

PROGRESS IN ADAPTING k-NN METHODS FOR FOREST MAPPING AND ESTIMATION USING THE NEW ANNUAL FOREST INVENTORY AND ANALYSIS DATA

Reija Haapanen, Kimmo Lehtinen, Jukka Miettinen, Marvin E. Bauer, and Alan R. Ek ¹

ABSTRACT.—The k-nearest neighbor (k-NN) method has been undergoing development and testing for applications with USDA Forest Service Forest Inventory and Analysis (FIA) data in Minnesota since 1997. Research began using the 1987-1990 FIA inventory of the state, the then standard 10-point cluster plots, and Landsat TM imagery. In the past year, research has moved to examine potentials for improving cover type and volume mapping and estimation with the new annual FIA data, notably the new four-subplot cluster plot, and Landsat ETM+. Major findings to date point to the difficulty of choosing the number of neighbors (k). A value of k between 1 and 3 seems appropriate for mapping. A larger number of neighbors reduces the overall estimation error, but it also leads to a reduction in the producer's accuracy. Additionally, using multiple image dates for an area typically improves results considerably. Recent results with the new four-subplot cluster plot data show that stratification of the data into upland/lowland strata, use of thermal bands, and a plot location optimization all improve mapping and estimation results. Finally, segmentation algorithms show potential for improving mapping and the k-NN estimation process. A C-language program package for applying the k-NN method to forest inventory has also been developed.

BACKGROUND

A non-parametric method for estimation of forest variables has been used operationally in the Finnish National Forest Inventory since the early 1990s (Muinonen and Tokola 1990; Tokola and others 1996; Tomppo 1990, 1993). This method estimates forest variables by calculating a weighted average of measurements from a number of field sample plots. Weights are assigned according to the distance in spectral space, defined by the satellite image bands, between the pixel under classification and field sample plots. The k-nearest neighbor (k-NN) method can estimate multiple forest variables simultaneously and is a simple but powerful way to extend a wide range of field data to landscapes. The method can also preserve the covariance structure of forest variables and thus

produce maps that appear very realistic in terms of their spatial pattern.

Adapting the Finnish-developed k-NN technique to the FIA four-subplot cluster plot data and to the complex structure of Lake States forests has been a main study objective. The research poses significant challenges, but if these can be met, the combination of multispectral satellite imagery and existing field plot-based inventory data can be used to create detailed maps of volume, basal area, cover type, and annual change on an objective basis, at low cost, and with useful precision and a high degree of automation.

Additionally, the k-NN method can produce very local estimates with much improved precision. With k-NN, the information contained in field samples (FIA plots) is propagated across the entire population under the assumption that similar forest conditions exist across the satellite imagery and the measured spectral-radiometric-temporal responses of pixels in the imagery are dependent on the forest conditions.

¹ Research Fellow, Applications Programmer, Research Fellow, and Professors, respectively, Department of Forest Resources, College of Natural Resources, University of Minnesota, 1530 Cleveland Avenue N., Saint Paul, MN 55108. Phone: (612) 624-3400, fax: (612) 625-5212; e-mail: aek@forestry.umn.edu

PREVIOUS RESULTS WITH THE OLD FIA DESIGN

The adaptation of the k-NN method for the forests of Lake States began with the old FIA 10-point cluster plot design (see Franco-Lopez and others 2001). This research led to important findings on the behavior of the k-NN method in complex mixed species forests. The following list includes the most important results.

1. Larger numbers of neighbors (k) reduce the overall estimation error, but also reduce the producer's accuracy.
2. Using multiple (up to three) satellite images from different dates significantly improves results.
3. For the measure of distance between pixels in spectral space, the Euclidean distance proved superior to the Mahalanobis distance.
4. Filtering had little effect, e.g., pixel-based results were judged equivalent or better than using a 3 x 3 filter.
5. The best estimation accuracies at the pixel level with the old plot design were typically 60 to 85 percent root mean square error (RMSE) for volume and 45 to 55 percent correctly classified into the FIA cover type classification (14 forest cover types, USDA 2000).

STUDY AREA AND MATERIALS FOR THE RECENT STUDIES WITH THE NEW FIA DESIGN

Most studies have been conducted in northeastern Minnesota, specifically the FIA Aspen-Birch Survey Unit (unit 1). The area is approximately 29,748 km². However, the wall-to-wall extension of volume, basal area, and cover type has been developed for unit 1 and parts of unit 2 (Northern Pine Survey Unit).

Field Data

The new FIA four-subplot cluster plot (USDA 2000) data were used in the k-NN estimation. So far, however, only the first-year (1999) panel or set of plots has been available. The second-year panel (2000) will be added to the database when it arrives. The 1999 panel provided 180 forested plots in the study area, which led to 720 subplots, nearly all of which were forested. Mean volume for these plots was approximately 62 m³/ha (890 ft³/acre).

Image and Digital Map Data

The predominant imagery used was Landsat 7 ETM+ satellite images; when these were not available because of clouds, Landsat 5 TM images were used. Typically, two to four image dates were available for each scene used in the study. The satellite image database preprocessed for k-NN classification covered FIA unit 1 and parts of units 2 and 3 (Central Hardwoods Survey Unit) in Minnesota. The database included 13 satellite images: 8 Landsat 7 ETM+ images and 5 Landsat 5 TM images.

Digital map data are used to remove nonforest land use classes from the estimation and mapping process and to improve the accuracy of estimates (upland/lowland stratification). A mid-1990s Minnesota land use/land cover map with 30-m pixel size was used for this purpose (<http://lucy.lmic.state.mn.us>).

C-Program Package

A C-language program package for k-NN method estimation has been developed and tested. The package includes all the necessary tools for estimation of both categorical and continuous variables using Landsat data and the new FIA four-subplot cluster plot data. Estimation software includes a cross-validation program for error estimation (for both continuous and class variables), an optimization program to find the best weights for different channels in satellite images, and a mapmaker for implementing the actual k-NN classification. The programs also include features that allow geographic and other constraints in neighbor selection, band selection, optimal band weighting, and use of different mask files. These features are useful for both research and operational k-NN estimation. A manual for using the programs for k-NN-based estimation protocols has also been published (Haapanen and others, in prep. (c)), and it is now available on the Web at: <http://www.cnr.umn.edu/FR/publications/staffpapers/index.html>.

A set of programs for segmentation (polygon development) has also been developed. At the moment a preliminary program is ready (Haapanen and others, in prep. (c)). This software is modeled after approaches used in Finland. The segmentation may be used before or after the k-NN classification, e.g., to reduce the noise and make the stand borders clearer. There are some obvious application interests for forest cover type mapping, and we are communicating with cooperators on such potentials. However, the approach needs more exploration before it is clear just how it may aid k-NN applications.

METHODS

K-nearest Neighbor Method

For estimation with Euclidean distances, consider the spectral distance $d_{p_i,p}$ computed in the feature space from the pixel to be estimated p to each pixel p_i for which the ground measurement or class is known. For each pixel p , take the k -nearest field plot pixels (in the feature space) and denote the distances from the pixel p to the nearest field plot pixels by $d_{p_1,p}, \dots, d_{p_k,p}$ ($d_{p_1,p} \leq \dots \leq d_{p_k,p}$). The estimate of the variable value for the pixel p is then expressed as a function of the k closest units, with $k \sim 1-10$, and each such unit is weighted according to a distance function in the feature space. A commonly used function for weighting distances is

$$w_{(p_i)p} = \frac{1}{d_{(p_i)p}^t} \bigg/ \sum_{j=1}^k \frac{1}{d_{(p_j)p}^t} \quad (1)$$

with $t=2$.

The estimate of the variable m for pixel p is then

$$\hat{m}_p = \sum_{i=1}^k w_{(p_i)p} \cdot m_{(p_i)} , \quad (2)$$

where $m_{(p_i)}$, $i=1, \dots, k$, is the value of the variable m in sample plot i corresponding to the pixel $p_{(i)}$ which is the i th closest pixel (of “known” pixels) in the spectral space to the pixel p . Additionally, for estimation of class variables such as forest cover type, the modal cover type of the k nearest neighbors serves as the estimator. Readers may note that the overall method is a form of post-stratification.

Selection of Parameters

To obtain reliable estimates with the k -NN estimation method, it is important to have a rather large field plot sample. Theoretically speaking, one seeks to have all of the forest conditions represented in the sample. Without such representation, the estimation is limited in terms of effectively propagating all of the forest conditions across the subject area or population. For Nordic forest conditions, the understanding to date is that at least 500 sample plots are needed (Nilsson 1997), and we concur on this number for Minnesota. Beyond that, the most important parameters to be selected are the distance metric, the number of nearest neighbors k , parameters related to the digital elevation model, and stratification of the image data (Katila and Tomppo 2000).

Upland/Lowland Stratification

As has been found in Finland, the spectral response of lowland differs from that of uplands. In Minnesota this may be due in part to species concentrations (e.g., black spruce on lowlands) and possibly site factors. As a consequence, we employed upland/lowland stratification to improve the accuracy of estimates.

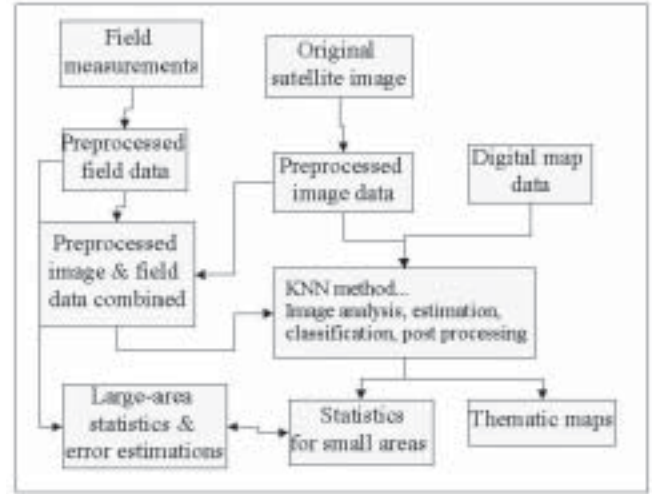


Figure 1.—Flowchart of the k -NN estimation and classification process.

The Location of Pixels

The pixel-level estimates are sensitive to field plot location and rectification errors. Errors can be mitigated to some extent by choice of the value of k . However, a high value of k shifts the estimates of the variables towards their mean values and reduces the variation. A quasi-optimization procedure for identifying the best match of pixel and field plot location has been developed, and preliminary results seem promising. The method is simplified to consider each plot and its nearby pixels, one plot at a time. This is in contrast to the approach developed by Halme and Tomppo 2001, which considers the entire inventory of plots and their locations simultaneously.

The quasi-optimization involved identifying a 3×3 pixel area around the center subplot location. The entire four-subplot cluster was then moved in the same direction, and the correlation between subplot volume and selected spectral bands was calculated. In effect, we sought the location that maximized the above correlation on the plot over the four

subplots. Also, when we calculated the correlations for a four-subplot cluster for a given test location, we moved all the subplots in tandem, i.e., given the center subplot location, and assumed the other subplots to be in their correct location relative to each other. Image dates were treated separately and allowed different optimal locations.

Finally, using the software noted above, we have been examining the potentials of segmentation approaches in conjunction with k-NN methods. The last part of the paper describes this work and plans.

RESULTS AND DISCUSSION

The FIA Four-Subplot Cluster

The four-subplot cluster plot has some interesting features when used in k-NN classification (fig. 2). Since the subplots are very close to each other, the spectral values for pixel values tend to be very similar within a cluster. In cover type classification, the entire four-subplot cluster was often of the same cover type and thus the nearest neighbors were usually found within the cluster. This result can clearly impact inventory-wide calculations and interpretation of overall accuracies in classifications.

Upland/Lowland Stratification

The upland/lowland division was made using the Minnesota land use/land cover map. As anticipated, the stratification improved the cover type accuracies for both strata (fig. 3).

Multitemporal Images

Using more than three image dates did not seem to improve the results (fig. 4). In this case the three images were March, April, and May. The fourth image added (a July date) did not improve results.

Larger Study Area Versus Multiple Dates

A comparison was made between two, three, and four image dates. The two image date data set included the study area and a portion of the adjacent Landsat scene south of the primary study area. The three and four image date data sets did not extend beyond the primary study area. Results indicated only the oak, elm-ash-cottonwood, and maple-basswood cover types benefited from an increase in number of field plots as judged by the producer's accuracy. The increase in the number of field plots with two image dates was over 45 percent for these three

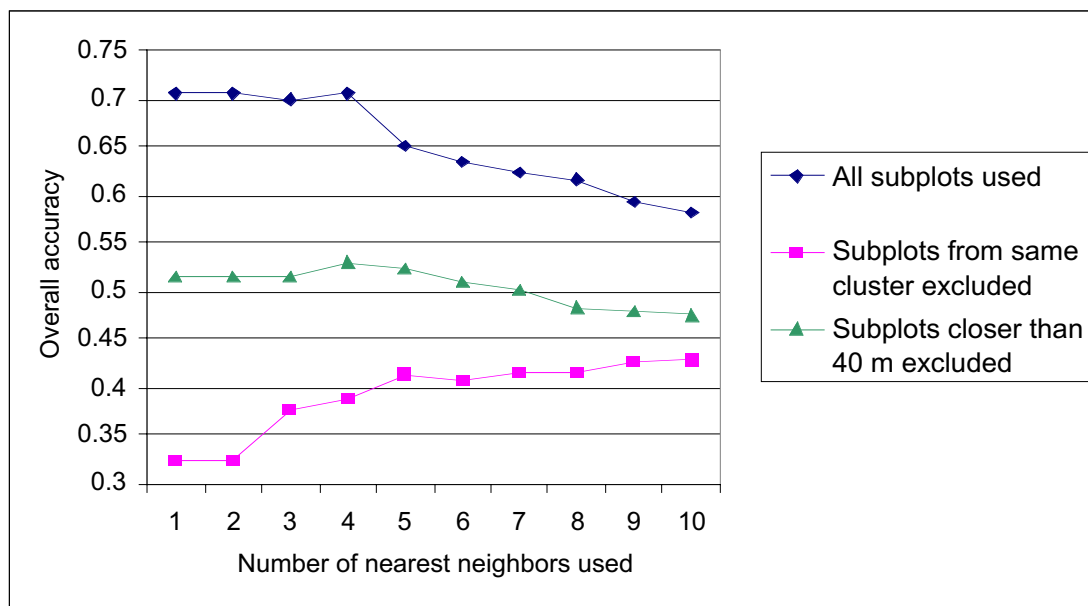


Figure 2.—Effect on overall accuracy of forest cover type estimation from limiting the inclusion of subplots.

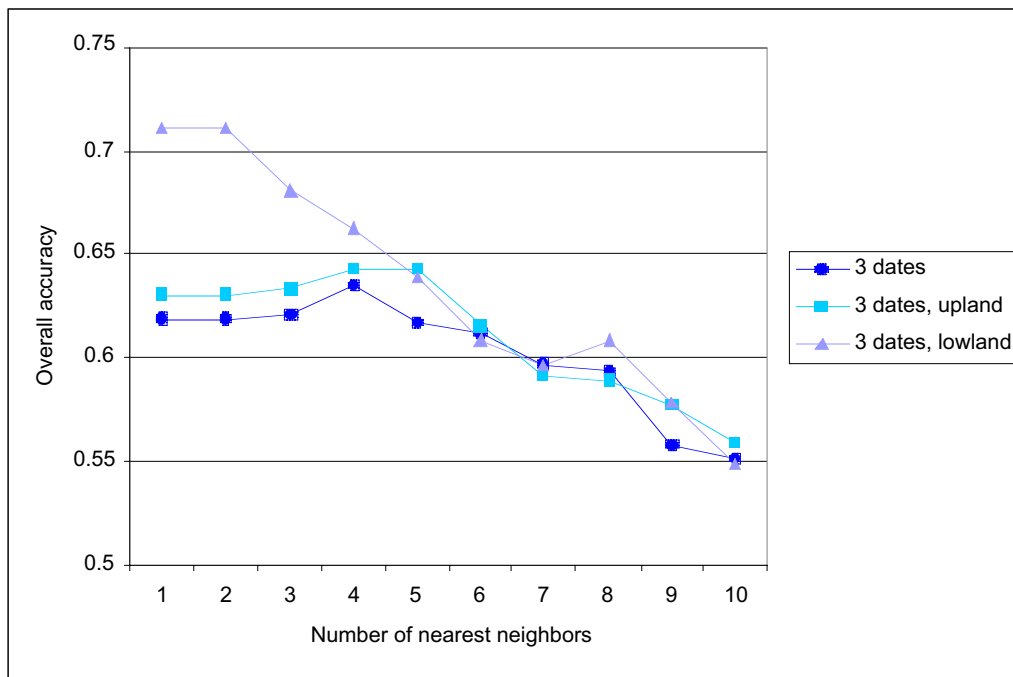


Figure 3.—Effect on overall accuracy of forest cover type estimation with stratification of field data into upland and lowland classes, northeastern Minnesota, 1999 FIA data, $n = 717$ subplots.

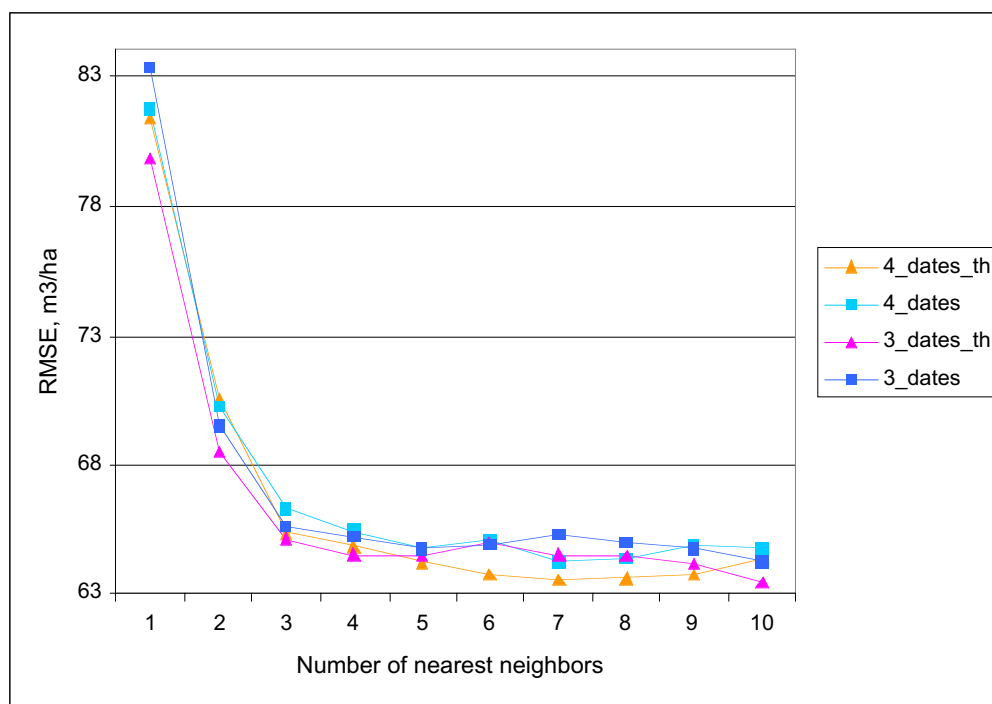


Figure 4.—RMSE for volume estimation for three and four dates of imagery (March, April, May, and July) with thermal bands, $n = 685$ subplots.

classes, and the accuracies changed as follows: oak 44 to 64 percent, elm-ash-cottonwood 47 to 52 percent, and maple-basswood 56 to 60 percent. However, the accuracies of most of the other 13 classes fell, indicating that an additional optimal image date (in this case the March image) was more important than the additional plot data available when only two dates were considered (fig. 5).

bands improved results. The benefits for cover type classification were also evident when subplots from the same cluster (those within 70 m) as neighbors were prohibited (overall accuracy 29 percent without and 33 percent with thermal bands). In volume estimation, the results with use of thermal bands improved only for $k = 1$.

Weighting of Spectral Bands

The weighting of spectral bands improved the cover type classification accuracy from 70 to 82 percent for $k = 1$, when all subplots were included in the analysis. The volume RMSEs dropped from 80.7 to 69.4 m^3/ha with similar parameters. However, the determination of weights was not straightforward. Results for different cycles (optimizations based on different—arbitrarily chosen—starting points) show local optima, thus the approach is not without some problems.

Thermal Bands

Thermal bands were found to add important information to the classifications for forest cover type classification and to a lesser degree those for volume (fig. 6). When all subplots were included in the error estimation process, the thermal

Forest/Nonforest Classification

The viability of using k -NN for FIA specified forest/nonforest classification has been tested, with preliminary results showing 86 percent accuracy. A paper describing these results is in preparation.

Plot Location Optimization

Plot location optimization was tested for both volume and cover type. The same plot locations, obtained by studying the correlation between volume and selected image bands, were used for estimating both variables. The cover type accuracies did not benefit from the location optimization, but for volume, the results were slightly better than the band optimization results described earlier. When the data were stratified into upland and lowland, the plot location optimization gave the best results for upland, but no optimization performed best on lowland. However, the differences in results for lowland were

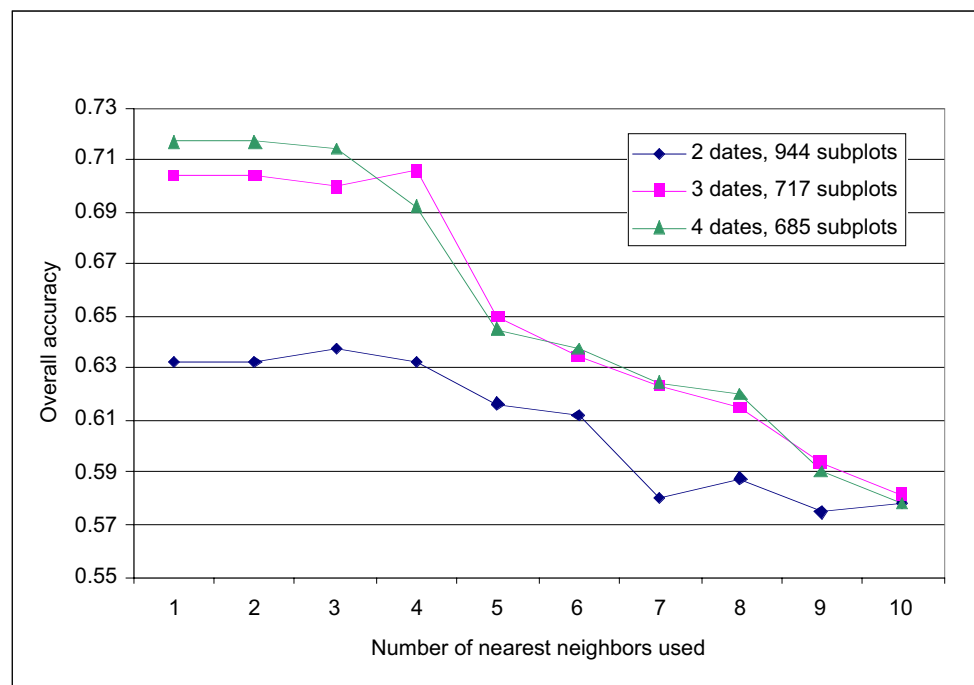


Figure 5.—Comparison of overall accuracies for forest cover type classification when two (April and May), three (March added), or four (July added) image dates are used; the number of subplots is 944, 717, and 685, respectively. Thermal bands included except for the July image (Landsat 5).

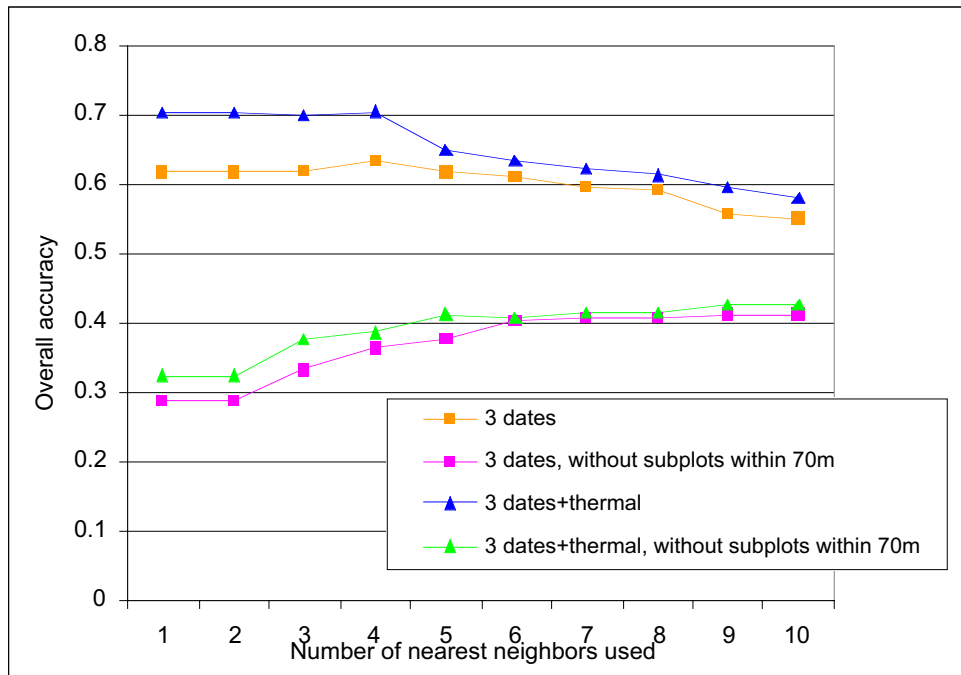


Figure 6.—Accuracy of cover type classification with and without thermal bands, 717 subplots.

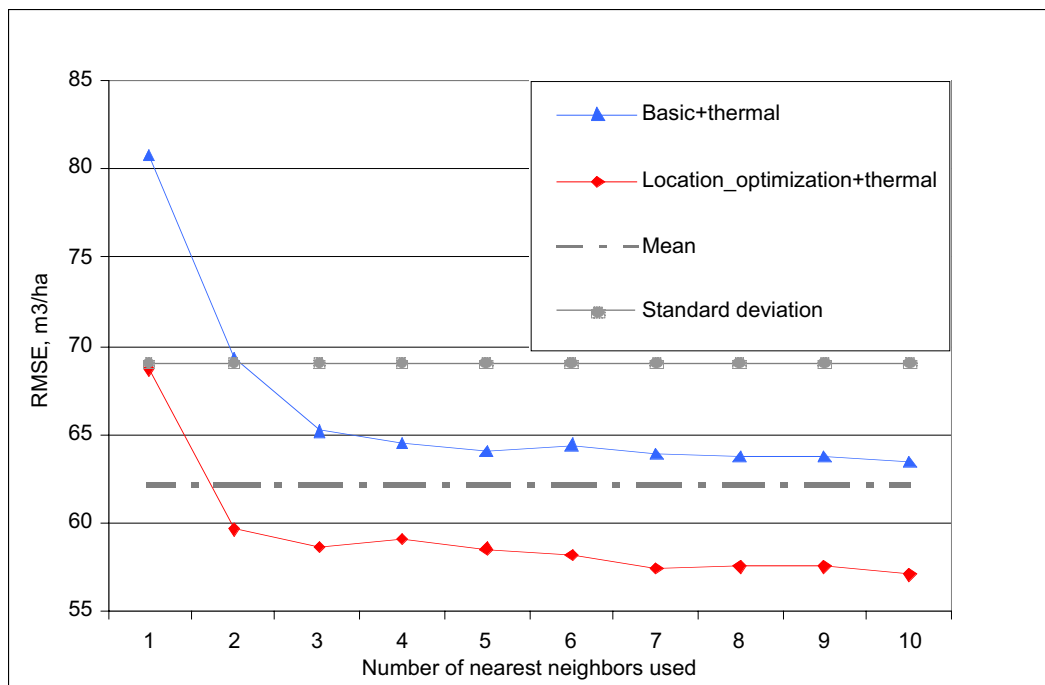


Figure 7.—Effect of plot location optimization on volume estimation RMSEs, 717 subplots.

small. Figure 7 describes the effect of plot location optimization on volume estimation RMSE. Note that plot location optimization was always superior to no optimization.

k-NN and Segmentation Approaches

Using segmentation with the k-NN method has not received much attention in research. In theory there are two ways to use segmentation with k-NN. First, one could conduct a k-NN assignment of field plot data and then segment the results and label the resulting polygons by average or modal k-NN values. This could be an effective way to produce preliminary stand delineations and cover type maps. Alternatively, one could segment the image and then conduct a modified k-NN estimation on a polygon or other basis to assign field plot data.

In our studies, two different segmentation algorithms that have been successfully used in Scandinavian forests (Haapanen and Pekkarinen 2000, Hagner 1990) have been implemented (in software) and are undergoing testing:

- (1) T-ratio method by Hagner (1990)
- (2) Narendra and Goldberg method (1980)

These two algorithms can also be used one after another to improve results: the segmentation can be run with Narendra and Goldberg's method and then fine tuned with the T-ratio method. Our studies have just started and there are few results to present at this point, except that realistic appearing stand- level segmentation now seems feasible for section, township, and perhaps larger areas with Landsat TM level resolution.

NEXT STEPS

Plans for the near future include extending k-NN methods to incorporate a second year of FIA data, which will allow a larger sample size on both a plot and subplot basis. This addition will facilitate determination of the best approaches (plot or subplot basis) to improve confusion matrices and to investigate segmentation approaches in conjunction with k-NN methods. Additionally, we plan to examine these methods for possible use in improving existing FIA design/ estimation and potential design improvements.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support for this work from the University of Minnesota's College of Natural Resources and the Minnesota Agricultural Experiment Station, the Minnesota Department of Natural Resources, NASA, the USDA Forest Service North Central Research Station, and NCASI. The technical assistance of Dr. Ronald E. McRoberts and Dr. Mark H. Hansen of the North Central Research Station is especially appreciated. We also extend our thanks to Dr. McRoberts and Christopher Edgar, Department of Forest Resources, University of Minnesota, for their very constructive review comments on this paper.

LITERATURE CITED AND REFERENCES

- 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.
- Haapanen, R.; Ek, A.R. 2001. Software and instructions for kNN applications in forest resources description and estimation. Staff Pap. 15. St. Paul, MN: Department of Forest Resources, College of Natural Resources, University of Minnesota. (see publications list at www.cnr.umn.edu/FR)
- Haapanen, R.; Pekkarinen, A. 2000. Utilising satellite imagery and digital detection of clear cuttings for timber supply management. In: *International Archives of Photogrammetry and Remote Sensing*. 33(B7): 481-488.
- Haapanen, R.; Ek A.R.; Bauer, M.E. In prep. (a) K-nearest neighbors based estimation and mapping of forest stand volume and cover type using the new FIA plot design. To be submitted to *Remote Sensing of Environment*. 43 p. Draft.
- Haapanen, R.; Ek A.R.; Bauer, M.E. In prep. (b) Delineation of forest/non-forest and other land-use/cover classes using k nearest neighbor classification. To be submitted to *Remote Sensing of Environment*. 14 p. Draft.

- Haapanen, R.; Lehtinen, K.; Ek, A.R. In prep. (c). Software and instructions for segmentation applications in forest resources description and estimation. Staff Pap. St. Paul, MN: Department of Forest Resources, College of Natural Resources, University of Minnesota. 15 p. Draft.
- Hagner, O. 1990. Computer aided forest stand delineation and inventory based on satellite remote sensing. In: The usability of remote sensing for forest inventory planning. Proceedings from SNS/IUFRO workshop; 1990 February 26-28; Umeå, Sweden. Umeå, Sweden: Remote Sensing Laboratory, Swedish University of Agricultural Sciences: 94-105.
- Halme, M.; Tomppo, E. 2001. Improving the accuracy of multi-source forest inventory estimates – a multi-criteria approach. *Remote Sensing of Environment*. 77: 1-7.
- Katila, M.; Tomppo, E. 2000. Selecting estimation parameters for the Finnish multisource National Forest Inventory. *Remote Sensing of Environment*. 76: 1-17.
- Muinsonen, E.; Tokola, T. 1990. An application of remote sensing for communal forest inventory. In: The usability of remote sensing for forest inventory and planning. In: Proceedings from SNS/IUFRO workshop. Rep. 4. Umeå, Sweden: Remote Sensing Laboratory, Swedish University of Agricultural Sciences: 35-42.
- Narendra, P.; Goldberg, M. 1980. Image segmentation with directed trees. *IEEE Transactions on pattern analysis and machine intelligence*. Pami-2: 185-191.
- Nilsson, M. 1997. Estimation of forest variables using satellite image data and airborne lidar. *Acta Universitatis Agriculturae Suecicae. Silvestria* 17. Ph.D. Dissertation.
- Tokola, T.; Pitkänen, J.; Partinen, S.; Muinsonen, E. 1996. Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. *International Journal of Remote Sensing*. 17(12): 2333-2351.
- Tomppo, E. 1990. Designing a satellite image aided National Forest Survey in Finland. In: The usability of remote sensing for forest inventory planning. Proceedings from SNS/IUFRO workshop; 1990 February 26-28; Umeå, Sweden. Umeå, Sweden: Remote Sensing Laboratory. Swedish University of Agricultural Sciences: 43-46.
- Tomppo, E. 1993. Multi-source national forest inventory of Finland. In: Nyyssönen, A.; Poso, S.; Rautala, J., eds. Proceedings of Ilvessalo symposium on national forest inventories. Res. Pap. 444. Helsinki, Finland: The Finnish Forest Research Institute: 53-61.
- USDA Forest Service. 2000. Forest inventory and analysis national core field guide. Vol. I: Field data collection procedures for phase 2 plots. Version 1.4.