conclusions

Our study documented efforts to refine the process of kNN forest attributes classification using FIA plot data and Landsat 7 ETM+ imagery. We outlined three steps in the classification process that highlighted mapping efficiency improvements.

First, our analysis indicated that using transformed divergence may provide an objective way to reduce the dimensionality of the feature set without compromising classification accuracy. Second, the MPS algorithm proved to significantly reduce the number of distance calculations needed to classify each pixel. Third, the proposed database-assisted mapping provides a way to store and retrieve computationally expensive information.

Unlike the MPS algorithm, the discretization step needed for database-assisted mapping requires the analyst to compromise between increased mapping efficiency and loss of classification accuracy. Depending on the structure of the dataset and the degree of discretization, the loss of classification accuracy can be minimal.

Literature Cited

Bauer, M.E.; Burk, T.E.; Ek, A.R.; Coppin, P.R.; *et al.* 1994. Satellite inventory of Minnesota's forest resources. Photogrammetric Engineering and Remote Sensing. 60(3): 287–298.

Chmielewski, M.R.; Grzymala-Busse, J.W. 1996. Global discretization of continuous attributes as preprocessing for machine learning. International Journal of Approximate Reasoning. 15: 319–331.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment. 37: 35–46.

Efron, B.; Tibshirani, R.J. 1993. An introduction to the bootstrap. New York: Chapman & Hall. 436 p.

Franco-Lopez, H.; Ek, A.R.; Bauer, M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment. 77(3): 251–274.

Gong, G. 1986. Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. Journal of the American Statistical Association. 81: 108–113.

Haapanen, R.; Ek, A.R.; Bauer, M.E.; Finley, A.O. 2004. Delineation of forest/nonforest and other land-use/cover classes using k nearest neighbor classification. Remote Sensing of Environment. 89(3): 265–271.

McRoberts, R.E.; Nelson, M.D.; Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. Remote Sensing of Environment. 82: 2–3, 457–468.

Ra, S.W.; Kim, J.K. 1993. A fast mean-distance-ordered partial codebook search algorithm for image vector quantization. IEEE Transactions on Circuit and System, Part II: Analog and Digital Signal Processing. 40(9): 576–579.

Stehman, S.V. 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment. 62: 77–89.

Swain, P.H.; Davis, S.M., eds. 1978. Remote sensing: the quantitative approach. New York: McGraw-Hill. 396 p.

U.S. Department of Agriculture (USDA) Forest Service. 2000. Forest inventory and analysis national core field guide, volume 1: field data collection procedures for phase 2 plots, version 1.4. U.S. Department of Agriculture, Forest Service. Internal report. On file at U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis, Washington Office.