Remote Sensing of Environment 163 (2015) 13–22

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data

Ronald E. McRoberts [a,*], Erik Næsset [b], Terje Gobakken [b]

[a] Northern Research Station, U.S. Forest Service Saint Paul, MN, USA
[b] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 28 January 2015
Accepted 24 February 2015
Available online 26 March 2015

Keywords:
Distance metric
Precision

ABSTRACT

Nearest neighbors techniques calculate predictions as linear combinations of observations for a selected number of population units in a sample that are most similar, or nearest, in a space of auxiliary variables to the population unit requiring the prediction. Nearest neighbors techniques have been shown to be particularly effective when used with forest inventory and remotely sensed data. Recent attention has focused on developing an underlying foundation consisting of diagnostic tools, inferential extensions, and techniques for optimization.

For a study area in Norway, forest inventory and airborne laser scanning data were used with the k-Nearest Neighbors technique to estimate mean aboveground biomass per unit area. Optimization entailed reduction of the dimension of feature space, deletion of influential outliers, and selection of optimal weights for the weighted Euclidean distance metric. These optimization steps increased the proportion of variability explained in the reference set by as much as 20%, reduced confidence interval widths by as much as 35%, and produced standard errors that were as small as 3% of the estimate of the mean.

Published by Elsevier Inc.

1. Introduction

Nearest neighbors techniques are multivariate, non-parametric approaches to estimation. Population unit predictions are calculated as linear combinations of sample observations for units that are nearest or most similar in a space of auxiliary variables to units for which predictions are desired. Nearest neighbors techniques have received considerable attention for areal estimation and mapping of forest attributes, particularly when used with forest inventory and remotely sensed data. Eskelson et al. (2009) reviewed applications that impute missing values in forestry databases, and McRoberts (2012) reviewed the variety of scientific applications and the current state of the art for continuous response variables, and provided a catalog of international forestry applications.

Implementation of nearest neighbors techniques requires choices for three parameters: (1) a value for k, the number of nearest neighbors, (2) a scheme for weighting neighbors when calculating predictions, and (3) a distance or similarity metric. The choices are often guided by assessments of results obtained for various combinations of parameters which, in turn, rely on diagnostics related to the quality of predictions, analysis of residuals, extrapolations, inferences, and ease of implementation. The term *k-Nearest Neighbors* (k-NN) is generic and is used to refer to nearest neighbors techniques regardless of the number of neighbors, weighting scheme, or distance metric. Variations such as *Most Similar Neighbor* (MSN) (Moeur & Stage, 1995) and *Gradient Nearest Neighbor* (GNN) (Ohmann & Gregory, 2002) are simply k-NN using different distance metrics.

The popularity of k-NN for use with forest inventory and remotely sensed data motivated a workshop on k-NN applications under the auspices of COST Action FP1001. COST (European Cooperation on Science and Technology) is a European framework for promoting and facilitating scientific cooperation among European scientists and researchers (COST, 2014). COST Action FP1001 focused on European approaches for using multi-source national forest inventory (NFI) data to improve information on the potential supply of wood resources (COST FP1001, 2014). Within the Action, Working Group 1 focused on harmonizing NFI estimates obtained using different sampling designs and estimation techniques; Working Group 2 focused on improving estimates of wood resources by combining NFI and remotely sensed data; and Working Group 3 focused on inventory volume and consumption information. Working Group 2 conducted a comprehensive review of the k-NN literature and summarized the results with respect to multiple features including response variables, auxiliary variables, distance metrics, accuracy measures, and most cited publications. In addition, a tutorial in the form of a series of lectures on the use and optimization of nearest neighbors techniques for forest inventory applications was presented at a Task Force meeting of Working Group 2 in Dublin, Ireland, 17–19 September 2013.

The overall objectives of the study were to document and articulate further the principles and techniques presented in the Task Force lectures using an empirical example. The context for the example was

* Corresponding author at: Northern Research Station, U.S. Forest Service, 1992 Folwell Avenue, Saint Paul, Minnesota 55108, USA. Tel.: 651 649-5174.
E-mail address: rmcroberts@fs.fed.us (R.E. McRoberts).

http://dx.doi.org/10.1016/j.rse.2015.02.026
0034-4257/Published by Elsevier Inc.

estimation of mean aboveground biomass (AGB) per unit area as the response variable using the k-NN technique with forest inventory and airborne laser scanning (ALS) data. The specific technical objectives focused on techniques for optimizing the k-NN technique for the purpose of increasing the accuracy and precision of the estimate of mean AGB per unit area.

2. Data

The study area was in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway (Fig. 1), and was defined as the geographic area represented by the 1259-km^2 portion of the Latin Square sampling design used by the Norwegian NFI that was inventoried between 2005 and 2007 (Tomter, Hylen & Nilsen, 2010). The dominant tree species in the study area are Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.). Field measurements were acquired for 145 circular, 250-m^2, Norwegian NFI field plots measured between 2005 and 2007. On each plot, all trees with a diameter at-breast-height (dbh, 1.3 m) of at least 5 cm were callipered. The biomass of each sample tree was estimated using species-specific statistical models with species, dbh, and height as independent variables (Braastad, 1966; Brantseg, 1967; Vestjordet, 1967). Plot-level AGB was estimated as the sum of biomass estimates for individual trees on the plot, scaled to a per unit area basis, and considered to be an observation without error (McRoberts & Westfall, 2014). A variogram analysis indicated no meaningful spatial correlation among AGB observations for different plots. Differential Global Navigation Satellite Systems (GPS and Russian GLONASS) were used to determine the locations of plot centers to sub-meter accuracy.

ALS data acquired between 15 July 2006 and 12 September 2006 from a height of approximately 1700 m and with an average aircraft speed of 75 m/s produced an average density of 0.7 pulses/m^2. The study area was tessellated into 250-m^2 squares that served as population units and that were of the same area as the plots. For each plot and population unit, height distributions were estimated for first echoes with heights greater than 2 m, and two sets of ALS metrics were calculated (Gobakken & Næsset, 2008). Heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions were denoted h_{10}, h_{20}, …, h_{100}, respectively. Canopy densities were calculated as the proportions of echoes with heights greater than 0%, 10%, …, 90% of the range between 2 m above the ground and the 95th percentile height and denoted d_0, d_{10}, …, d_{90}, respectively.

3. Nearest neighbors techniques

3.1. Terminology and notation

For notational purposes, Y commonly denotes a possibly multivariate vector of response variables, and **X** denotes a vector of auxiliary variables with observations for all population units. In the terminology of nearest neighbors techniques, the auxiliary variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the set of sample population units for which observations of both response and feature variables are available is designated the *reference set* and is denoted R with size denoted n; and the set of population units for which predictions of response variables are desired is designated the *target set* and is denoted U with size denoted N. All population units for both the reference and the target set are assumed to have a complete set of observations for all feature variables.

For continuous response variables such as AGB, the nearest neighbors prediction, $\hat{y}_i$, for the ith target unit is calculated as,

$$\hat{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i \qquad (1)$$

where $\{y_j^i, j = 1, 2, \ldots, k\}$ is the set of response variable observations for the k reference set units that are most similar or nearest to the ith target

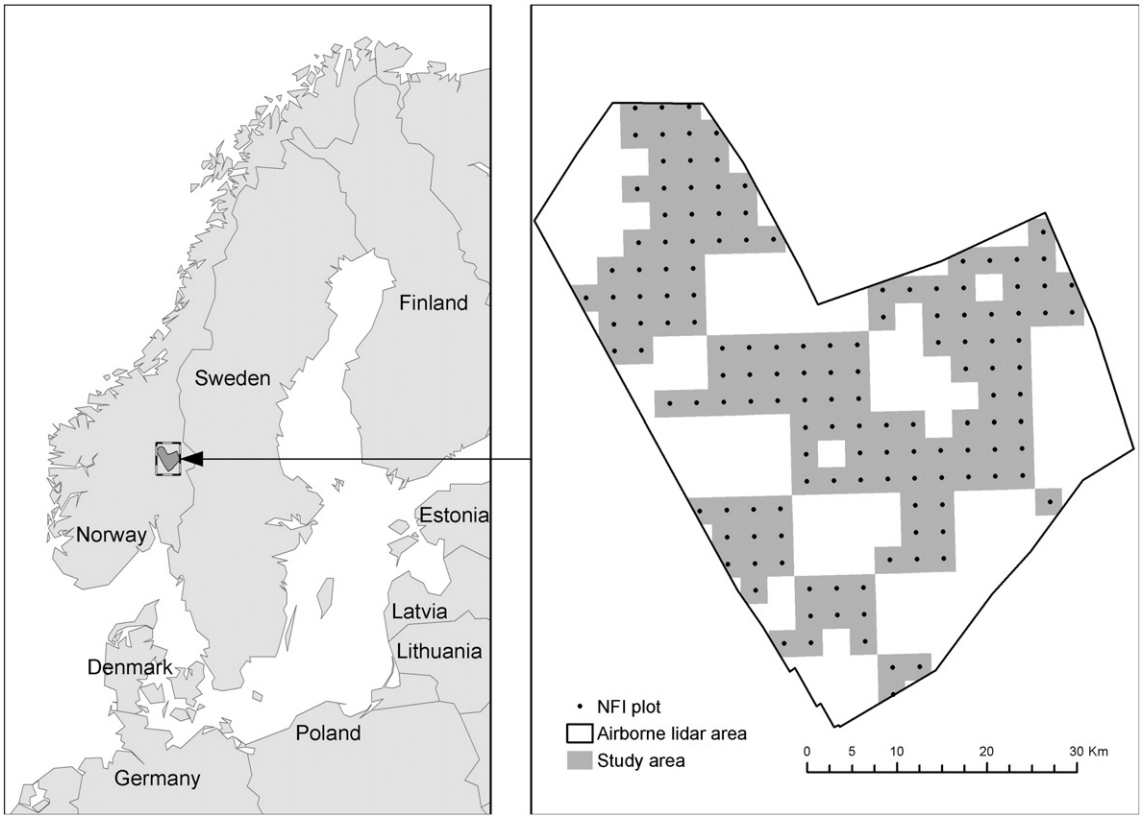

Fig. 1. Study area in Hedmark County, Norway.

unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the jth nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1$. For categorical variables such as forest/non-forest and forest type, the predicted class of the ith target unit is the most heavily weighted class among the k nearest neighbors, a weighted median or mode in case of ordinal scale variables, and a mode in the case of nominal variables.

3.2. Optimization

3.2.1. Leave-one-out analyses

A common approach for optimizing and assessing k-NN results in the reference set is through use of the *leave-one-out* technique (Lachenbruch & Mickey, 1986). With this technique, each reference set unit is deleted in sequence and predicted using the remaining reference set units. The leave-one-out technique shares features with the jackknife technique and is a nearly unbiased error estimator (Elisseef & Pontil, 2002). Common metrics for assessing results with the leave-one-out technique are based on the sum of squared differences between reference set observations and predictions for continuous response variables or overall accuracy for categorical response variables.

3.2.2. Number of neighbors, k

The value of k may be selected to optimize multiple criteria either individually or in combination. For k-NN variations that permit $k > 1$, smaller values of k are generally preferred as a means of reducing complexity and computational intensity. However, caution must be exercised when selecting small values of k because such values may yield root mean square errors that are greater than the standard deviations of the response variable observations, meaning that the overall mean as a prediction for every target unit would better maximize accuracy than the k-NN predictions.

The most intuitive approach is to select the value of k that optimizes a criterion, C (e.g., prediction error, class accuracy), in the reference set as assessed using the leave-one-out technique. When the relationship between the response and feature variables is weak and when the reference sets are large, the value of k that optimizes C is often large. However, the C versus k curve is often relatively flat in the vicinity of the optimal value of k so that a smaller value may be selected with no appreciable adverse impact on C. McRoberts, Nelson and Wendt (2002) proposed selecting the smallest value of k for which C does not differ by more than a pre-selected percentage (1% or 5%) from the optimal value of C.

3.2.3. Neighbor weighting

The most common approach to weighting neighbors in the calculation of predictions is to weight neighbors inversely proportional to a power of the distance, d_{ij}, between the jth reference unit and the ith target unit,

$$w_{ij} \propto d_{ij}^{t},$$

where $t \leq 0$. Commonly, albeit arbitrarily, selected values are often $t = 0$, $t = -1$, or $t = -2$. Other than McRoberts (2012), no reports of attempts to optimize the selection of t are known.

3.2.4. Distance metrics

Many familiar distance metrics can be expressed in matrix form as,

$$d_{ij} = \left[\left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{M} \left(\mathbf{X}_i - \mathbf{X}_j\right)\right]^{1/2}, \quad (2)$$

where i denotes a target unit for which a prediction is desired, j denotes a reference unit, $\mathbf{X}_i$ and $\mathbf{X}_j$ are vectors of observations of feature variables, and **M** is a square, positive definite matrix. When **M** is the identity matrix, Euclidean distance results; when **M** is a non-identity diagonal matrix, weighted Euclidean distance results; and when **M** is the inverse of the covariance matrix of the feature variables, Mahalanobis distance results. Metrics based on canonical correlation and canonical correspondence can also be expressed in matrix form (LeMay & Temesgen, 2005; Ohmann, Gregory & Roberts, 2014).

When considering nearest neighbors distance metrics, two feature space properties are particularly relevant. The first property, characterized by Bellman (1961) as the *curse of dimensionality*, is that the multi-dimensional size of feature space increases exponentially as the number of feature variables increases linearly. Important detrimental consequences follow as the dimension of feature space increases: (i) nearest neighbors are at greater distances from target units (Schaal, Vijayakumar & Atkeson, 1998); (ii) the distance to the nearest neighbor approaches the distance to the farthest neighbor (Beyer, Goldstein, Ramakrishnan & Shaft, 1998) which mitigates the beneficial effects of distance-based neighbor weighting; and (iii) extrapolations beyond the ranges of the feature variables in the reference set are more probable (McRoberts, 2009).

The second property is that inclusion of feature variables that are unrelated to the response variables has detrimental effects. Langley and Iba (1993) and Blum and Langley (1997) characterize such feature variables as *irrelevant*. Unlike regression analyses for which irrelevant variables usually have little effect on predictions because their corresponding coefficient estimates are often not statistically significantly different from zero, irrelevant feature variables introduce randomness into nearest neighbor distances, thereby introducing randomness into the selection of nearest neighbors which, in turn, results in selection of spurious neighbors and less accurate predictions. Langley and Iba (1993) showed that, on average, the size of the reference set necessary to obtain a specified level of accuracy grows very fast with increases in the number of irrelevant feature variables. Thus, optimization of nearest neighbors distance metrics must consider reducing the number of feature variables, particularly by eliminating irrelevant feature variables.

The two primary approaches to reducing the adverse effects of the curse of dimensionality and irrelevant feature variables are reduction of the number of feature variables and weighting feature variables in proportion to their relevance in the distance metric. Dimension reduction techniques, also characterized as feature selection, are of two kinds: transformations based on techniques such as principal components and canonical correlation analysis and selection of a subset of feature variables that optimizes a selected criterion. Genetic algorithms (GAs) (Holland, 1975) have been used both to select subsets of feature variables and to weight them. GAs are adaptive heuristic search algorithms that simulate evolutionary processes in natural selection by intelligently exploiting random searches within defined search spaces. For nearest neighbors applications, GAs start with a randomly selected set of values for the elements of **M** and then use genetic operators such as inheritance, selection, mutation, and crossover to find values that optimize a selected criterion. Tomppo and Halme (2004), McRoberts (2008), Tomppo, Gagliano, De Natale, Katila and McRoberts (2009), and Holopainen et al. (2010) all illustrate the use of GAs to determine the optimal elements for weighted Euclidean distance represented by diagonal matrices, **M**.

3.3. Inference

Prediction techniques such as k-NN are often used only to describe relationships among variables and to construct maps. Although maps, and the prediction techniques on which they depend, may be useful for depicting an estimate of the spatial distribution of a resource, maps make no statements regarding parameters related to the resource. Further, map accuracy measures convey no direct information regarding the precision or accuracy of the estimates of parameters obtained by aggregating map predictions (McRoberts, 2011). For inventory purposes, the ultimate objective is an inference for a population parameter. The Oxford English Dictionary defines *inference* as "to accept from evidence

or premises" (Simpson & Weiner, 1989). In a sampling framework, inference requires expression of the relationship between a population parameter, μ, and its estimate, $\hat{\mu}$ in probabilistic terms (Dawid, 1983). For estimation purposes, these expressions take the form of confidence intervals,

$$\hat{\mu} \pm t_{1-\alpha} \cdot \sqrt{\hat{\mathrm{Var}}(\hat{\mu})}, \tag{3}$$

where $1 - \alpha$ denotes the probability that confidence intervals constructed in this manner include the true value of the parameter. Two approaches to inference are relevant for k-NN applications, probability- or design-based inference and model-based inference.

3.3.1. Probability-based inference

Probability-based inference, also characterized as design-based inference, is based on three assumptions: (1) population units are selected for a sample using a probability-based randomization scheme, (2) the probability of selection for each population unit is positive and known, and (3) the observation of the response variable for each population unit is a constant or fixed value. Estimators are derived to correspond to sampling designs and are typically unbiased, meaning that the expectation of estimates over all possible samples that could be obtained with the sampling design is the true value of the population parameter. However, the estimate obtained with any particular sample could still deviate substantially from the true value.

Hansen, Madow and Tepping (1983) apparently coined the term *probability-based* as an alternative to the more familiar term *design-based*. Because the basis for inference is not just a *design* for sampling, but rather a probability-based design, the term *probability-based* is considered by some to better characterize the basis for inference.

Probability-based, model-assisted estimators use models based on auxiliary data to enhance inferences but rely on probability samples for validity. Baffetta, Fattorini, Franeschi and Corona (2009) and Baffetta, Corona and Fattorini (2011) illustrate probability-based, model-assisted inference using the k-NN technique. For this study, the model-assisted regression estimators described by Särndal, Swensson, and Wretman (1992, Section 6.5) were used,

$$\hat{\mu} = \frac{1}{N}\sum_{i \in U} \hat{y}_i - \frac{1}{n}\sum_{i \in R}(\hat{y}_i - y_i) \tag{4}$$

and

$$\hat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{n(n-1)}\sum_{i \in R}(\hat{y}_i - y_i)^2. \tag{5}$$

The first term of Eq. (4) is simply the mean of k-NN predictions over all population units, and the second term is an estimate of bias which, when subtracted, compensates for systematic prediction error.

3.3.2. Model-based inference

With model-based inference, as with probability-based, model-assisted inference, a model is used to calculate predictions for population units. However, model-based inference is based on a different set of assumptions: (1) for each population unit, the observation of the response variable is a random realization from a distribution of possible values, rather than a constant value as is the case for probability-based inference, and (2) randomization enters through the realization of observations rather than from selection of population units into a sample. Validity for model-based inference resides in correct model specification, not a probability sample. Thus, an important advantage of model-based inference is that probability samples are not necessary, although they may be preferable. This advantage makes model-based inference particularly appealing for small areas and for remote inaccessible areas for which sufficiently large probability samples are not possible. In addition, with model-based inference, variances can be estimated for all population unit predictions which is not the case for probability-based inference. A disadvantage is that no simple compensation for systematic model prediction error is included in the estimator. Thus, careful assessment of correct model-specification is crucial.

With model-based inference, the mean and standard deviation of Y for the ith population unit are denoted μ_i and σ_i, and an observation is denoted,

$$y_i = \mu_i + \varepsilon_i, \tag{6}$$

where ε_i is the deviation of the observation from the mean of its distribution. The estimator of μ_i is $\hat{\mu}_i = \hat{y}_i$ from Eq. (1), and the estimator of the population mean is,

$$\hat{\mu} = \frac{1}{N}\sum_{i \in U} \hat{\mu}_i. \tag{7}$$

McRoberts, Tomppo, Finley and Heikkinen (2007) derived parametric estimators for σ_i^2 and $\mathrm{Var}(\hat{\mu})$ that accommodate spatial correlation among reference set observations. In the absence of spatial correlation, as is the case for this study, and for $t = 0$,

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{k}\left(y_j^i - \hat{\mu}_i\right)^2}{k-1}. \tag{8}$$

The general form of a model-based estimator of the variance of $\hat{\mu}$ is,

$$\hat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{i \in U}\hat{\mathrm{Var}}(\hat{\mu}_i) + 2\sum_{i <}\sum_{j \in U}\hat{\mathrm{Cov}}\left(\hat{\mu}_i, \hat{\mu}_j\right)\right] \tag{9}$$

(McRoberts, 2006, 2009). The covariance between estimates of the ith and jth means, $\hat{\mu}_i$ and $\hat{\mu}_j$, may be estimated as,

$$\hat{\mathrm{Cov}}\left(\hat{\mu}_i, \hat{\mu}_j\right) \approx \frac{m_{ij}\hat{\sigma}_i\hat{\sigma}_j}{k^2}, \tag{10}$$

where

$$\hat{\mathrm{Var}}(\hat{\mu}_i) = \hat{\mathrm{Cov}}(\hat{\mu}_i, \hat{\mu}_i),$$

and m_{ij} is the number of nearest neighbors common to the predictions for both the ith and jth population units. If the ith and jth population unit predictions share no common nearest neighbors, $\hat{\mathrm{Cov}}\left(\hat{\mu}_i, \hat{\mu}_j\right) = 0$.

The parametric approach to model-based variance estimation is complex, and because Eq. (9) requires double summations over all population units, it is also computationally intensive. Resampling estimators such as the bootstrap are well-suited for complex and nonparametric applications. McRoberts, Magnussen, Tomppo and Chirici (2011) and McRoberts (2012) described a bootstrapping approach characterized as *bootstrapping pairs* that entails randomly drawing samples with replacement from the reference set using a sampling design that mimics the original sampling (Efron & Tibshirani, 1994). For each bootstrap sample, b, the k-NN technique is applied to predict the response variable for each population unit, and the estimate of the population mean is calculated as $\hat{\mu}_b^{\mathrm{Boot}}$ using Eq. (7). The overall bootstrap estimator of the population mean, μ, is,

$$\hat{\mu}^{\mathrm{Boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b^{\mathrm{Boot}}, \tag{11}$$

with variance estimator,

$$\hat{\mathrm{Var}}\left(\hat{\mu}^{\mathrm{Boot}}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\mu}_b^{\mathrm{Boot}} - \hat{\mu}^{\mathrm{Boot}}\right)^2, \quad (12)$$

where B is the total number of bootstrap samples. Of importance, B must be large enough that both $\hat{\mu}^{\mathrm{Boot}}$ and $\hat{\mathrm{Var}}\left(\hat{\mu}^{\mathrm{Boot}}\right)$ stabilize.

3.4. Diagnostics

McRoberts (2009) reported k-NN diagnostic tools for univariate response variables. Unlike prediction techniques such as regression, no k-NN predictions can be smaller or larger than the smallest or largest observations in the reference set. One consequence, particularly for large values of k, is that when using the leave-one-out technique in the reference set, predictions corresponding to the smallest observations are too large and predictions corresponding to the largest observations are too small. The degree to which this lack of fit phenomenon affects overall estimation should not be ignored. An appropriate diagnostic is a graph of observations versus predictions.

The lack of fit problem may be exacerbated when predictions must be extrapolated in the target set beyond the range of the feature variables in the reference set. McRoberts (2009, Section 6.4) illustrates several diagnostics to assess the degree to which extrapolations may be a problem.

Outliers are defined as observations that differ from other observations to such a degree that they raise questions as to whether they are from a different population, whether sampling is faulty, or whether excessive measurement error is involved (Kendall & Buckland, 1982). For k-NN techniques, the adverse effects of outliers may be more serious than for other prediction methods. For example, with regression, the prediction for a particular population unit is based on parameter estimates obtained using observations for the entire sample, whereas for k-NN, the prediction for a particular population unit is based on only a small subset of the sample. Thus, the effects of a single outlier may be substantial, particularly if the outlier is frequently selected as a neighbor and if k is small.

When spatial data are combined from multiple sources and multiple dates, possibilities for outliers increase substantially. For example, because of locational errors, ground plots and either image pixels or ALS echoes may not be correctly matched. In the case of image pixels for medium resolution imagery such as Landsat, the ground plots are typically not complete samples of the pixels. Also, ground plots may have been subject to considerable disturbance between the date the remotely sensed data were acquired and the plot measurement dates. In addition, plots with similar AGB observations but different age distributions, different vertical compositions, or different health conditions may have quite different feature variable signatures.

Reference set observations that cause substantial changes in predictions, regardless of their outlier status, are characterized as *influential observations* (Belsley, Kuh & Welsch, 1980). Outliers that are also influential observations and that are frequently used as neighbors are of particular concern because of their potential adverse effects. Useful diagnostics for identifying candidates for influential outliers include standardized residuals, sums of neighbor weights, and change in the accuracy criterion resulting from deleting particular reference observations (Belsley et al., 1980).

The effects of the curse of dimensionality and irrelevant feature variables can often be assessed using the same diagnostics. One approach entails calculating the mean over all predictions of distances to the nearest neighbor and the mean over all predictions of the ratio of the distance to the nearest neighbor and the distance to the farthest neighbor. If these diagnostics increase substantially as additional feature variables are added, the results can be attributed to too many and/or irrelevant feature variables. A somewhat unique feature of nearest neighbors techniques resulting from the effects of irrelevant feature variables is that as the number of feature variables increases, a criterion such as root mean square error will often initially decrease, reach an optimal level, and then increase. Reducing the dimension of feature space to the number of feature variables that optimizes a criterion is a form of feature selection that contributes to reducing the effects of the curse of dimensionality and irrelevant feature variables.

4. Analyses

4.1. Overview

The analyses as described in the following sections entailed multiple steps. First, for each number of feature variables, the particular combination with the greatest accuracy was selected. Second, potential influential outliers in the reference set were identified and some were deleted from the reference set. Third, using the revised reference set, the GA algorithm was used to select optimal or near-optimal weights for the diagonal of the weighted Euclidean distance metric. Fourth, the degree to which optimization of the distance metric in the reference set could be expected to be realized in the target set was evaluated. Finally, the population mean and its standard error were estimated using both model-assisted and model-based estimators. For the model-based estimators, both parametric and bootstrap approaches to estimating the standard error were used. Details for the operational implementation of these steps follow.

4.2. Initial analyses

The criterion used to assess the relationship between reference observations and their k-NN prediction was root mean square error (RMSE) calculated using the leave-one-out technique as,

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SS_{res}}}{n - n_{var}}}, \quad (13)$$

where

$$\mathrm{SS_{res}} = \sum_{i \in R}\left(y_i - \hat{y}_i\right)^2$$

and n_{var} is the number of feature variables. The first step in the analyses was to determine values of k and t with the unweighted Euclidean distance that minimized RMSE in the reference set using a leave-one-out analysis for all combinations of all numbers of feature variables. For each number of feature variables with $n_{var} \leq 10$, the particular combination with the smallest RMSE was selected for further analysis. A pseudo-R^2 was expressed as,

$$R^{2*} = \frac{\mathrm{SS_{mean}} - \mathrm{SS_{res}}}{\mathrm{SS_{mean}}}, \quad (14)$$

where

$$\mathrm{SS_{mean}} = \sum_{i \in R}\left(y_i - \bar{y}\right)^2$$

and

$$\bar{y} = \frac{1}{n}\sum_{i \in R} y_i.$$

The term *pseudo*-R^2 is used because the formal definition of R^2 requires a linear model with an intercept (Anderson-Sprecher, 1994).

4.3. Identifying influential outliers

Three diagnostics were used to identify influential outliers in the reference set. First, following the analyses using the unweighted Euclidean distance metric described in Section 4.2, residuals were calculated as $\varepsilon_i = y_i - \hat{y}_i$; the triples $(y_i, \hat{y}_i, \varepsilon_i)$ were ordered with respect to $\hat{y}_i$; and the triples were aggregated into ordered groups of sizes 10–15. For each group, g, the mean observation, $\bar{y}_g$; the mean prediction $\bar{\hat{y}}_g$; the mean residual, $\bar{\varepsilon}_g$, and standard deviation of the residuals s_g, were calculated. A model of the relationship between s_g and the corresponding mean prediction $\bar{\hat{y}}_g$ was formulated as,

$$s_g = \beta_1\left[1 - \exp\left(\beta_2 \cdot \bar{\hat{y}}_g\right)\right] + \varepsilon_g, \tag{15}$$

and fit to the data (Fig. 2). Standardized residuals were then calculated by dividing each residual by the prediction of its standard deviation calculated using the model of Eq. (15).

Second, for each number of feature variables, $n_{var} \leq 10$, the selected combination of feature variables that minimized RMSE with corresponding k and t was applied to the target set using the unweighted Euclidean distance metric. For purposes of comparing sums of weights for different values of k, the weights for each prediction were rescaled to sum to k. The sum of weights, W_i, for each reference unit over all target predictions was calculated and divided by $\tau = \frac{k \cdot N}{n_{ref}}$ which is the mean weight expected for each reference unit. Reference units with $\frac{W_i}{\tau} \geq 1$ are used more frequently than expected, whereas reference set units with $\frac{W_i}{\tau} \leq 1$ are used less frequently than expected.

Third, the effect on RMSE of sequentially deleting each reference set unit was calculated for each number of feature variables, $n_{var} \leq 10$, using the selected combination of feature variables and corresponding k and t. For each reference unit, the ratio of the residual sum of squares, SS_{res}, following deletion and the SS_{res} with no deletions was calculated; reference set units with ratios less than 1 were potential influential units. Finally, primary candidates for influential outliers were reference observations for which the SS_{res} ratio was small, the standardized residual was large, and $\frac{W_i}{\tau}$ was large.

4.4. Distance diagnostics

For the selected combinations of each number of feature variables, the mean distance to the nearest neighbor and the mean of the ratio of the distance to nearest neighbor and the distance to the farthest neighbor were calculated using both unweighted and weighted Euclidean distances.

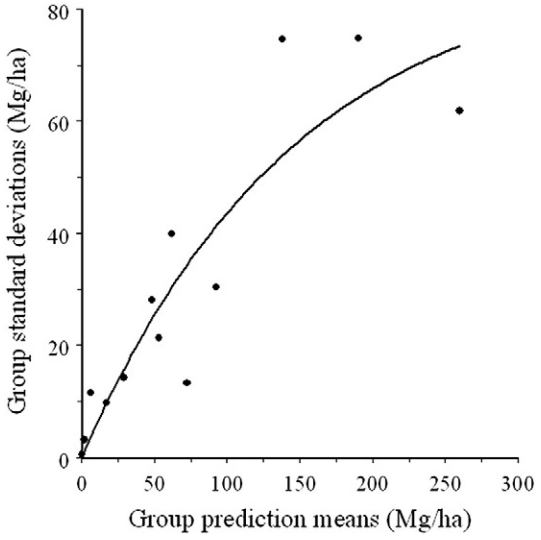

Fig. 2. For three feature variables, group standard deviations versus means of group predictions obtained using Eq. (15).

4.5. Optimizing the distance metric

A GA approach was used to select optimal values of k, t, and diagonal values for the distance matrix corresponding to the weighted Euclidean distance metric. Optimization of a full matrix in the reference set, as opposed to a diagonal matrix, would optimize a selected criterion to a greater degree than any other matrix-based distance metric including the canonical correlation analysis, canonical correspondence analysis, and Mahalanobis metrics. However, when more than a small number of feature variables is used, optimization of a full matrix is computationally very intensive. Thus, optimization of a weighted Euclidean distance metric represents a partial step toward full optimization.

However, optimization in the reference set does not guarantee optimization in the target set. In fact, if optimization in the reference set results in the equivalent of regression overfitting, the result may be detrimental to target unit predictions. The degree to which optimization in the reference set produced optimization in the target set was assessed using a modified cross validation approach. First, the reference set was randomly divided into test reference and test target sets of equal size. The values of k and t that produced the smallest RMSE for the unweighted Euclidean distance metric in the test reference set were determined and then applied to the test target set. Residual sums of squares for both the test reference and the target sets were calculated. The GA was then applied to the test reference set to select optimal values of k, t, and diagonal values for the distance metric which were then applied to the test target set. SS_{res} corresponding to the weighted Euclidean distance metric was then calculated for both the test reference and test target sets. For each replication, differences between the unweighted and weighted R^{2*} in the test reference set and between the unweighted and weighted R^{2*} in the test target set were calculated, and then the ratio of the two differences was calculated. This diagnostic characterizes the proportional gain in R^{2*} realized in the target set as a result of optimizing the weighted Euclidean distance metric in the reference set.

If the ratio is negative, the effects of optimization are actually detrimental in the target set; such a result can be attributed to overfitting in the reference set and is most likely to occur when the relationship between the response and feature variables is weak and/or when the effects of irrelevant feature variables are severe. If the ratio is positive, at least some gain resulting from optimization in the reference set is realized in the target set, and if the ratio is close to 1.0, then gain realized in the target set is comparable to the gain achieved in the reference set. For this study, the 10th, 50th, and 90th percentiles of the distribution of this ratio over 250 replications were calculated.

4.6. Inference

For combinations of selected numbers of feature variables, the model-assisted and the model-based estimators were used to estimate the population mean and its standard error. For the model-based approach, both parametric and bootstrap approaches to estimating standard errors were used.

5. Results and discussion

5.1. Analyses using the unweighted Euclidean distance metric

When using the unweighted Euclidean distance metric, the smallest values of RMSE corresponded to the largest values of R^{2*} which were in the range 0.72 to 0.75 for $3 \leq n_{var} \leq 10$ (Table 1). These R^{2*} values are comparable to or greater than those reported by Kankare et al. (2013) and those derived from statistics reported by Holopainen et al. (2010) and Breidenbach, Næsset and Gobakken (2012).

Table 1
Optimization results for variable combination with smallest RMSE for each number of feature variables.

No. of feature variables (n_{var})	Unweighted Complete reference set				Unweighted 2 reference observations deleted				Weighted 2 reference observations deleted			
	k	t	RMSE	R^{2*}	k	t	RMSE	R^{2*}	k	t	RMSE	R^{2*}
1	11	−0.19	48.6297	0.706	8	−0.42	42.7567	0.776	–	–	–	–
2	6	0.00	47.3556	0.723	6	−0.50	37.7536	0.826	4	0.00	30.0681	0.890
3	6	0.00	46.1881	0.739	4	−0.30	36.2797	0.841	5	−0.04	29.5586	0.894
4	6	0.00	46.2970	0.739	4	−0.31	36.2915	0.842	5	0.00	29.3615	0.896
5	6	0.00	45.8659	0.746	4	−0.24	36.1075	0.844	5	0.00	29.4977	0.896
6	6	0.00	45.9960	0.746	5	−0.49	35.7004	0.849	5	0.00	29.8828	0.894
7	6	0.00	45.7941	0.750	5	−0.50	35.8449	0.849	5	0.00	29.5793	0.897
8	6	0.00	45.9492	0.751	4	−0.31	35.9485	0.849	5	0.00	30.0481	0.895
9	6	0.00	46.1179	0.751	4	−0.28	36.0727	0.849	5	0.00	30.1505	0.895
10	6	0.00	46.2883	0.751	4	−0.29	36.2320	0.849	5	0.00	30.2668	0.895

For each number of feature variables, $2 \leq n_{var} \leq 10$, reference units were ranked with respect to the SS_{res} ratio as described in Section 4.3, and both the standardized residual and $\frac{W_i}{\bar{\tau}}$, also described in Section 4.3, were noted. Reference units with SS_{res} ratios less than 0.9, standardized residuals greater than 2, and $\frac{W_i}{\bar{\tau}} > 1$ were of particular interest. The 0.9 threshold for the SS_{res} ratio is arbitrary, but the 2.0 threshold for the standardized residuals is approximately the 95th percentile of the assumed Gaussian residual distribution, and as previously noted, $\frac{W_i}{\bar{\tau}} > 1$ indicates a reference unit that is used as a neighbor more frequently than the average.

The results were similar for all numbers of feature variables with reference units 69, 55, and 34 all identified as influential outliers (Table 2). Reference units 69 and 34 were initially deleted from the reference set, but reference unit 55 was initially retained because it was one of only a few reference units with large AGB observations and because its frequency of use as a neighbor was relatively small. The three metrics were then recalculated, including refitting the model of Eq. (15), and the assessment was repeated. The result was that the candidacy of reference unit 55 as an influential outlier was greatly diminished. Therefore, for further analyses, only reference units 69 and 34 were permanently deleted from the reference set.

Table 2
Candidates for influential outliers.

No. feature variables	Reference unit	SS_{res} ratio[a]	Standardized residual	Weight ratio[b]	AGB observation
2	69	0.8327	2.1256	1.0385	32.68
	55	0.8486	2.4672	0.5356	406.68
	34	0.8904	1.8863	1.9604	45.24
	13	0.9553	1.9581	1.1359	238.80
3	69	0.8229	2.2667	1.0356	32.68
	55	0.8397	2.5287	0.5356	406.68
	34	0.8903	1.9644	1.8686	45.24
	108	0.9422	1.6727	0.8877	38.60
4	69	0.8224	2.3444	1.0338	32.68
	55	0.8393	2.5593	0.5353	406.68
	34	0.8961	2.0058	1.8472	45.24
	108	0.9450	1.5904	0.9242	38.60
5	69	0.8150	2.3509	1.0319	32.68
	55	0.8351	2.5956	0.5369	406.68
	34	0.8934	2.0249	1.7915	45.24
	13	0.9351	2.1324	1.4526	238.80
6	69	0.8147	2.3810	1.0294	32.68
	55	0.8349	2.5957	0.5382	406.68
	34	0.8932	2.0354	1.7723	45.24
	13	0.9388	2.0959	1.4611	238.80
7	69	0.8117	2.1911	1.0285	32.68
	55	0.8332	2.5561	0.5386	406.68
	34	0.8915	1.9504	1.7684	45.24
	13	0.9500	2.2577	1.4687	238.80
8	69	0.8116	2.1528	1.0269	32.68
	55	0.8321	2.5365	0.5388	406.68
	34	0.8914	1.9278	1.7235	45.24
	13	0.9548	2.2690	1.4979	238.80

[a] Ratio of SS_{res} when observation is deleted from reference set to SS_{res} when reference observation is included in reference set.
[b] Ratio of sum of weights over all predictions to mean sum of weights over entire reference set ($\frac{W_i}{\bar{\tau}}$ from Section 4.2).

The selection of combinations of feature variables and values of k and t with the unweighted Euclidean distance metric that minimized RMSE for each number of feature variables, $2 \leq n_{var} \leq 10$, was repeated using the reference set with the two units deleted. Deletion of the two reference units produced substantially greater values of R^{2*} and corresponding smaller RMSEs (Table 1). In addition, values of k were slightly smaller and values of t were slightly larger in absolute value, both indicating that the selected neighbors were more similar to reference observations.

5.2. Analyses using the weighted Euclidean distance metric

For each number of feature variables, $2 \leq n_{var} \leq 10$, and using the combinations of feature variables selected following deletion of the influential outliers, the GA was used to select optimal or near optimal weights and corresponding values of k and t for use with the weighted Euclidean distance metric to minimize RMSE in the revised reference set (Table 2). The positive effects of using the weighted rather than unweighted Euclidean distance metric were to decrease RMSE by 34–36% and to increase R^{2*} by 19–23% (Table 1). Values of k were similar to those for the unweighted metric with outliers removed, but the values of t were uniformly 0. Large absolute values of t indicate that closer neighbors are weighted more heavily when calculating predictions and often suggest more accurate predictions. In this case, the smaller absolute values of t are attributed to compensating beneficial effects of values of t and the weighted Euclidean distance metric.

The feature variable weights for the weighted Euclidean distance metric are shown in Table 3. For small n_{var}, d_0 was consistently the most heavily weighted feature variable. Of interest, the optimization procedure selected only one height feature variable, and despite its quite small weight, it was selected for all numbers of feature variables.

The distance diagnostics confirmed that as the number of feature variables increased, distance to the nearest neighbor increased as did the ratio of the distance to the nearest neighbor and the distance to the farthest neighbor (Table 4). These diagnostics confirm the assertions of Schaal et al. (1998) and Beyer et al. (1998) as described in Section 3.2.4. However, the very small values of these diagnostics indicate that any adverse effects of the curse of dimensionality or irrelevant feature variables were negligible for $n_{var} \leq 10$.

The modified cross validation approach indicated that gains achieved in R^{2*} in the reference set as a result of using optimized Euclidean distance were also largely realized in the target set (Table 5). R^{2*} values for the test reference sets were comparable, albeit slightly smaller, than R^{2*} in the original reference set (Table 5, columns 2–4). Median gains achieved in R^{2*} in the test reference sets as a result of using the weighted

Table 3
Feature variable weights for weighted Euclidean distance metric.

Feature variable[a]	Number of feature variables									
	1	2	3	4	5	6	7	8	9	10
10	1.0000	0.0047	0.0021	0.0028	0.0028	0.0024	0.0022	0.0034	0.0024	0.0025
11	–	0.9953	0.7899	0.4341	0.4152	0.1184	0.2302	0.3634	0.2994	0.2343
12	–	–	0.2080	0.1324	0.1566	0.5203	0.1969	0.3020	0.2380	0.2459
13	–	–	–	–	0.0067	0.0325	0.0059	0.0339	0.0272	0.0100
14	–	–	–	–	–	–	–	0.0072	0.0124	0.0408
15	–	–	–	–	–	–	–	–	–	–
16	–	–	–	–	–	0.0261	0.0382	0.0506	0.0162	0.0551
17	–	–	–	–	–	0.3004	0.0820	0.0723	0.3087	0.0256
18	–	–	–	0.4307	0.4187	–	–	0.1671	0.0381	0.0579
19	–	–	–	–	–	–	–	–	0.0576	0.0676
20	–	–	–	–	–	–	0.4445	–	–	0.2602

[a] Feature variables: $1 = h_{10}$, $2 = h_{20}$, …, $10 = h_{100}$, $11 = d_0$, $12 = d_{10}$, …, $20 = d_{90}$ (Section 2).

Euclidean distance metric were in the range 0.06–0.09 which were also comparable but smaller than were achieved in the original reference set (Table 5, columns 5–7). The smaller values of R^{2*} and increase in R^{2*} are attributed to the test reference sets being smaller than the already rather small original reference set. The percentage gains in R^{2*} achieved in the test reference sets that were also realized in the test target sets depended on the number of feature variables. As the number of feature variables increased beyond $n_{var} = 3$, the distributions of percentage gains realized shifted to smaller values, meaning that less of the gains achieved in the test reference sets were realized in the test target sets. In fact, for $n_{var} \geq 4$, at least 10% of the percentage gains were negative, meaning that the weighted Euclidean distance metric caused R^{2*} to decrease rather than increase. This phenomenon is equivalent to overfitting a regression model and is attributed to optimizing on features of the reference sets that were not present in the target sets. Although the latter result may be unique to this dataset, it suggests an additional adverse effect of the curse of dimensionality.

Table 4
Distance diagnostics[a].

Number of feature variables	Mean $\frac{d_1}{d_{nref}}$	Mean distance to kth neighbor		Proportion of kth neighbors at same distance as 1st neighbor	
		k = 1	k = 5	k = 2	k = 5
Unweighted, complete reference set					
1	0.0005	0.0001	0.0001	0.9396	0.8503
2	0.0002	0.0752	0.8775	0.1134	0.1063
3	0.0003	0.0979	0.9311	0.1139	0.1063
4	0.0004	0.1100	0.9507	0.1140	0.1063
5	0.0004	0.1255	0.9946	0.1140	0.1063
6	0.0005	0.1364	1.0185	0.1140	0.1063
Unweighted, 2 reference observations deleted					
1	0.0001	0.0429	0.8098	0.1558	0.1063
2	0.0002	0.0760	0.8800	0.1147	0.1063
3	0.0003	0.0979	0.9311	0.1139	0.1063
4	0.0004	0.1061	0.9433	0.1139	0.1063
5	0,0004	0.1226	0.9829	0.1141	0,0163
6	0.0005	0.1379	1.0134	0.1144	0.0163
Weighted, 2 reference observations deleted					
1	0.0001	0.0429	0.8098	0.1558	0.1063
2	0.0032	0.0109	0.0344	0.1175	0.1063
3	0.0039	0.0106	0.0318	0.1179	0.1063
4	0.0046	0.0181	0.0500	0.1174	0.1063
5	0.0046	0.0227	0.0626	0.1174	0.1063
6	0.0048	0.0261	0.0727	0.1173	0.1063

[a] d_i denotes distance to the ith nearest neighbor.

5.3. Inferences

For $n_{var} \geq 2$, the estimates of the study area mean and its standard error differed very little (Table 6). For the model-assisted regression estimators, estimates of the mean ranged from 79.65 to 82.19 Mg/ha and estimates of the standard error ranged from 2.43 to 2.51 Mg/ha. As proportions of the estimated mean, the standard errors were approximately 0.03 which may be regarded as quite acceptable, particularly for a relatively small reference set.

Similarly, for $n_{var} \geq 2$, the model-based estimates of the mean ranged from only 80.66 to 81.80 Mg/ha, the parametric estimates of the standard error ranged from 3.03 to 3.36 Mg/ha, and the bootstrap estimates of the standard error ranged from 2.67 to 3.19 Mg/ha. Because the unbiasedness of model-based estimators depends on correct model specification, and because there is no bias correction term as with the model-assisted estimators, careful assessment of the correspondence between the k-NN predictions and their corresponding reference observations is essential with model-based estimators. Graphs of AGB observations versus their corresponding k-NN predictions indicated no serious lack of fit (Fig. 3); the results were similar for graphs of group means calculated as described in Section 4.3 (Fig. 4). In addition, bias estimates for the model-assisted estimators were small (Table 6).

Differences between the parametric and bootstrap standard error estimates merit comment. For $k \geq 25$, McRoberts et al. (2011) and McRoberts (2012) reported similar values for parametric and bootstrap estimates. However, Magnussen, McRoberts and Tomppo (2009) reported accurate parametric estimates for $k > 25$, but less accurate estimates for $k < 7$; the results of this study confirm the finding of Magnussen et al. (2009). As per Eq. (8), the parametric variance estimators for individual population units are based on differences among the k selected neighbors. For most statistical applications, sample sizes of at least five and preferably greater are recommended for estimating variances, but for this study, $k = 5$ was the optimal value for use with the weighted Euclidean distance metric and, therefore was the sample size used for calculating population unit variances. For a sample of individual population units, comparisons of parametric and bootstrap

Table 5
Optimization gains.

No. of feature variables	$R^{2*}_{ref\text{ - }unwtd}$ (Percentile)			$R^{2*}_{ref-wtd} - R^{2*}_{ref-unwtd}$ (Percentile)			$\frac{R^{2*}_{tgt-wtd} - R^{2*}_{tgt-unwtd}}{R^{2*}_{ref-wtd} - R^{2*}_{ref-unwtd}}$ (Percentile)		
	10th	50th	90th	10th	50th	90th	10th	50th	90th
1	–	–	–	–	–	–	–	–	–
2	0.735	0.791	0.846	0.049	0.088	0.140	0.125	0.832	2.700
3	0.761	0.819	0.855	0.038	0.074	0.113	0.023	0.674	2.306
4	0.759	0.814	0.857	0.037	0.077	0.120	−0.038	0.607	2.268
5	0.766	0.817	0.859	0.036	0.073	0.118	−0.055	0.574	2.137
6	0.771	0.818	0.861	0.031	0.068	0.109	−0.228	0.551	1.953

Table 6
Estimates of means and standard errors (SE).

No. of feature variables	Model-assisted			Model-based		
	Bîas	$\hat{\mu}_{MA}$	$SE(\hat{\mu}_{MA})$	$\hat{\mu}_{Mod}$	$SE(\hat{\mu}_{Mod})$	$SE(\hat{\mu}_{Boot})$
1	1.2267	88.7260	3.7281	89.9527	4.0385	3.7705
2	2.5999	79.1959	2.7335	81.7957	3.1010	3.1908
3	2.0578	78.5994	2.6375	80.6572	3.0307	2.6731
4	2.0702	79.4270	2.7912	81.4972	3.3571	2.8363
5	1.8669	79.6402	2.7955	81.5071	3.3521	2.7791
6	1.5289	79.5174	2.8677	81.0463	3.2459	2.9583

estimates of standard errors for individual population units indicated that the parametric estimates were usually considerably larger. Therefore, for this study with k = 5, the bootstrap standard error is regarded as more reliable. As a proportion of the estimated mean, the bootstrap standard error was also approximately 0.03.

For comparison purposes, both the model-assisted and model-based estimates of the mean were similar to those reported by McRoberts, Næsset and Gobakken (2013) who used a nonlinear regression model of the relationship between AGB and the ALS metrics for this same dataset. For this study, the simple random sampling, plot-based estimate of the mean was 74.75 Mg/ha with a standard error of 7.55 Mg/ha. The differences between both the model-assisted and the model-based estimates of the mean and the simple random sampling-based estimate are attributed to different distributions of the feature variables in the reference set relative to the target set. For example, for $n_{var} = 3$, d_0 is the dominant feature variable with proportional weight of 0.7899 (Table 3). The distribution of the values of d_0 in the reference set is much more heavily skewed toward small values than is the distribution in the target set. Finally, the much smaller model-assisted and model-based estimates of the standard error relative to the simple random sampling estimate attest to the utility of the auxiliary ALS data for improving the precision of the estimate of mean AGB.

Although the model-assisted and model-based estimates of the standard error were similar, caution must be exercised when formally comparing them because their estimators are based on such different underlying assumptions. In a practical sense, the model-assisted estimators are much easier to apply and typically produce acceptable estimates for sufficiently large sample sizes. In addition, whereas the model-assisted estimates of the standard error decrease as sizes of study areas and their corresponding sample sizes increase, such is not necessarily the case for model-based estimators (McRoberts, Næsset & Gobakken, 2013). However, for small or inaccessible areas for which only small probability samples or perhaps no probability samples are possible, model-based estimators may be the only alternative.

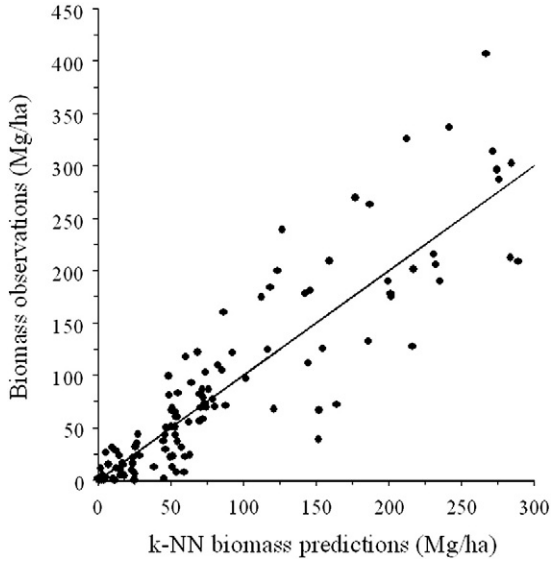

Fig. 3. Plot-level AGB observations versus k-NN predictions for three feature variables.

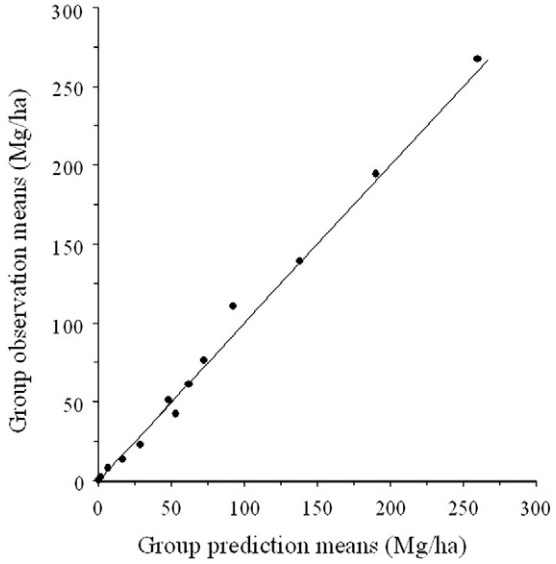

Fig. 4. Means of group AGB observations versus means of group k-NN AGB predictions for three feature variables.

In summary, k-NN accuracies obtained using the weighted Euclidean distance metric were essentially indistinguishable with respect to RMSE and R^{2*} for $2 \leq n_{var} \leq 4$ (Table 1). However, accuracy gains realized in the target set by optimizing the weighted Euclidean distance metric in the reference set were notably greater for $n_{var} = 3$ than for other values (Table 5). Both model-assisted and model-based bootstrap estimates of the standard error were either smallest or nearly smallest for $n_{var} = 3$ (Table 6). Although the estimate of the mean was slightly smaller for $n_{var} = 3$ than for other values of n_{var}, the differences both in absolute terms and with respect to the standard error estimates were small. Finally, smaller values of n_{var} contribute to minimizing the effects of the curse of dimensionality and irrelevant feature variables. Therefore, using $n_{var} = 3$, approximate 95% confidence intervals calculated as per Eq. (3) are 74.70–84.61 Mg/ha for the model-assisted estimators and 75.31–86.00 Mg/ha for the model-based estimators. The combined effects of deleting the two influential outliers and optimization of the weighted Euclidean distance metric were to reduce widths of two-standard error confidence intervals from 15.24 to 9.91 Mg/ha for the model-assisted estimators and from 16.31 to 9.69 Mg/ha for the model-based estimators.

6. Conclusions

Six conclusions may be drawn from the study. First, airborne laser scanning data have substantial utility for increasing the precision of forest AGB estimates, and second, the k-NN technique is an effective method for predicting forest AGB from the combination of forest inventory and airborne laser scanning data. Third, optimization of the Euclidean distance metric via the genetic algorithm increased the quality of k-NN predictions and the precision of estimate of the population mean. These three conclusions confirm previous findings, although studies using k-NN with airborne laser scanning data have not been extensively reported. Fourth, reduction of the dimension of feature space had no adverse effects on estimates. Investigations related to this issue have been incorporated in other studies, but mostly as peripheral factors. In particular, the effects of the curse of dimensionality and

irrelevant observations are not known to have been articulated or investigated previously.

The fifth and sixth conclusions are unique to this study. The fifth conclusion is that identification and deletion of influential outliers increased the accuracy of predictions and, subsequently, the precision of large area estimates. This issue has not been addressed in the forestry or remote sensing literature, particularly as it pertains to the use of the k-NN technique. The sixth conclusion is that a large proportion of gain achieved by optimization in the reference was realized in the target set, but more so with small numbers of feature variables. This issue has simply not been investigated previously, meaning that the conclusion is unique to this study.

Although broad generalizations are inappropriate based on only a single study, this study area and dataset are sufficiently generic that similar results for other should not be unexpected. Nevertheless, replication of the study for other study areas and other response and feature variables would contribute to further advancing the state of the science for k-NN applications.

Acknowledgments

The authors thank the following for careful reviews and insightful comments on the paper: Professor Gherardo Chirici, University of Molise, Italy; Drs. Daniel McInerney and Frank Barrett, Irish Forest Service; Dr. Ambrose Berger, Federal Research and Training Centre for Forests, Natural Hazards and Landscape, Austria; and Professor Erkki Tomppo, Finnish Forest Research Institute. Funding for the Task Force meeting at which much of the material in this article was first presented was provided by COST Action FP1001.

References

Anderson-Sprecher, R. (1994). Model comparisons and R^2. *The American Statistician*, *48*(2), 113–117.

Baffetta, F., Corona, P., & Fattorini, L. (2011). Design-based diagnostics for k-NN estimators of forest resources. *Canadian Journal of Forest Research*, *41*, 59–72.

Baffetta, F., Fattorini, L., Franeschi, S., & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, *113*(3), 463–475.

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton, New Jersey: Princeton University Press.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics.* New York: Wiley.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1998). When is "nearest neighbor" meaningful? In C. Beeri, & P. Buneman (Eds.), *Proceedings of the 7th International Conference on Database Theory (ICDT), January 10–12, 339 1999, Jerusalem, Israel* (pp. 217–235).

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*, 245–271.

Braastad, H. (1966). Volume tables for birch. *Meddelelser Norske SkogforsVes*, *21*, 265–365 (In Norwegian with English summary).

Brantseg, A. (1967). Volume functions and tables for Scots pine. South Norway. *Meddelelser Norske SkogforsVes*, *22*, 689–739 (In Norwegian with English summary).

Breidenbach, J., Næsset, E., & Gobakken, T. (2012). Improving k-nearest neighbor predictions in forest inventories by combining high and low density airborne laser scanning data. *Remote Sensing of Environment*, *117*, 358–365.

COST (2014). European Cooperation in Science and Technology. Available at: http://www.cost.eu/ (last accessed 22 January 2014)

COST FP1001 (2014). *USEWOOD*. Improving data and information on potential supply of wood resources, a European approach from multisource national forest inventories. Available at: https://sites.google.com/site/costactionfp1001/ (last accessed 22 January 2014)

Dawid, A. P. (1983). Statistical inference I. In S. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 4.* (pp. 89–105). New York: Wiley.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap.* Boca Raton, FL: Chapman and Hall/CRC.

Elisseef, A., & Pontil, M. (2002). Leave-one-out error and stability of learning algorithms with applications. *Advances in learning theory: Methods, models, and applications* (pp. 111–1130). NATO Advanced Study Institute on Learning Theory and Practice.

Eskelson, B. N. I., Temesgen, H., LeMay, V., Barrett, T. N., Crookston, N. L., & Hudak, A. T. (2009). The role of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, *24*(3), 235–246.

Gobakken, T., & Næsset, E. (2008). Assessing effects of laser point density, ground sampling intensity, and field plot sample size on biophysical stand properties derived from airborne laser scanner data. *Canadian Journal of Forest Research*, *38*, 1095–1109.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: Michigan. University of Michigan Press.

Holopainen, M., Haapanen, R., Karjalainen, M., Vastaranta, M., Hyyppä, J., Yu, X., et al. (2010). Comparing accuracy of airborne laser scanning and TerraSAR-X radar images in the estimation of plot-level forest variables. *Remote Sensing*, *2*, 432–445.

Kankare, V., Vastaranta, M., Holopainen, M., Räty, M., Yu, X., Hyyppä, J., et al. (2013). Retrieval of forest aboveground biomass and stem volume with airborne scanning LiDAR. *Remote Sensing*, *5*, 2257–2274.

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

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

Langley, P., & Iba, W. (1993). Average-case analysis of a nearest neighbor algorithm. *Proceedings of the International Joint Conference on Artificial Intelligence-93* (pp. 889–894). France: Chambery.

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*(2), 109–119.

Magnussen, S., McRoberts, R. E., & Tomppo, E. O. (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.

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

McRoberts, R. E. (2008). Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. *Remote Sensing of Environment*, *112*, 2212–2221.

McRoberts, R. E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment*, *113*, 489–499.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures. *Remote Sensing of Environment*, *115*, 715–724.

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management*, *272*, 3–12.

McRoberts, R. E., Magnussen, S., Tomppo, E. O., & Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. *Remote Sensing of Environment*, *115*, 3165–3174.

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment*, *128*, 268–275.

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). Estimating areal means and variances using the k-Nearest Neighbors technique and satellite imagery. *Remote Sensing of Environment*, *111*, 466–480.

McRoberts, R. E., & Westfall, J. A. (2014). Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. *Forest Science*, *60*(1), 34–42.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor — an improved sampling inference procedure for natural resource planning. *Forest Science*, *41*(2), 337–359.

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.

Ohmann, J. L., Gregory, M. J., & Roberts, H. M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. *Remote Sensing of Environment*, *151*, 3–15.

Särndal, C. -E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer.

Schaal, S., Vijayakumar, S., & Atkeson, C. G. (1998). Local dimension reduction. In M. I. Jordan, J. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 1–7). Cambridge, MA: MIT Press.

Simpson, J. A., & Weiner, E. S. C. (1989). *The Oxford English Dictionary* (2nd ed.). Oxford: Clarendon Press, 923–924 (Preparers).

Tomppo, E. O., Gagliano, C., De Natale, F., Katila, M., & McRoberts, R. E. (2009). Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. *Remote Sensing of Environment*, *113*, 500–517.

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of k-NN estimation: A genetic algorithm approach. *Remote Sensing of Environment*, *92*, 1–20.

Tomter, S. M., Hylen, G., & Nilsen, J. -E. (2010). Development of Norway's National Forest Inventory. In E. Tomppo, T. Gschwantner, M. Lawrence, & R. E. McRoberts (Eds.), *National Forest Inventories — Pathways for Common Reporting* (pp. 411–424). Springer.

Vestjordet, E. (1967). Functions and tables for volume of standing trees. Norway spruce. *Meddelelser Norske SkogforsVes*, *22*, 539–574 (In Norwegian with English summary).