Remote Sensing of Environment 113 (2009) 489–499

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Diagnostic tools for nearest neighbors techniques when used with satellite imagery

Ronald E. McRoberts

Northern Research Station, U.S. Forest Service, St. Paul, MN, USA

ARTICLE INFO

Article history:
Received 7 February 2008
Received in revised form 6 May 2008
Accepted 5 June 2008

Keywords:
Bias
Homoscedasticity
Outliers
Influential observations
Extrapolations
Forest inventory

ABSTRACT

Nearest neighbors techniques are non-parametric approaches to multivariate prediction that are useful for predicting both continuous and categorical forest attribute variables. Although some assumptions underlying nearest neighbor techniques are common to other prediction techniques such as regression, other assumptions are unique to nearest neighbor techniques. Graphical diagnostic tools are proposed to evaluate the assumptions and to address issues of bias, homoscedasticity, influential observations, outliers, and extrapolations. The tools are illustrated using results obtained from applying the k-Nearest Neighbors technique with Landsat imagery and forest inventory ground observations.

Published by Elsevier Inc.

1. Introduction

Nearest neighbors techniques are intuitive, non-parametric approaches to either univariate or multivariate prediction based on the similarity in a covariate space between the unit for which a prediction is desired and units for which observations are available. Nearest neighbors techniques have gained popularity in forestry applications for imputing values for missing observations in databases (LeMay & Temesgen, 2005; Moeur & Stage, 1995; Temesgen et al., 2002), mapping forest attributes using the spectral values of satellite imagery (Franco-Lopez et al., 2001; McRoberts et al., 2002, 2007; Trotter et al., 1997), increasing the precision of probability-based forest inventory estimates (Baffetta et al., 2009-this issue; Katila & Tomppo, 2005; McRoberts et al., 2002; Nilsson et al., 2005; Tomppo, 1991, 1996), and model-based inference (Magnussen et al., 2009-this issue; McRoberts et al., 2007).

Despite the popularity and utility of nearest neighbors techniques, diagnostic tools for evaluating the assumptions underlying these techniques and the characteristics of their predictions have not been widely reported or commonly used. Because predictions of continuous variables obtained using nearest neighbors techniques share features with predictions obtained using regression techniques, an investigation of regression diagnostics merits consideration. Atkinson (1986) defined regression diagnostics as "the name given to a collection of techniques for detecting disagreement between a regression model and the data to which it is fitted." Brownlee (1965), Draper and Smith (1966), and Box et al. (1978) all discussed diagnostic techniques using graphical approaches facilitated by the use of computers and computer software packages. Belsey et al. (1980) and Cook and Weisberg (1982) were among the first to attempt comprehensive approaches to diagnostic analyses. Both textbooks emphasized the analysis of residuals and the use of graphical methods, and both addressed detection of influential observations, outliers, and departures from model assumptions such as linearity and homoscedasticity (homogeneity of residual variance). In a review of these two texts, Atkinson (1984) provided a background and history of diagnostics techniques. Cook and Weisberg (1983) followed their textbook with a discussion of diagnostic techniques for assessing homoscedasticity, and Chatterjee and Yilmaz (1992) reviewed diagnostic techniques for identifying and evaluating influential observations. The previous citations all include substantial sets of references on diagnostics techniques and applications for use with a single response variable.

Concepts such as bias, homoscedasticity, influential observations, and outliers are as relevant for nearest neighbors techniques as they are for regression techniques. Further, bias and homoscedasticity may be assessed for nearest neighbors techniques using methods similar to those used for regression techniques. However, measures associated with other concepts such as influential observations and extrapolation beyond the range of predictor variables require approaches unique to nearest neighbors techniques. Therefore, the objective of this study was to develop and illustrate graphical diagnostic tools for nearest neighbors techniques when used with satellite imagery to predict continuous forest attribute variables. Some of the tools illustrated are common to other statistical methods, but others are designed specifically to accommodate unique aspects of nearest neighbors techniques. With one exception (Section 3.6), all illustrations are for a single response variable. Although diagnostic tools for multiple continuous response variables and for categorical variables are also required for nearest neighbors techniques, an investigation of them is beyond the scope of this study.

E-mail address: rmcroberts@fs.fed.us.

0034-4257/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2008.06.015

The order of the following sections is somewhat atypical because of the necessity of first proposing diagnostic tools for nearest neighbors techniques and then illustrating the tools using applications with actual data. Following a brief overview of nearest neighbors techniques in Section 2, diagnostic tools for these techniques are proposed in generic terms without regard to particular data in Section 3. Section 4 includes a description of a dataset combining forest inventory ground data and Landsat satellite imagery, and Section 5 provides a brief overview of the analyses. Section 6 illustrates the results obtained for the proposed diagnostic tools using the study data, and Section 7 documents the study conclusions.

2. Nearest neighbors techniques

Let **Y** denote a possibly multivariate vector of response variables with observations for a sample of size *n* from a finite population of size *N*, and let **X** denote a vector of ancillary variables with observations for all population units. In the terminology of nearest neighbors techniques, the set of population units for which observations of both response and ancillary variables are available is designated the *reference set*; the set of population units for which predictions of response variables are desired is designated the *target set*; and the space defined by the ancillary variables, **X**, is designated the *feature space*. All elements of both the reference and target set are assumed to have a complete set of observations for all feature space variables.

For continuous response variables, the *k*-NN prediction for the *i*th target set element is,

$$\tilde{y}_i = \frac{1}{W_i} \sum_{j=1}^{k} w_{ij} y_j^i, \tag{1}$$

where $\{y_j^i, j=1,2,\ldots, k\}$ is the set of response variable observations for the k reference set elements that are nearest to the ith target set element in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the jth nearest neighbor with $W_i = \sum_{j=1}^{k} w_{ij}$. For categorical variables the predicted class of the ith target set element is the most heavily weighted class among the k nearest neighbors. Frequent choices for the weights are $w_{ij}=d_{ij}^t$ where $t\in[-2, 0]$. Many distance measures may be expressed in matrix form as,

$$d_{ij} = (\mathbf{X}_i - \mathbf{X}_j)' \, \mathbf{M} \, (\mathbf{X}_i - \mathbf{X}_j), \tag{2}$$

where i denotes a target set element for which a prediction is sought, j denotes a reference set element, $\mathbf{X}_i$ and $\mathbf{X}_j$ are the vectors of observations of feature space variables for the ith and jth elements, respectively, and **M** is a square matrix. Popular choices for **M** include the identity matrix which results in squared Euclidean distance and a diagonal matrix which results in weighted squared Euclidean distance. Squared Mahalanobis distance results when **M** is the inverse of the covariance matrix of the feature space variables (Kendall & Buckland, 1982). Other choices for **M** are based on canonical correlation analyses (LeMay et al., 2008; Moeur & Stage, 1995) or canonical correspondence analyses (Ohmann & Gregory, 2002). Chirici et al. (2008) and Hudak et al. (2008) evaluated additional distance metrics.

3. Diagnostic tools

In general, the approach to developing diagnostic tools focuses on a single continuous response variable and emphasizes graphical methods. In addition, where feasible and useful, the tools are based on the analysis of residuals calculated as differences between observations and predictions for reference set elements.

3.1. Homoscedasticity

As with many statistical techniques, an assessment of homoscedasticity is important for comparative and inferential purposes and for standardizing residuals to facilitate diagnostic analyses. For biological data, variances of observations and variances of residuals often increase as observations and predictions increase. Therefore, this study emphasizes assessment of variances or standard deviations with respect to response variables, in particular the relationships between standard deviations of residuals with respect to response variable predictions. The approach includes four steps: (1) order the reference set elements with respect to predictions; (2) select boundaries that produce ordered groups of reference set elements; (3) calculate the means of the residuals and predictions for each group; and (4) graph the group standard deviations against the group means. The size of the ordered groups is arbitrary, but it should be large enough to obtain reliable estimates of within group standard deviations; a guideline is that the groups should include at least 10 reference set elements. If the residual variance is homogeneous, then the standard deviations should lie along a line with slope 0. Formal statistical tests for assessing homoscedasticity for a single response variable may be found in Cook and Weisberg (1983).

The quality of nearest neighbors predictions is often evaluated using root mean square error,

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n}}, \tag{3}$$

where y_i and $\tilde{y}_i$ are the response variable observation and nearest neighbors prediction, respectively, for the ith reference set element, and n is the number of reference set elements. RMSE can be a useful measure for selecting optimal values of nearest neighbors operational parameters such as the distance metric, weighting schemes, and the k-value for particular datasets. However, when comparing the quality of predictions among different datasets using RMSE, the distributions of response variable observations for the two data sets should be homoscedastic. For multiple datasets with identical heteroscedastic relationships between residual standard deviations and response variable predictions, datasets with greater proportions of large response variable observations will have greater RMSEs. Therefore, in the absence of homoscedasticity, RMSE should generally not be used to compare the quality of predictions for different datasets.

An additional motivation for assessing homoscedasticity is estimation of the standard deviations of residuals to facilitate standardization of residuals preparatory to an analysis of outliers and influential observations.

3.2. Outliers and influential observations

Outliers are defined as values of observations that are so different from the rest of the observations that they raise questions as to whether they are from a different population or whether aspects of the sampling and/or the data are faulty (Kendall & Buckland, 1982). One method for detecting outliers is to calculate standardized residuals as the ratios of residuals and their standard deviations and then to assess the probability of observing the standardized residuals under specified distributional assumptions, usually Gaussian. Influential observations are those whose inclusion or exclusion cause substantial changes in predictions. For nearest neighbors techniques, the influence of reference set elements may be partially assessed using the sum of weights used to calculate predictions for either all reference set elements or all target set elements. When constant weights ($w_{ij}=1$) from Eq. (1) are assigned to each neighbor, the sum of the weights is simply the number of times each reference set element is selected as a nearest neighbor. Influential reference set elements and candidate outliers may be assessed jointly by graphing standardized residuals versus sums of weights for individual reference set elements. The effects of reference set elements with large absolute values of standardized residuals that are also used as nearest

neighbors large numbers of times should be assessed by comparing RMSEs obtained by including and excluding them as reference set elements. Formal statistical tests for detecting outliers for a single response variable may be found in Chatterjee and Yilmaz (1992) and Belsey et al. (1980).

3.3. Bias assessment

Nearest neighbors techniques are inherently biased because no prediction may be smaller or larger than the smallest or largest response variable observation in the reference set, respectively. When $k>1$ or when the correlation between response and feature space variables is weak, this effect may be exacerbated. Thus, bias assessment is crucial when using nearest neighbors techniques. A simple approach to assessing bias is to graph residuals for reference set elements against predictions for those same elements. If the predictions are unbiased, then the residuals should lie along a line with slope 0. The rationale for examining the relationship between the residuals and the predictions, rather than the residuals and the observations, is that whereas the residuals and observations are likely to be correlated, the residuals and the predictions do not exhibit such correlation (Draper & Smith, 1981, p.147).

If the reference set is large, the relationship between the residuals and the predictions may be visually masked by the sheer quantity of data. An alternative is to order the reference set elements with respect to their predictions and then graph means of residuals for ordered groups of reference set elements against means of predictions for the same groups. If the predictions are unbiased, then the means of residuals should also lie along a line with slope 0. A second alternative is to graph observations for reference set elements against corresponding predictions, either singly or in groups. If the predictions are unbiased, then the observations should lie along the 1:1 line which has intercept 0 and slope 1. Finally, a third alternative is to graph a representative sample of residuals or observations against their respective predictions. Similar approaches could be used with ancillary variables rather than predictions.

When the feature space is spatial in nature and reference set elements have been selected using a probability-based sampling design, bias may be partially assessed using an areal approach conducted in four steps: (1) select small areas of interest (AOI) that contain a minimum number of reference set elements; (2) calculate the probability-based (design-based) estimate of the mean and standard error of the estimate of the mean for the response variable observations for reference set elements with centers in each AOI; (3) calculate the mean of predictions for all target set elements with centers in each AOI; and (4) graph the means of the response variable observations and their associated 2-standard error confidence intervals against the means of predictions for the target set elements. If the predictions are unbiased, the means of the response variable observations should lie along the 1:1 line, and the 1:1 line should cross the 2-standard error confidence intervals. The latter feature follows because the 1:1 line depicts the means of target set predictions which, if they are unbiased, should be within the 2-standard error confidence interval for the probability-based means. Two premises underlie this approach: first, nearest neighbors predictions and means dependent on them cannot be assumed to be unbiased, and second, for appropriate sampling designs, probability-based means are known to be at least asymptotically unbiased (Cochran, 1977). Thus, a reasonable and feasible approach for assessing the bias of means for AOIs based on nearest neighbors predictions is to compare them to corresponding probability-based means for the same AOIs. McRoberts (2006) and McRoberts et al. (2007) illustrate this approach with AOIs of different sizes.

Caution must be exercised in the selection of the size of the AOIs. If they are too large, then random selection of neighbors will produce means comparable to large sample probability-based means. Conversely, if the AOIs are too small, the sizes of the confidence intervals for the probability-based means will be so large that the power of any test to detect differences will be small. Although the minimum number of reference set elements per AOI is arbitrary, it should be large enough that the assumption of asymptotic unbiasedness for the probability-based estimates is approximately satisfied. Results obtained using this approach are only approximate because no accommodation is made for the uncertainty in the means of the predictions for the target set elements. Therefore, for purposes of evaluating bias, the approach is conservative in the sense that it may indicate bias when none exists.

3.4. Extrapolations

All statistical prediction techniques are vulnerable to poor performance when predictions are extrapolated beyond the range of the independent or predictor variables. Nearest neighbors techniques are particularly vulnerable because, unlike regression techniques, there is no predictive parametric curve that can be easily extended. In addition, there is no predictive parametric curve to bridge gaps in the distribution of feature space variables. Further, the situation for gaps is exacerbated when using small values of k, particularly $k=1$. Thus, for nearest neighbors techniques, the assumption that the distribution of observations of the feature space variables for the reference set is comparable to the distribution for the target set is crucial.

Several methods for comparing the distributions of observations of feature space variables are feasible and useful. An initial, albeit naïve, approach is to compare the range of observations for each feature space variable for the reference set and target sets. Target set pixels with observations for one or more feature space variables beyond the range of the variables for the reference set require extrapolations. This approach, however, does not take into account the proportion of target set elements for which extrapolations are necessary. A more informative extension of this approach entails counting the number of pixels in the target set requiring extrapolations. This approach and its extension are characterized as naïve because they do not account for the possibility that feature space variables may be correlated. Strong correlations increase the likelihood that observations for a pair of feature space variables for a target set pixel may be within reference set ranges for both variables considered singly but outside their two-dimensional boundary when considered jointly. Thessler et al. (2005) describe an approach for accommodating correlation among feature space variables when identifying pixels requiring extrapolations. With this approach, reference set observations for pairs of feature space variables are graphed, and convex hulls are constructed around the observations. Predictions for target set pixels outside the convex hull are also considered extrapolations. Although this approach is computationally intensive, it can be implemented and automated using geographic information system techniques. The two-dimensional convex hull is an improvement over the naïve approach, but a convex hull of the same dimension as the feature space is required to identify all pixels requiring extrapolations.

Insight into the consequences of extrapolations may be obtained by analyzing the nearest neighbors used for prediction. For example, if a large proportion of nearest neighbors for an extrapolation are similar or of the same category, then the consequences of the extrapolation may be minimal.

3.5. Distribution of reference set elements in feature space

A concern with nearest neighbors techniques is that distances to nearest neighbors can be relatively large in regions of feature space where reference set elements are sparsely distributed; examples include large gaps and regions near the edges of observed feature space. The effects of this phenomenon could include both bias and greater residual variance. An approach to assessing these effects is to

graph response variable observations, residuals, and standard deviations of residuals against feature space metrics such as distance to nearest neighbor, mean distance to the k nearest neighbors, and distance to feature space center.

A related issue pertains to the combined effects of the value of k and the distribution of reference set elements in feature space on the ecological consistency of predictions for multiple response variables. When the k nearest neighbors are in close proximity in feature space, the observations of the response variables for those neighbors may be expected to be similar, and the multivariate means over those k neighbors may be expected to represent a condition that could reasonably be expected to occur naturally. However, when k is large, particularly in regions of feature space where reference set elements are sparsely distributed, the possibility of selecting ecologically incompatible neighbors is greater. For example, a combination of high volume, low tree density neighbors and low volume, high tree density neighbors could produce a prediction that does not occur naturally.

3.6. Preservation of covariances

Some authors have asserted that covariances among observations are preserved in nearest neighbors predictions when using small k-values (Hudak et al., 2008; LeMay & Temesgen, 2005; McRoberts et al., 2002; Moeur & Stage, 1995; Tomppo et al., 2002). Despite the frequency of this assertion, no analyses of its validity are known to have been reported. This study addresses two aspects of the assertion: (1) preservation of covariances among multiple response variables, and (2) preservation of spatial covariances for individual response variables. Covariances among pairs of response variables are estimated as,

$$\hat{\mathrm{Cov}}(y_p, y_q) = \frac{1}{n}\sum_{i=1}^{n}\left(y_{pi}-\bar{y}_p\right)\left(y_{qi}-\bar{y}_q\right), \tag{4}$$

where p and q index response variables, and $\bar{y}_p$ and $\bar{y}_q$ are the means of observations or predictions for the pth and qth response variables, respectively. Spatial covariances for individual response variables are estimated using empirical semi-variograms calculated as,

$$\gamma(d) = \frac{1}{2\|N(d)\|}\sum_{N(d)}\left(y_{pi}-y_{pj}\right)^2, \tag{5}$$

where $N(d)$ denotes a collection of pairs of reference or target set elements for the p^{th} response variable whose Euclidean separation distances in geographic space are within a given neighborhood of d, and $\|N(d)\|$ denotes the number of pairs in $N(d)$ (Cressie, 1993).

A reasonable assumption is that the degree to which covariances are preserved is related to the quality of predictions. The assumption may be evaluated by comparing covariances obtained for predictions for $k=1$ using the first nearest neighbor and covariances obtained for predictions for $k=1$ using (for example) the fifth nearest neighbor alone, the 10th nearest neighbor alone, and the 20th nearest neighbor alone. If the similarity between covariances for observations and covariances for predictions decreases when moving from the first to the 20th nearest neighbor, then poorer quality predictions may be inferred to have a detrimental effect on the preservation of covariances.

4. Data

4.1. Satellite imagery

The study area was defined by the row 27, path 27, Landsat Thematic Mapper (TM) scene in northern Minnesota, USA (Fig. 1). Land use for the study area consists of forest land dominated by aspen–birch and spruce–fir associations (McRoberts et al., 2008), wetlands, water, and agriculture.

Imagery for the scene was obtained from the Multi-Resolution Characterization 2001 (MRLC 2001) land cover mapping project (Homer et al., 2004) of the U.S. Geological Survey. The imagery was characterized by several salient features: (1) a combination of Landsat 5 TM and Landsat 7 ETM+ data, (2) geometrically and radiometrically corrected, (3) cubic convolution resampling to 30 m×30 m spatial resolution, (4) visible and infrared bands (1–5, 7), and (5) conversion to at-satellite reflectance. Imagery for three dates corresponding to early, peak, and late vegetation green-up (Yang et al., 2001) were March 2000, July 1999, and October 1999. Preliminary analyses indicated that normalized difference vegetation index (NDVI) (Rouse et al., 1973) and the tasseled cap (TC) transformations (brightness, greenness, and wetness) (Crist & Cicone, 1984; Kauth & Thomas, 1976) were superior to both the spectral band data and principal component transformations with respect to predicting the probability of forest cover. Thus, 12 satellite image-based variables were used, NDVI and the three TC transformations for each of the three image dates.

4.2. Forest inventory data

The Forest Inventory and Analysis (FIA) program of the U.S. Forest Service conducts the national forest inventory of the USA. The program has established field plot centers in permanent locations using a sampling design that is assumed to produce a random, equal probability sample (Bechtold & Patterson, 2005; McRoberts et al., 2005). The sampling design is based on a tessellation of the USA into approximate 2400-ha (6000-ac) hexagons and features a permanent plot at a randomly selected location in each hexagon. Some states, including Minnesota in which the study area is located, provide additional funding to double the sampling intensity to approximately one plot per 1200 ha.

Each 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. In general, locations of forested or previously forested plots are determined using global positioning system receivers, and locations of non-forested plots are verified using aerial imagery and digitization methods.

Field crews measure diameter at-breast-height (dbh, 1.37 m, 4.5 ft) and height for all trees with dbh of at least 12.7 cm (5 in). Statistical models are used with data for individual trees to predict tree volume. Data for all trees are aggregated to produce per unit area estimates of basal area (BA, m^2/ha), growing stock volume (VOL, m^3/ha), and tree density (TD, tree count/ha) for subplots. In addition, field crews visually estimate the proportions of subplot areas that satisfy specific ground land use conditions. Subplot estimates of proportion forest area (FOR) are obtained by collapsing land use conditions into forest and non-forest classes consistent with the FIA definition of forest land: primarily area of at least 0.4 ha (1 ac), external crown-to-crown width of at least 36.58 m (120 ft), stocking of at least 10%, and forest land use. Data for 2266 plots or 9064 subplots observed between October 1999 and September 2004 were available for the study area.

4.3. Combining FIA data and satellite imagery

The spatial configuration of the FIA subplots with centers separated by 36.58 m and the 30-m×30-m spatial resolution of the TM/ETM+ imagery permits individual subplots to be associated with individual image pixels. The subplot area of 167.87 m^2 is approximately 19% of the 900 m^2 pixel area and is assumed to be sufficient to characterize entire pixels containing subplot centers.

For many of the diagnostic analyses, only the reference set observations and their corresponding predictions are used. However, for identifying influential reference set elements, assessing areal bias,

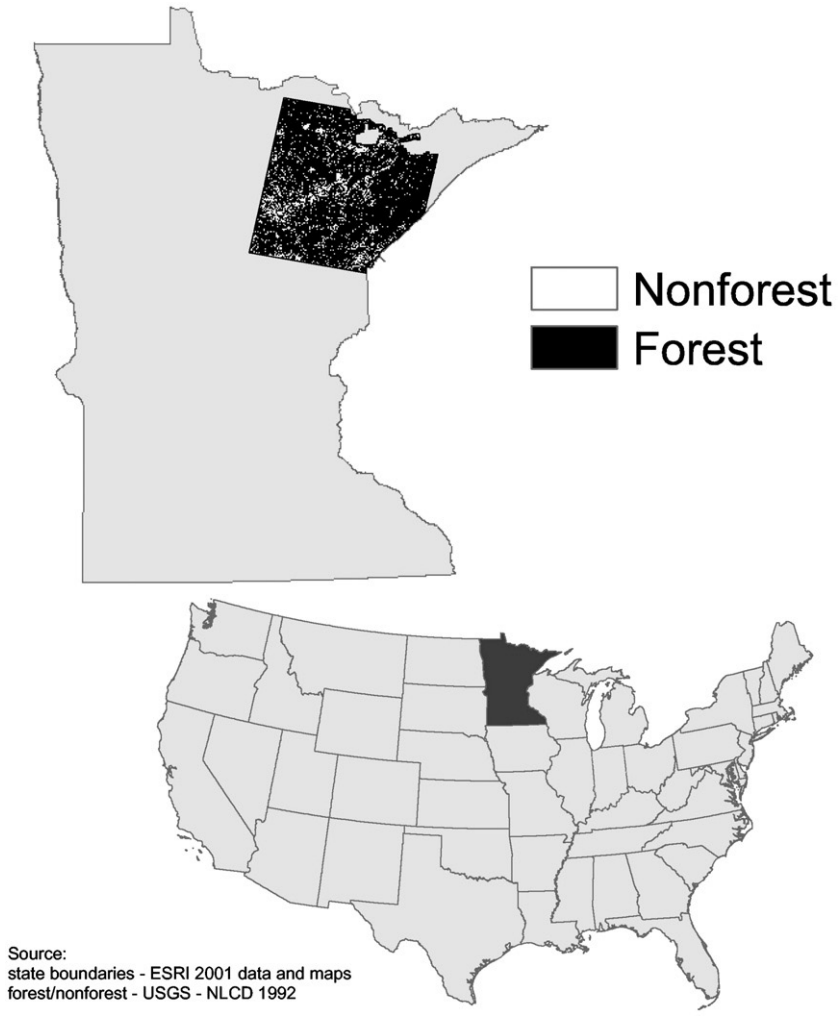

Fig. 1. Study area.

identifying pixels requiring extrapolations, and assessing preservation of covariances, predictions for target set elements also may be necessary. For these analyses, 15 AOIs of radius 15 km with centers selected using a systematic grid of locations were used.

5. Analyses

For this study, the population is the set of 30-m×30-m Landsat pixels for an AOI; $\mathbf{Y}_i$ consists of observations of FOR, BA, VOL, and TD for the FIA subplot with center in the ith pixel; and $\mathbf{X}_i$ is the vector of observations of the feature space variables for the ith pixel. The feature space variables are NDVI and the three TC transformations for each of the three image dates. In general, only the VOL response variable is used to illustrate the diagnostic tools, although all four response variables are used when illustrating tools for assessing the preservation of covariances.

A basic implementation of the k-NN technique was used with Euclidean distance, constant weighting of nearest neighbors ($w_{ij}=1$), and $1 \le k \le 50$. The leave-one-out cross validation technique (Lachenbruch & Mickey, 1986) was used to calculate predictions for each reference set pixel subject to the constraint that nearest neighbors could not be selected from among pixels corresponding to subplots of the same FIA plot as the pixel for which the prediction was calculated because of the high correlation among subplot observations for the same plot.

Selection of the value for k is a subjective decision guided by multiple, competing, objective considerations. On the one hand, small values of k reduce computational intensity, better preserve observed covariances among response variables, and provide a measure of protection against ecological inconsistency. On the other hand, larger values of k may be necessary to bridge gaps in feature space and to reduce prediction bias, particularly for data sets with weak relationships between response and feature space variables. For the latter condition, the value of k that minimizes a criterion such as RMSE may be very large. In addition, when multiple response variables are to be predicted, the value of k that minimizes a function of RMSEs for the individual variables may be considered; e.g.,

$$C = \sum_{j=1}^{J} \frac{SS_{j,\text{mean}} - SS_{j,\text{nn}}}{SS_{j,\text{mean}}},$$

where j indexes response variables, $SS_{j,\text{mean}}$ is the sum of squared deviations of observations from their mean for the jth variable, and $SS_{j,\text{nn}}$ is the sum of squared nearest neighbor residuals for the same variable. To minimize bias and simultaneously accommodate the generally accepted principle of selecting the smallest reasonable value of k, McRoberts et al. (2002) proposed selecting the smallest value of k for which RMSE is only a small proportion larger than the minimum RMSE. Although different values of k were selected to illustrate different diagnostic tools for this study, selection of the

value of k was guided primarily by the subjective decision to minimize bias in predictions.

6. Results

6.1. Homoscedasticity

As is common for biological variables such as VOL, the graph of the standard deviation of residuals against mean predictions clearly indicates heteroscedasticity (Fig. 2). Therefore, the RMSE obtained for VOL using nearest neighbor techniques for this study should not be compared to RMSEs obtained for other data sets unless the distributions of observations are very similar. In the absence of homoscedasticity, the effects on RMSE may be considerable when samples with different proportions of small and large observations are obtained from identical distributions. For the analysis of influential observations and outliers, standardized residuals were calculated by dividing observed residuals by predictions of their standard deviations obtained from a statistical model fit to the data used to assess homoscedasticity (Fig. 2).

6.2. Outliers and influential observations

Reference set pixels in the upper and lower right of the graph of standardized residuals versus number of times used as a neighbor should be carefully scrutinized as candidates for influential observations and/or outliers (Fig. 3). Assuming standardized residuals follow a Gaussian (0, 1) distribution, the proportion of standardized residuals with absolute values greater than 2.0 and 3.0 should be less than 0.045 and 0.001, respectively. However, when response and feature space variables are observed independently, at different times, and on sample units of different sizes, a greater proportion of large standardized residuals may be observed than would be expected under the Gaussian assumption. For example, an FIA subplot is only approximately a 19% sample of a TM pixel, so the subplot observation may not adequately represent the entire pixel. In addition, if the subplot observation date differs from the image date, the pixel spectral values may not represent observed ground conditions. Also, the subplots may be misregistered to the image pixels causing incorrect spectral data to be associated with the ground observation. Under these and similar conditions, outliers may present a serious problem. Finally, the standard deviations of residuals used to calculate standardized residuals are estimated from a curve fit to group data. Of necessity, these estimates have uncertainty associated with them, and that uncertainty propagates to the standardized residuals.

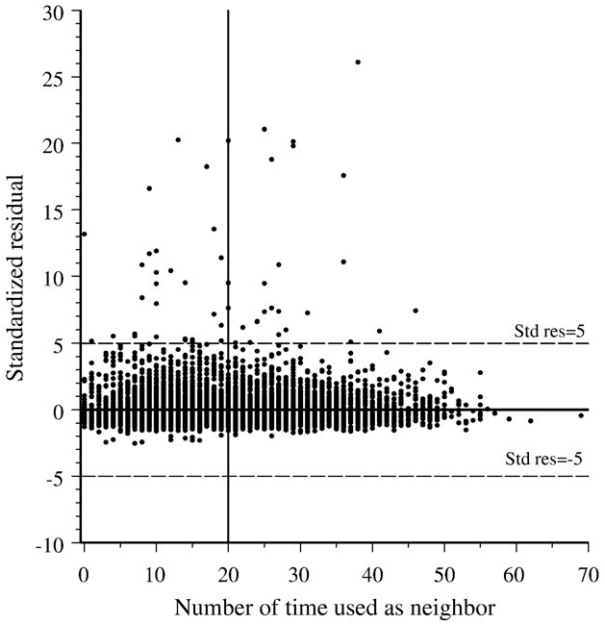

Fig. 3. Influential observations and candidate outliers, $k=20$.

An assessment of the effects on RMSE of deleting reference set pixels with large standardized residuals, particularly those used a large number of times, was conducted. For $k=5$, deletion of all reference set pixels with absolute values of standardized residuals greater than 3.0 produced a proportional reduction in the size of the reference set of only approximately 0.03 but a proportional reduction in RMSE of more than 0.20 (Fig. 4). For larger k-values, the

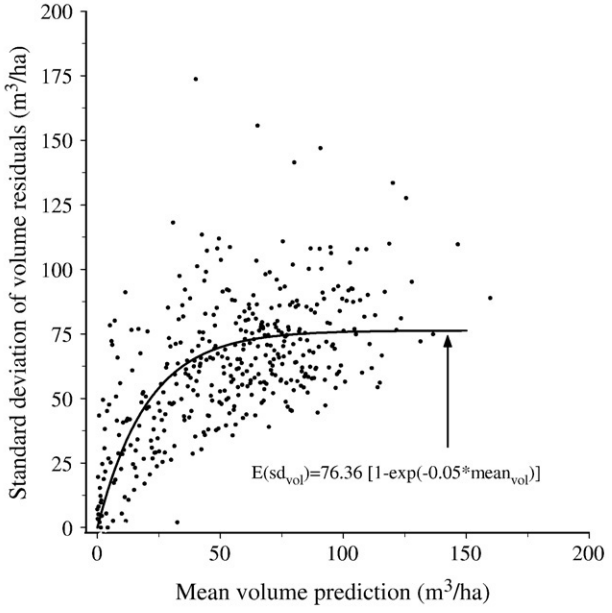

Fig. 2. Residual standard deviation versus mean prediction for VOL, $k=20$, group size=20.

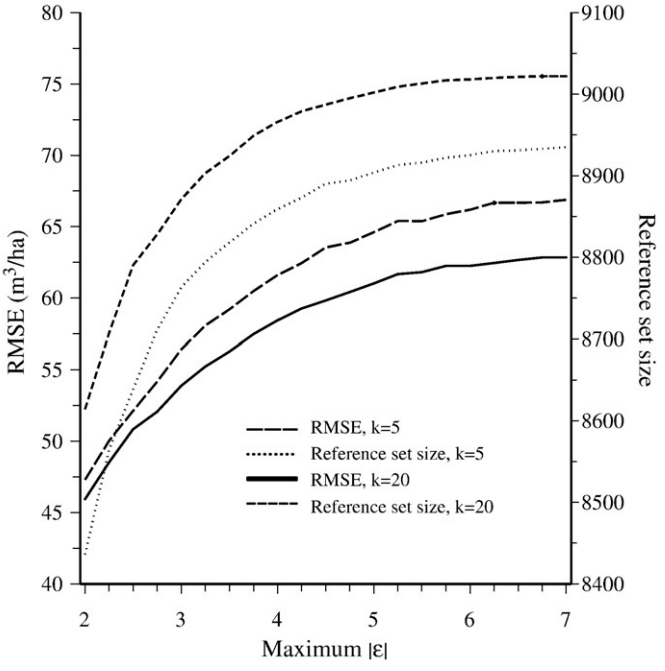

Fig. 4. RMSE and reference set size when deleting reference set pixels with absolute values of standardized residuals as great or greater than $|\varepsilon|$.

proportional reduction in RMSE was smaller because the effects of individual outliers among nearest neighbors are reduced when larger numbers of neighbors are used in calculating predictions. However, even for $k=20$, some standardized residuals were much larger in absolute value than 3.0 and, in addition, some of them should be considered influential because they were used as nearest neighbors many times more than $k=20$ as would be expected (Fig. 3). Because of the relatively small number of candidate outliers, but their relatively large effect on RMSE, observations with very large absolute values of standardized residuals should be seriously considered for deletion from the reference set.

However, before deleting influential reference set pixels that are also candidate outliers, several relevant questions should be posed: are potential influential observations and/or candidate outliers concentrated in a particular region of feature space; are they concentrated in a particular region of the continuum of the response variable observations; and if deletion of such a reference set pixel improves RMSE obtained using the leave-one-out technique, is the improvement due to better predictions or elimination of difficult-to-predict reference set pixels? For this study, influential reference set pixels were similar to the rest of the reference set, but the candidate outliers tended to have large VOL observations. Thus, no reference set pixels were deleted because of concern that the quality of predictions for target set pixels would decrease.

6.3. Bias assessment

For many statistical prediction techniques, the recommended approach for assessing bias is to graph residuals against predictions. However, when response variables have defined lower and/or upper limits, this approach may not be useful. For example, because VOL has a lower limit of 0, residuals corresponding to nearest neighbor predictions of $\tilde{y}=0$ cannot be negative (Fig. 5). A more informative approach for nearest neighbors techniques is to graph reference set observations against corresponding predictions or means of ordered groups of observations against means of predictions for the same ordered groups (Fig. 6). Graphs of means of observations for reference set observations against corresponding means of predictions exhibited bias for small and large predictions, although the effect was less pronounced for $k \geq 20$.

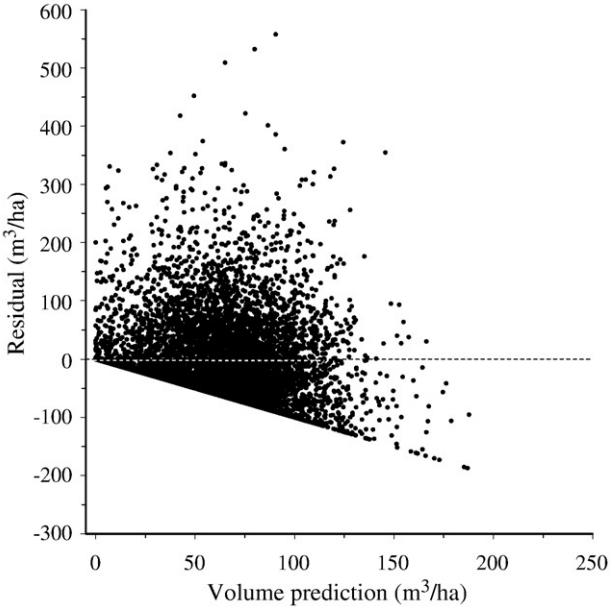

Fig. 5. Residuals versus predictions for VOL, $k=20$.

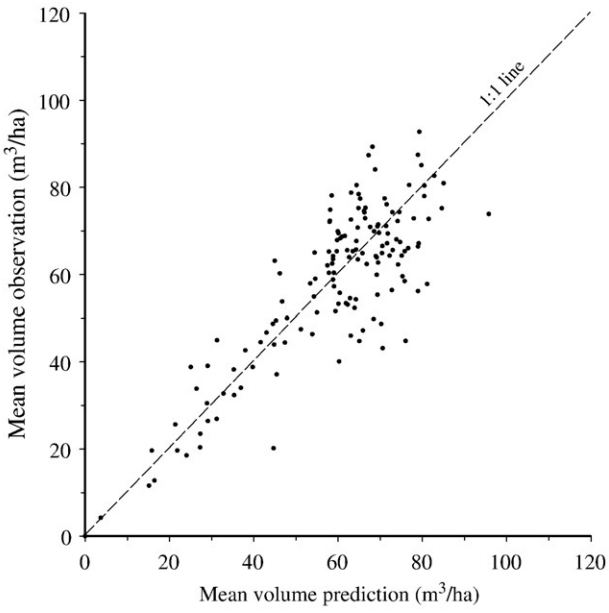

Fig. 6. Mean observation versus mean prediction for VOL, $k=20$, and group size=20.

Based on the foregoing analyses of influential observations, outliers, and bias, $k=20$ was selected for calculating the nearest neighbor predictions for VOL (Fig. 7). Although $k=45$ yields the minimum RMSE=64.68, $k=20$ is less than half $k=45$ and yields RMSE=65.40 which is only 1% larger than the minimum. The RMSE versus k graph illustrates a precaution previously noted by McRoberts et al. (2002) that for small k-values, RMSE may actually be greater than the standard deviation around the mean.

For an areal bias assessment, means of pixel predictions obtained using $k=20$ were generally similar to the probability-based means and were within 2-standard error probability-based confidence intervals for 10-km radius AOIs (Fig. 8). Correlations between probability-based plot means and means of pixel predictions were $\rho=0.64$ for FOR, $\rho=0.77$ for VOL, $\rho=0.79$ for BA, and $\rho=0.82$ for TD, suggesting strong correspondence between the two sets of estimates.

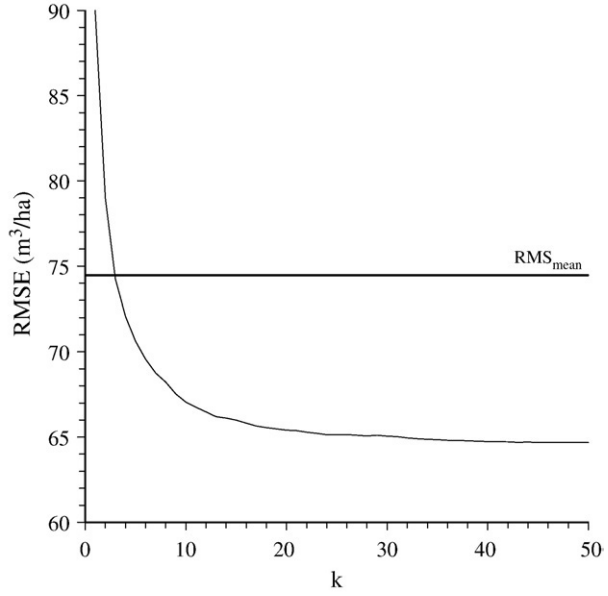

Fig. 7. RMSE versus k for VOL.

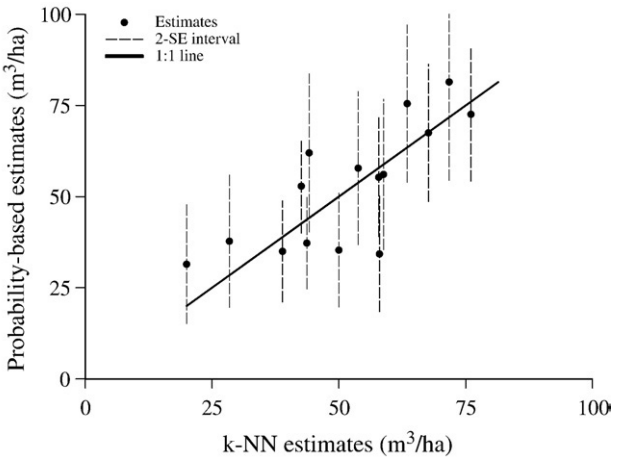

Fig. 8. Probability-based plot means and 2-standard error confidence intervals versus *k*-NN pixel-based areal means for VOL, *k*=20.

6.4. Extrapolations

The naïve approach to assessing extrapolations revealed that the ranges for at least some feature space variables were considerably greater for the target set than for the reference set (Fig. 9). For the 12 feature space variables, the ratios of the ranges of spectral values for the reference set to those for the target set were between 0.53 and 0.95 with a mean of 0.70. Although these differences suggest the possibility of a large number of extrapolations beyond the reference set data, a count of the proportion of pixels requiring extrapolations based on this naïve approach was small, approximately 0.001. However, the two-dimensional convex hull approach identified a much greater proportion of pixels requiring extrapolations, approximately 0.02 (Fig. 10). As expected, some target set pixels requiring extrapolations were identified for multiple pairs of feature space variables. Of target set pixels requiring extrapolations, 36% were identified by only a single pair of feature space variables, 18% were identified by two pairs, and 10% were identified by three pairs. For

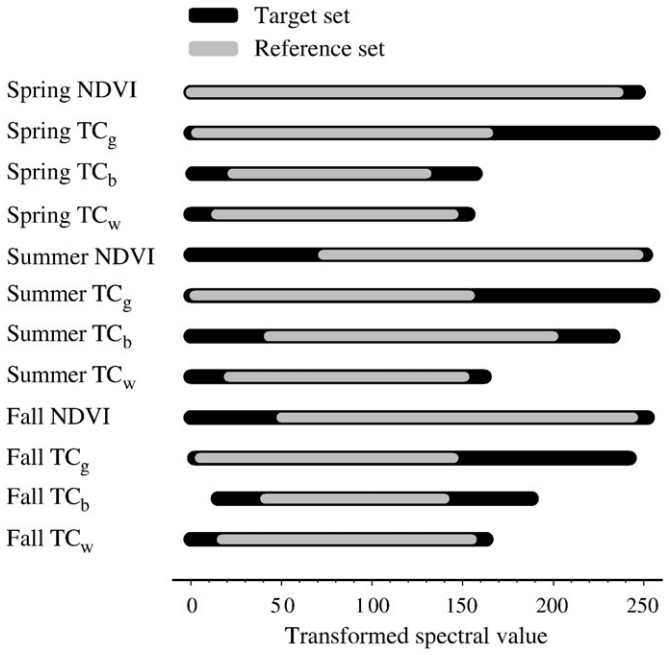

Fig. 9. Ranges of feature space variables for reference and target sets (NDVI = normalized difference vegetation index; TC = tasseled cap; subscripts: g = greenness, b = brightness, w = wetness).

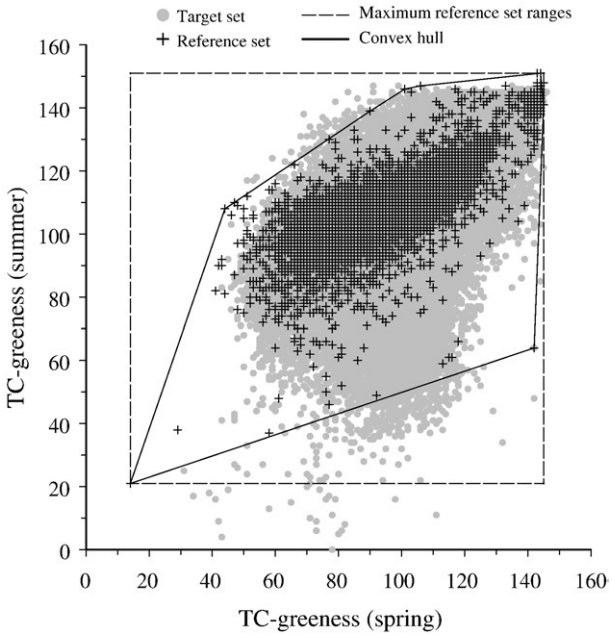

Fig. 10. Extrapolation analyses for a single AOI. Pixels outside the dashed line box are identified as requiring extrapolations by the naïve approach; pixels between the dashed line box and the solid line polygon are additionally identified as requiring extrapolations by the convex hull approach but not by the naïve approach.

target set pixels requiring extrapolations, the *k*=20 nearest neighbor reference set pixels were generally non-forest: 42% had 20 of 20 non-forest nearest neighbors, 7% had 19 of 20 non-forest nearest neighbors, 4% had 18 of 20 non-forest nearest neighbors, and 6% had 17 of 20 non-forest nearest neighbors. Slightly less than 30% had more forest nearest neighbors than non-forest nearest neighbors. Therefore, because the proportion of target set pixels requiring extrapolations was small, and the majority of extrapolations were based on nearest neighbors that were predominantly non-forest, any detrimental effects of extrapolations were ignored for this study. Nevertheless, several aspects of identifying extrapolations merit emphasis: (1) the naïve approach based on simply comparing ranges substantially underestimates the number of extrapolations; (2) the

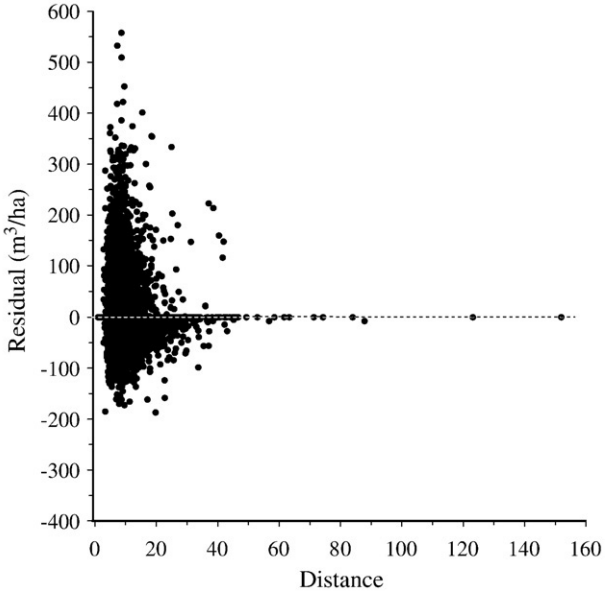

Fig. 11. Residual versus feature space distance to first nearest neighbor for VOL, *k*=20.

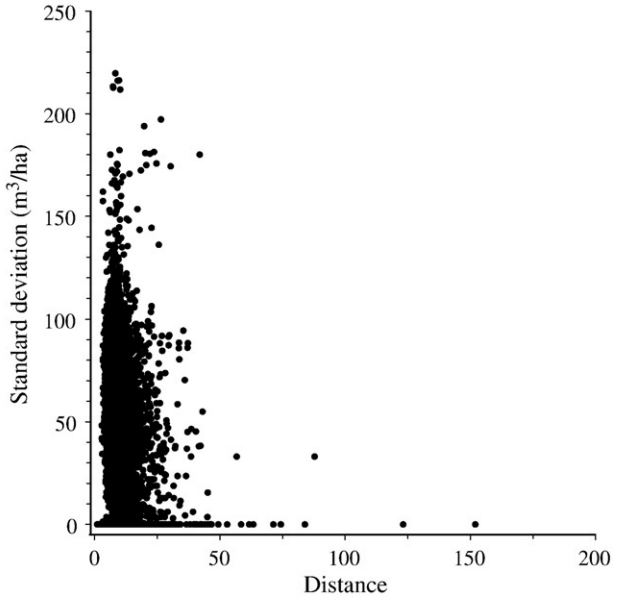

Fig. 12. Standard deviation among $k=20$ nearest neighbors for VOL versus mean feature space distance to 20 nearest neighbors.

two-dimensional convex hull approach is an improvement, but probably still misses some pixels requiring extrapolations; and (3) reduction of the feature space dimension should be considered to reduce the number of extrapolations and to reduce the computational intensity necessary to identify them.

For applications with geographically uncertain plot locations or clusters of plots extending over a relatively small number of contiguous satellite image pixels, means of spectral values for blocks of pixels may be assigned to reference set elements rather than spectral values for individual pixels. For these applications, means of spectral values must then be assigned to target set pixels in the same manner as they are assigned to reference set pixels. Assignment of means of spectral data to reference set pixels but not to target set pixels exacerbates the effects of differences in ranges of spectral feature space variables and may cause the number of extrapolations to increase substantially. For example, ranges of means of spectral values for 3×3 blocks of pixels centered at pixels containing FIA subplot centers were, on average, 16% less than ranges of spectral values for the single pixels containing the subplot centers.

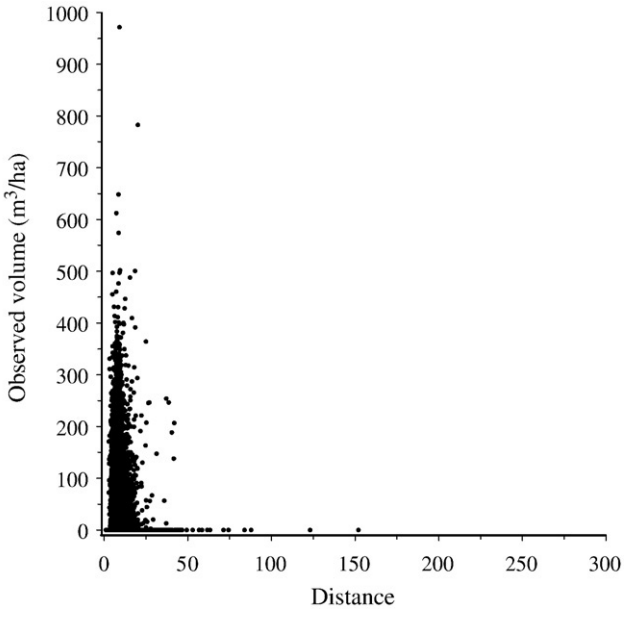

Fig. 13. Observation versus distance to feature space center for VOL.

Table 1
Correlations among response variable observations for the reference set

	FOR	VOL	BA	TD
FOR	1.00	0.41	0.45	0.47
VOL		1.00	0.97	0.76
BA			1.00	0.86
SD				1.00

Table 2
Ratios of covariances among response variable predictions to response variable observations for the reference set

	Reference set				AOIs			
	FOR	VOL	BA	TD	FOR	VOL	BA	TD
k=1								
FOR	0.9562	0.9462	0.9453	0.9576	0.8531	0.9144	0.9234	0.9311
VOL		0.9847	0.9759	0.9849		1.0656	1.0744	1.0640
BA			0.9717	0.9835			1.0860	1.0759
TD				0.9933				1.0825
k=5								
FOR	0.6500	0.7217	0.7263	0.7415	0.5613	0.6472	0.6574	0.6735
VOL		0.3939	0.4145	0.4791		0.3950	0.4203	0.4814
BA			0.4291	0.4777			0.4412	0.4873
TD				0.4743				0.4854
k=20								
FOR	0.5542	0.6357	0.6428	0.6608	0.4741	0.5751	0.5871	0.6080
VOL		0.2388	0.2638	0.3353		0.2402	0.2677	0.3386
BA			0.2815	0.3343			0.2887	0.3417
TD				0.3286				0.3385
k=1 using 5th nearest neighbor								
FOR	0.9713	0.9814	0.9931	1.0161	0.8183	0.8524	0.8664	0.8765
VOL		0.9315	0.9633	1.0229		1.0214	1.0407	1.0408
BA			0.9922	1.0410			1.0672	1.0631
TD				1.0703				1.0770
k=1 using 20th nearest neighbor								
FOR	0.9347	0.8993	0.9068	0.9230	0.8239	0.8748	0.8863	0.9025
VOL		0.8963	0.9155	0.9548		1.0753	1.0824	1.0780
BA			0.9359	0.9707			1.0972	1.1080
TD				0.9942				1.1441

6.5. Distribution of reference set pixels in feature space

A concern with nearest neighbors techniques is that sparseness of reference set pixels near the edges of observed feature space and/or at large distances from the center of that space may affect the quality of predictions. However, graphs of residuals, observations, and standard deviations of VOL among the 20 nearest neighbors against distance to nearest neighbor, mean distance to the 20 nearest neighbors, and distance to feature space center indicated no apparent ill effects (Figs. 11–13). In fact, these graphs suggest that near the edges of observed feature space, the VOL surface is rather uniform with only small or 0 values. This phenomenon may be attributed to the fact that land cover in the study area is mostly forested and, as a result, pixels with forest cover are near the center of feature space, whereas pixels with non-forest cover are at the edges of observed feature space. Further, because the study included a variety of non-forest land covers such as water, roads, and agriculture, all with different spectral signatures, spectral values of pixels with non-forest cover may be diverse with the result that these pixels may be found in different directions from feature space center.

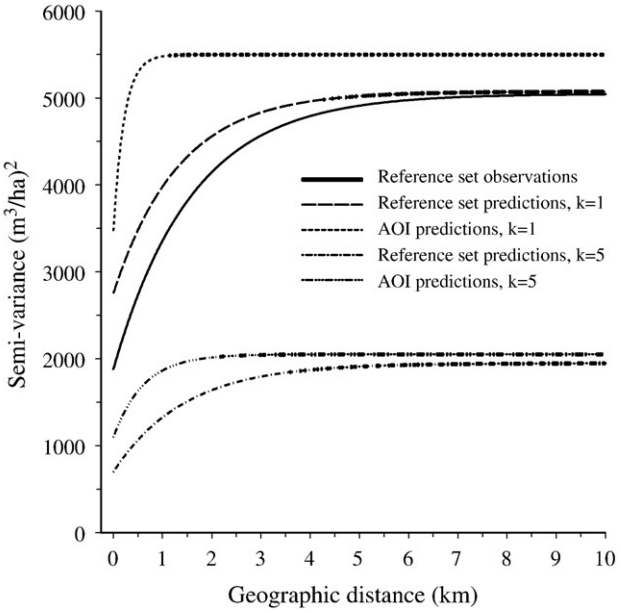

Fig. 14. Spatial semi-variograms.

6.6. Preservation of covariances

Comparisons of covariances among FOR, VOL, BA, and TD for observations and predictions indicate that covariances are preserved only for $k=1$ (Tables 1 and 2). For $k=1$, the ratios of covariances for the observations for reference set pixels and covariances for predictions for reference set pixels ranged from approximately 0.95 to 1.00; for $k=5$, the ratios ranged from approximately 0.40 to 0.75; and for $k=20$, the ratios ranged from approximately 0.25 to 0.65. With $k=1$, and using the fifth and 20th nearest neighbors alone as predictions, ratios were usually between 0.90 and 1.00, although they were smaller for the 20th nearest neighbor than for the fifth nearest neighbor. The latter result suggests that the degree to which covariances are preserved for $k=1$ decreases as the quality of predictions decreases. For multiple pixel AOIs, covariances among observations for reference set pixels tended to be preserved in predictions for $k=1$ but not for $k>1$. For $k=1$, ratios ranged from approximately 0.85 to 1.10; for predictions based on the fifth nearest neighbor, the ratios ranged from approximately 0.80 to 1.10; and for predictions based on the 20th nearest neighbor, the ratios ranged from approximately 0.80 to 1.15.

Preservation of spatial covariances for VOL was assessed using semi-variograms for observations and predictions for reference set pixels and predictions for target set pixels (Fig. 14). In general, spatial covariances tended to be preserved for $k=1$ but not for $k>1$. In addition, covariances among reference set observations were preserved better for reference set predictions than for target set predictions. However, because few FIA plots were co-located in closer proximity than a few km, considerable uncertainty may be associated with the portions of the semi-variograms corresponding to small geographic distances.

7. Conclusions

Several important conclusions can be drawn from this study. First, an assessment of influential reference set elements and candidate outliers should be conducted whenever the spatial locations of sample units are questionable and whenever the spatial locations of sample units for which response variables are observed must be registered to the spatial locations of units for which feature space variables are observed. Second, bias assessments may be conducted for pixels and for AOIs consisting of multiple pixels. For this study, any effects of pixel-level bias did not result in statistically significant bias at the areal level. Third, because of the nature of nearest neighbor techniques, extrapolations should always be expected, although their number and effects generally cannot be anticipated. For this study, the proportion of target set pixels requiring extrapolations was small, and their effects were ignored. However, identifying pixels requiring extrapolations was a computationally intensive task. Fourth, before making any assertions regarding preservation of covariances by using small k-values, the validity of the assertion should be checked. For this study, covariances among observations of response variables were well-preserved only for predictions for reference set pixels using $k=1$. However, even with $k=1$, the analyses suggested that covariances were less well-preserved with less accurate predictions and less well-preserved for target set predictions than for reference set predictions. Further, spatial covariances were preserved only under very specific conditions. Finally, although the utility of these diagnostic tools may be generalized to all nearest neighbors analyses, the specific findings for this study should not be generalized beyond the particular data used.

Any exposition and/or illustration of diagnostic tools for nearest neighbors techniques that does not address categorical and multivariate response variables is necessarily incomplete. Thus, ample opportunity exists for future research in this area.

Acknowledgements

The author gratefully acknowledges the assistance of Lisa G. Mahal, University of Nevada, Las Vegas, and Greg C. Liknes, Northern Research Station, U.S. Forest Service, for GIS assistance in implementing and automating the two-dimensional convex hull procedure for identifying target set pixels requiring extrapolations. The author also acknowledges three anonymous reviewers whose comments and suggestions have contributed to a substantially improved manuscript.

References

Atkinson, A. C. (1986). Regression diagnostics. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, *Volume 7*. (pp.)New York: Wiley.

Atkinson, A. C. (1984). Two books on regression diagnostics. *The Annals of Statistics*, *12*(1), 392–401.

Baffeta, F., Fattorini, L., Franceschi, S., & Corona, P. (2009). Design-based approach to k-NN technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, *113*, 463–475 (this issue).

Bechtold, W. A., & Patterson, P. L. (Eds.). (2005). *The enhanced Forest Inventory and Analysis Program — national sampling design and estimation procedures*. General Technical Report SRS-80. Asheville, North Carolina: Southern Research Station, USDA Forest Service 85 pp.

Belsey, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and collinearity*. New York: Wiley.

Box, G. E. P., Hunter, J. S., & Hunter, G. W. (1978). *Statistics for experimenters*. New York: Wiley.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*, second edition. New York: Wiley.

Chatterjee, S., & Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. *Applied Psychological Measurement*, *16*(3), 209–227.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., & Maselli, F. (2008). Non-parametric and parametric methods using satellite imagery for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sensing of Environment*, *112*, 2686–2700.

Cochran, W. G. (1977). *Sampling techniques*, 3rd Edition. New York: Wiley.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, *70*, 1–10.

Cressie, N. A. C. (1993). *Statistics for spatial data*, revised edition. New York: Wiley.

Crist, E. P., & Cicone, R. C. (1984). Application of the tasseled cap concept to simulated thematic mapper data. *Photogrammetric Engineering and Remote Sensing*, *50*, 343–352.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York: Wiley.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis*, second edition New York: Wiley.

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*, 251–1709.

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 national landcover database for the United States. *Photogrammetric Engineering and Remote Sensing*, *70*, 829–840.

Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E., & Falkowski, M. J. (2008). Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, *112*, 2232–2245.

Katila, M., & Tomppo, E. (2005). Empirical errors of small area estimates from the multisource National Forest Inventory in Eastern Finland. *Silva Fennica*, *40*, 729–742.

Kauth, R. J., & Thomas, G. S. (1976). The Tasseled Cap — a graphic description of the spectral–temporal development of agricultural crops as seen by Landsat. *Proceedings of the symposium on machine processing of remotely sensed data* (pp. 41–51). West Lafayette, IN: Purdue University.

Kendall, M. G., & Buckland, W. R. (1982). *A dictionary of statistical terms*, 4th edition. London: Longman Group.

Lachenbruch, P. A., & Mickey, M. R. (1986). Estimation of error rates in discriminant analysis. *Technometrics*, *10*, 1–11.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science*, *51*, 109–119.

LeMay, V. M., Maedel, J., & Coops, N. C. (2008). Estimating stand structural details using nearest neighbour analyses to link ground data, forest cover maps, and Landsat imagery. *Remote Sensing of Environment*, *112*, 2578–2591.

Magnussen, S., McRoberts, R. E., & Tomppo, E. (2009). Model based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sensing of Environment*, *113*, 476–488 (this issue).

McRoberts, R. E. (2006). A model-based approach to estimating forest area. *Remote Sensing of Environment*, *103*, 56–66.

McRoberts, R. E., Bechtold, W. A., Patterson, P. L., Scott, C. T., & Reams, G. A. (2005). The enhanced Forest Inventory and Analysis Program of the USDA Forest Service: historical perspective and announcement of statistical documentation. *Journal of Forestry*, *103*(6), 304–308.

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*, 457–468.

McRoberts, R. E., Tomppo, E. O., Finley, A. O., & Heikkinen, J. (2007). Variance estimation for the k-Nearest Neighbors technique. *Remote Sensing of Environment*, *111*, 466–480.

McRoberts, R. E., Winter, S., Chirici, G., Hauk, E., Pelz, D. R., & Moser, W. K. (2008). Large-scale patterns of forest structural diversity. *Canadian Journal of Forest Research*, *38*, 429–438.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science*, *41*, 337–359.

Nilsson, M., Holm, S., Reese, H., Wallerman, J., & Engberg, J. (2005). Improved forest statistics from the Swedish National Forest Inventory by combining field data and optical satellite data using post-stratification. *Proceedings of ForestSat 2005, 31 May–03 June 2005, Borås, Sweden* (pp. 22–26). Skogsstyrelsen, National Board of Forestry Rapport 8a.

Ohmann, J. L., & Gregory, M. J. (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.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the great plains with ERTS. *Proceedings of the Third ERTS Symposium, NASA SP-351, Volume 1.* (pp. 309–317)Washington, DC: NASA.

Temesgen, H., LeMay, V. M., Froese, P. L., & Marshall, P. L. (2002). Imputing tree-lists from aerial attributes for complex stands of southeastern British Columbia. *Forest Ecology and Management*, *177*, 277–285.

Thessler, S., Ruokolainen, K., Tuomisto, H., & Tomppo, E. (2005). Mapping gradual landscape-scale floristic changes in Amazonian primary rain forests by combining ordination and remote sensing. *Global Ecology and Biogeography*, *14*, 315–325.

Tomppo, E. (1991). Satellite image based national forest inventory of Finland. *International Archives of Photogrammetry and Remote Sensing*, *28*(7-1), 419–424.

Tomppo, E. (1996). Multi-source National Forest Inventory of Finland. In R. Vanclay, J. Vanclay, S. Miina (Eds.), *New Thrusts in Forest Inventory. Proceedings of the Subject Group S4.02-00 'Forest Resource Inventory and Monitoring' and Subject Group S4.12-00 'Remote Sensing Technology', vol 1. IUFRO XX World Congress 6–12 Aug. 1995* (pp. 27–41) Tampere, Finland. EFI proceedings, 7, European Forest Instutie. Joensuu, Finland.

Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote Sensing of Environment*, *82*, 156–171.

Trotter, C. M., Dymond, J. R., & Goulding, C. J. (1997). Estimation of timber volume in a coniferous plantation forest using Landsat TM. *International Journal of Remote Sensing*, *18*, 2209–2223.

Yang, L., Homer, C. G., Hegge, K., Huang, C., Wylie, B., & Reed, B. (2001). A Landsat 7 scene selection strategy for a National Land Cover Database. *Proceedings of the IEEE 2001 International Geoscience and Remote Sensing Symposium, Sydney, Australia, CD-ROM, 1 disc*.