K-NEAREST NEIGHBOR IMPUTATION OF FOREST INVENTORY VARIABLES IN NEW HAMPSHIRE

Andrew Lister and Mike Hoppus
Research Foresters
USDA Forest Service, Forest Inventory and Analysis
Northeastern Research Station
11 Campus Blvd, Ste 200
Newtown Square, PA 19073
alister@fs.fed.us

Raymond L. Czaplewski
Mathematical Statistician
USDA Forest Service, Natural Resource Assessment,
Ecology, and Research
Rocky Mountain Research Station
2150 Centre Ave., Bldg. A
Fort Collins, CO 80526
rczaplewski@fs.fed.us

ABSTRACT
The *k*-nearest neighbor (*k*NN) method was used to map stand volume for a mosaic of 4 Landsat scenes covering the state of New Hampshire. Data for gross cubic foot volume and trees per acre were summarized from USDA Forest Service Forest Inventory and Analysis (FIA) plots and used as training for *k*NN. Six bands of Landsat satellite imagery and various topographic measures were assessed for their use as variables defining the dimensions of the *n*-dimensional prediction space. Results were generally poor due to the weak correlations between the independent and dependent variables. We discuss the technique and results, and suggest avenues for future research.

INTRODUCTION

Nearest neighbor imputation is a nonparametric estimation procedure. The technique works by measuring *n* predictor variables and one or more dependent variables on a set of observation units. The values of the dependent variables are assigned to locations in an *n*-dimensional prediction space, the dimensions of which are formed by the ranges of values of the *n* predictor layers. Values of the dependent variables then can be imputed to other locations in the prediction space by creating and summarizing distance matrices. For example, the nearest observed point to an unobserved point in prediction space is the first nearest neighbor, and its value can be imputed to that unknown point. The average (if continuous) or mode (if categorical) value of the first *k* neighbors also can be imputed to an unknown point. Distance metrics vary, but include Euclidean, Mahalanobis, and others (Hineburg et al., 2000).

Nearest neighbor imputation methods have been used to produce continuous maps of forest attributes with satellite and GIS data layers to create prediction space-defining dimensions (Franco-Lopez et al., 2001; McRoberts et al., 2002; Tomppo, 2002). Forest inventory plots, each of which represents about 2428 ha and contains detailed tree and stand-level information, are geo-referenced and combined with satellite images and other maps to create a set of training data, which are then imputed to unknown locations. The U.S. Department of Agriculture's Forest Service's Forest Inventory and Analysis Program (FIA) is interested in using satellite and GIS data to produce continuous maps that can be summarized for small areas. The goal of the current project is to assess the *k*-nearest neighbor method's ability to produce continuous maps of forest attributes.

METHODS

The study area was New Hampshire (Fig. 1). Total cubic foot gross volume (CFGVOL) and trees per acre (TPA) were summarized from 642 completely forested or completely nonforested inventory plots found in the FIA

plot database. FIA plots consist of an array of four 7.3-m radius subplots (Fig. 1). The predictor layers evaluated include six bands of spring and six bands of fall Landsat ETM satellite imagery collected between 1999 and 2001 and several digital elevation model-based measures including elevation, slope, transformed aspect (Roberts and Cooper, 1989), topographic (Gessler et al., 1995; Moore et al., 1993) and landform shape indices (McNab, 1989; McNab, 1993), and slope position. A subset of these predictor layers was chosen based on strongest correlation with the dependent variables (spring and fall bands 1, 2, 3 and 7) using correlation and scatterplot analysis. For each FIA plot, the associated set of predictor data was created by averaging values of pixels intersecting a 35m buffer around the center subplot using ERDAS Imagine software*. Distances between known plots and unknown pixel locations were calculated using a simple Euclidean distance measure:

$$D_{ab} = \left[\sum_{i=1}^{n} (a_i - b_i)^2 \right]^{0.5}$$

where D is distance, a and b are the values of predictor variables at pixels being evaluated, and n is the number of predictor variables used.

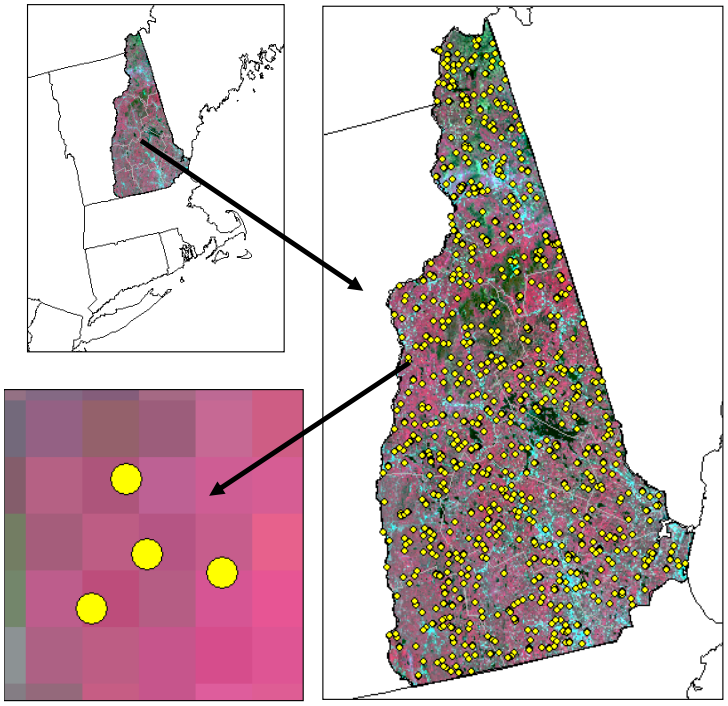

Figure 1. Map of the study area. Each of the 642 FIA plots is comprised of an array of four 7.3-meter-radius subplots – three of the subplots are located 36.6 meters from a center subplot and at 120 degrees from each other.

To objectively choose a number of nearest neighbors (*k*) to average for imputation to a location with an unknown value for volume and TPA, we graphed the relationship between number of neighbors and two metrics we use to quantify the accuracy of a prediction – mean absolute error (MAE) and squared Pearson's correlation coefficient (R^2). MAE is useful because it indicates the magnitude of the error relative to the range of data values; R^2 is useful because it is an index of the linear association of the actual and predicted values. We then subjectively chose a value for *k* beyond which little increase in accuracy was achieved with the addition of another neighbor. We created a map with this level of *k*, and used 70 FIA plots for accuracy assessment.

* The use of trade, firm or corporation names in this publication is for the information of the reader. Such use does not constitute an official endorsement or approval by the USDA or the Forest Service of any product or service to the exclusion of others that may be suitable.

RESULTS AND DISCUSSION

The correlation statistics and the resulting subset of layers chosen for the analysis are shown in Table 1. We chose to ignore collinearity in the predictor layers because the method is nonparametric and does not rely on tests of accuracy statistics against theoretical distributions, as is the case with parametric methods such as linear regression. Furthermore, we could not find mention of adjustments made for collinearity in other literature about the use of *k*NN in forestry.

Table 1. Correlation coefficients (*r*) and p values (p) for CFGVOL and TPA. Variables in bold were included in the final analysis.

	CFGVOL		TPA	
	r	p	*r*	p
Transformed aspect	0.02	0.62	0.05	0.19
Aspect	0.01	0.81	0.04	0.28
CTI	-0.17	<0.01	-0.21	<0.01
Easting	-0.05	0.14	-0.13	<0.01
Elevation	0.05	0.16	0.30	<0.01
Fall 1	-0.36	<0.01	-0.44	<0.01
Fall 2	-0.34	<0.01	-0.43	<0.01
Fall 3	-0.35	<0.01	-0.41	<0.01
Fall 4	0.06	0.13	-0.10	0.00
Fall 5	-0.20	<0.01	-0.36	<0.01
Fall 7	-0.31	<0.01	-0.42	<0.01
Northing	0.07	0.04	-0.07	0.06
Shape	-0.04	0.27	-0.07	0.04
Slope	0.14	<0.01	0.25	<0.01
Slope position	-0.02	0.63	0.06	0.08
Spring 1	-0.32	<0.01	-0.38	<0.01
Spring 2	-0.29	<0.01	-0.40	<0.01
Spring 3	-0.34	<0.01	-0.36	<0.01
Spring 4	0.06	0.09	-0.14	<0.01
Spring 5	-0.20	<0.01	-0.35	<0.01
Spring 7	-0.29	<0.01	-0.37	<0.01

Figures 2a-b show the relationships of MAE and R^2 vs. *k* for CFGVOL and Figures 2c-d show those for TPA. Both CFGVOL and TPA tend to show highest accuracy with respect to R^2 with an approximate value for *k* of 20. For MAE, the *k* value above which large reductions occur is slightly lower. The choice of the level of *k* to choose in a *k*NN analysis involves a tradeoff. Because both the plot data and the predictor layers have random error, averaging *k* predictions at each location tends to compress the variance of the final predictions. For example, in a situation where the average of the first 10 neighbors will be used to make the final map, large values for *k*=1-3 will not lead to a high estimate if *k*=4-10 are all values close to the mean. Generally, the impact of values of the dependent variable found in the tails of the distribution of training data tends to be diluted, especially when the distribution of training data approximates a normal distribution, as it does in the case of TPA and CFGVOL (Fig. 3). Figure 4, which depicts scatterplots of actual vs. predicted values for *k*=15, demonstrates this. Low true values tend to be overestimated and high true values are underestimated, resulting in a situation where values close to the mean of the distribution of training data are better predicted than those in the tails. This same phenomenon is described in a regression context in Cohen et al. (2003). In *k*NN, a tradeoff is therefore required: if the goal is to more accurately reproduce the variance of the training data in the set of estimates, then a smaller value of *k* would be desirable, but if overall accuracy is of interest, than a larger value is useful (Franco-Lopez, 2001). We chose a value of *k*=15 subjectively after inspecting Figure 2 and considering the disadvantages of choosing larger values for *k*.

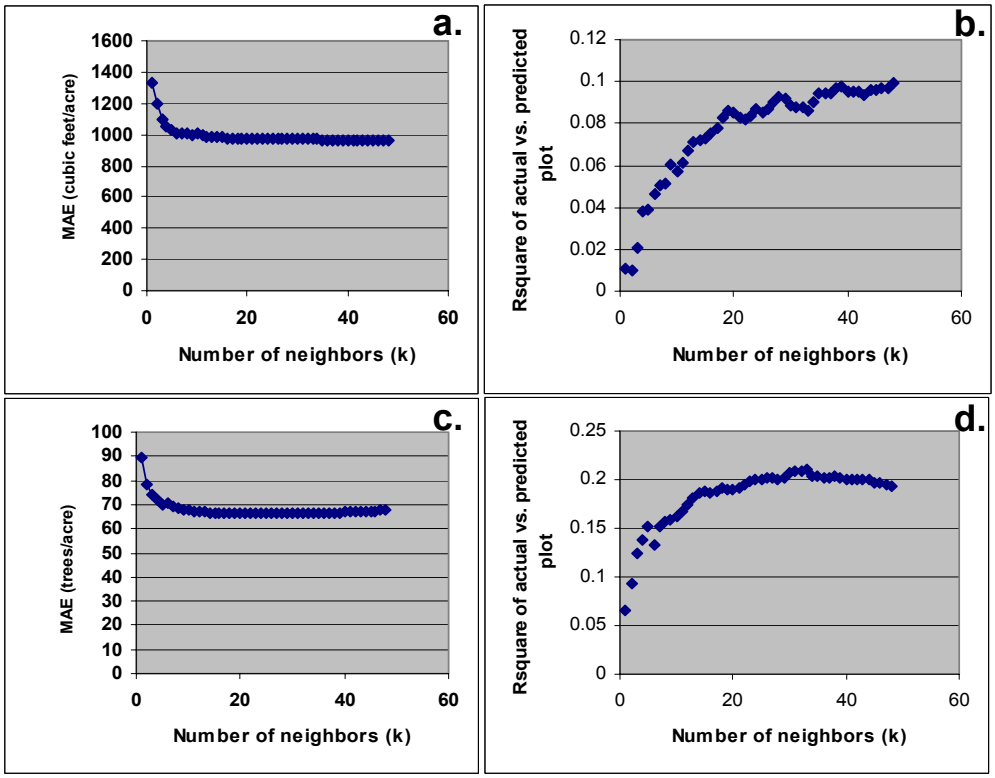

Figure 2. MAE (a) and R^2 (b) for CFGVOL vs. *k*; and MAE (c) and R^2 (d) for TPA vs. *k* for 642 plots at which predictions were generated.

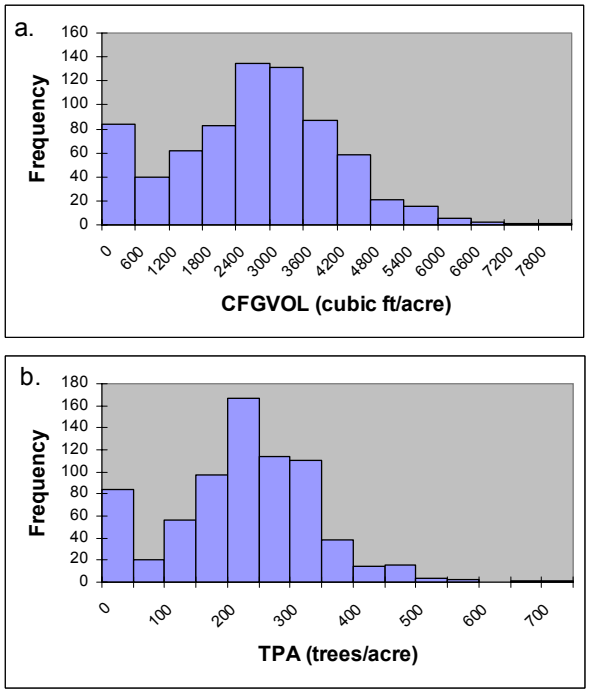

Figure 3. Frequency histograms of the training data for CFGVOL (a) and TPA (b).

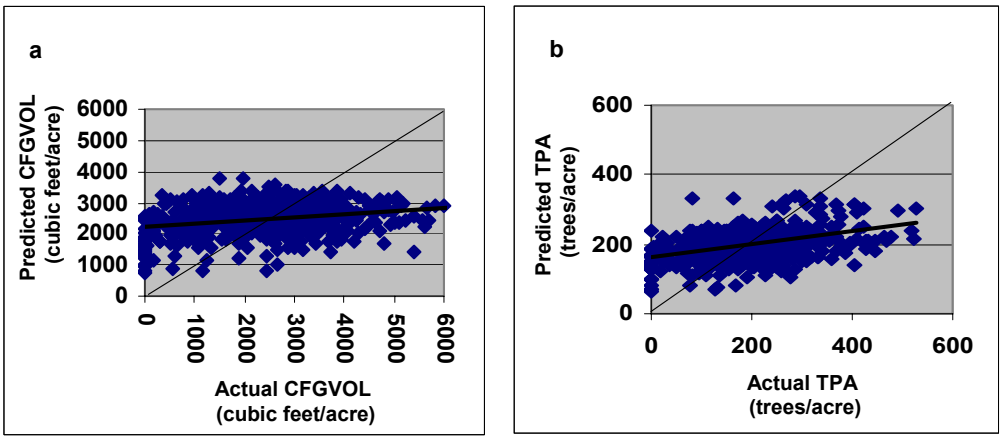

Figure 4. Scatterplots of actual vs. predicted values for CFGVOL (a) and TPA (b) for predictions at 642 plot locations. Bold line is best fit line through cloud of points (observed agreement), and thin line is the 1:1 line (perfect agreement).

The final maps of CFGVOL and TPA appear random at a coarse scale, but show patterns across the landscape at finer scales (Fig. 5). The accuracy statistics for the modeling and validation data are shown in Table 2. The relative error is the ratio of the observed MAE to the error obtained by averaging absolute deviations of each training datum from the mean of the set of training data. A relative error of 1 (or above) indicates that the model performed the same (or worse) than simply using the mean of the training data as each prediction. Smaller relative errors indicate the model performed better than simply applying the mean to each prediction.

The poor performance of the *k*NN technique (Table 2) is probably due to the weak relationship between the satellite data and any of the predictor data (Table 1). Correlation coefficients are generally low, suggesting that the locations in eight dimensional space defined by the values of the eight chosen predictor layers are not related to similarity of plots with respect to CFGVOL or TPA. The Landsat sensor responds to reflected light and many land-cover configurations can produce the same signal at the sensor. Also, TPA and CFGVOL are to some extent below-canopy phenomena; many different combinations of size class and tree density can yield the same amount of vegetation cover and satellite sensor response. The best satellite-based prediction methods typically use variables like canopy cover or leaf area index as dependent variables (e.g. Cohen et al., 2003). Also, using more complex distance measures and *k*-nearest neighbor algorithms has been shown to yield better accuracies (Ohmann and Gregory, 2002).

To improve our results, future work will use different predictor variables, weight certain plots based on their degree of relatedness to the dependent variable of interest, use different distance metrics, and calculate weighted averages of the training data contributing to a prediction, with weights arising from the relative magnitudes of the *k*-dimensional distances from each training datum to that zprediction. Also, other dependent variables, such as stocking or percent canopy cover, probably would be better correlated with our set of predictor layers.

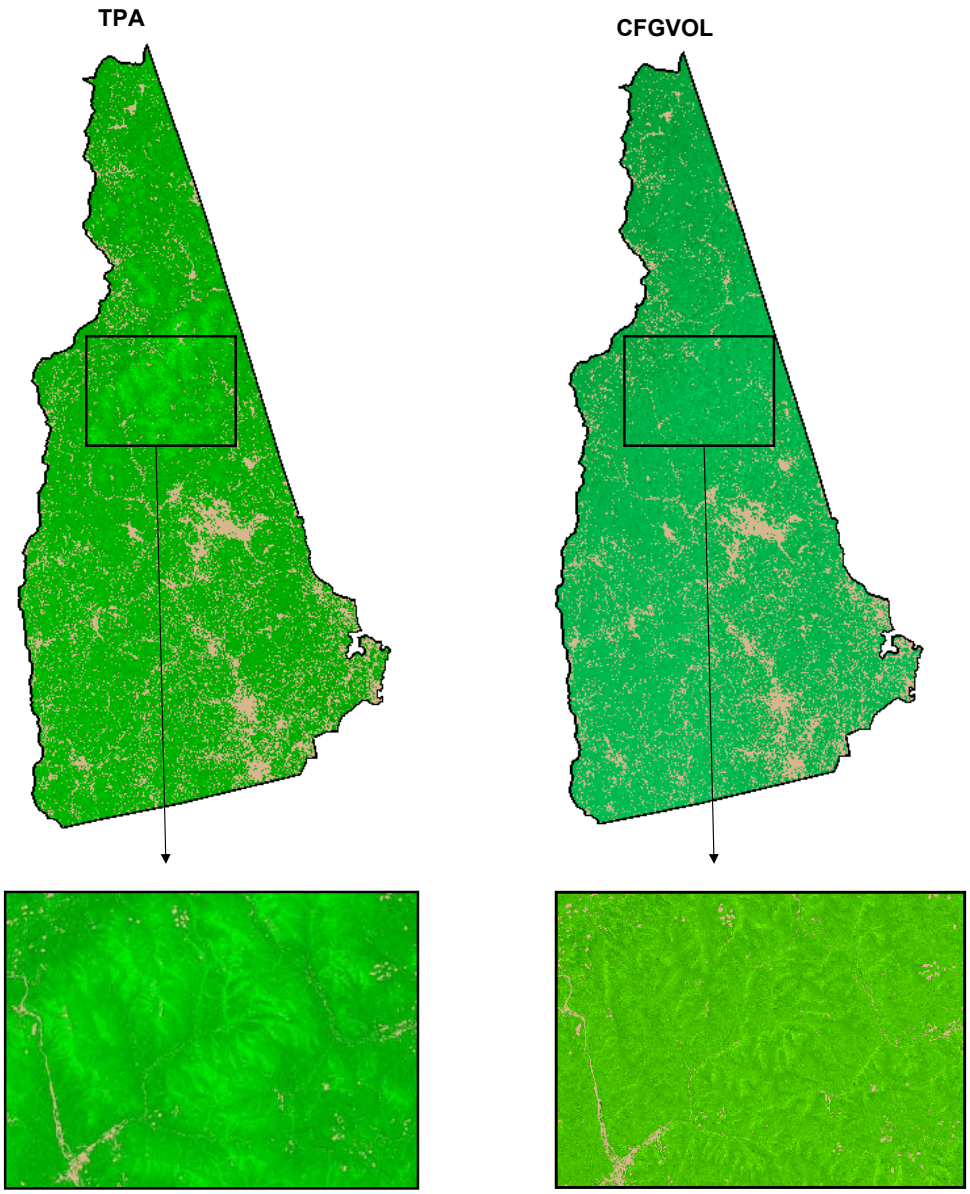

Figure 5. Final *k*-nearest neighbor maps for TPA and CFGVOL for New Hampshire (k=15). The area within the box is a subset of the map at a larger scale to reveal finer scale patterns in the predictions.

Table 2. Accuracy results of the model and validation for both TPA and CFGVOL. Relative error is the ratio of the observed MAE to the average absolute deviation of each training datum from the mean of the distribution of training data.

	model		validation	
	TPA	CFGVOL	TPA	CFGVOL
R^2	0.17	0.07	0.12	0.01
MAE	66.11	979.47	80.50	1036.50
Relative error	0.91	0.99	0.96	1.07

LITERATURE CITED

Cohen, W.B., Maiersperger, T.K., Gower, S.T., and D.P. Turner. (2003). An improved strategy for regression of biophysical variables and Landsat ETM+ data. *Remote Sensing of Environment*, 84: 561-571.

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

Gessler, P.E., Moore, I.D., McKenzie, N.J., and P.J. Ryan. (1995). Soil-landscape modeling and spatial prediction of soil attributes. *International Journal of GIS*, 9(4): 421-432.

Hineburg, A., Aggarwal, C.C., and D.A. Keim. (2000). What is the nearest neighbor in high dimensional spaces? In: Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, Kyu-Young Whang (eds.). VLDB 2000, *Proceedings of 26th international conference on very large data bases*, September 10-14, 2000, Cairo, Egypt, pp. 506-515.

McNab, H.W. (1989). Terrain shape index: quantifying effect of minor landforms on tree height. *Forest Science*, 35(1): 91 -104.

McNab, H.W. (1993). A topographic index to quantify the effect of mesoscale landform on site productivity. *Canadian Journal of Forest Research*, 23: 1100-1107.

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

Moore, ID., Gessler, P.E., Nielsen, G.A., and G.A. Petersen. (1993). Terrain attributes: estimation methods and scale effects. In: A.J. Jakeman M.B. Beck and M. McAleer (eds.). *Modeling change in environmental systems*. Wiley, London. pp. 189 - 214.

Ohmann, J.L. and M.J. Gregory. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, U.S.A. *Canadian Journal of Forest Research*, 32: 725-741.

Roberts, D.W. and S.V. Cooper. (1989). Concepts and techniques of vegetation mapping. In: *Land classifications based on vegetation: applications for resource management*, Gen. Tech. Rep. INT-257. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ogden, UT, pp. 90-96.

Tomppo, E. (1991). Satellite imagery-based national inventory of Finland. *International Archives of Photogrammetry and Remote Sensing*, 28: 419-424.