Forestry Advance Access published September 9, 2016

Forestry An International Journal of Forest Research

Institute of Chartered Foresters

*Forestry* 2016; 1–13, doi:10.1093/forestry/cpw035

# Optimizing nearest neighbour configurations for airborne laser scanning-assisted estimation of forest volume and biomass

**Ronald E. McRoberts[1]*, Qi Chen[2], Grant M. Domke[1], Erik Næsset[3], Terje Gobakken[3], Gherardo Chirici[4] and Matteo Mura[4]**

[1]*Northern Research Station, U.S. Forest Service, 1992 Folwell Avenue, Saint Paul, MN 55038, USA*
[2]*Department of Geography, University of Hawai'i at Mānoa, 422 Saunders Hall, 2424 Maile Way, Honolulu, HI 96822, USA*
[3]*Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432, Ås, Norway*
[4]*Department of Agricultural, Food and Forestry Systems, Università degli Studi di Firenze, Via San Bonaventura, 13-50145, Firenze, Italy*

*Corresponding author. Tel: +1-651-649-5174; E-mail: rmcroberts@fs.fed.us

*Received 7 April 2016*

Inferences for forest-related spatial problems can be enhanced using remote sensing-based maps constructed with nearest neighbours techniques. The non-parametric *k*-nearest neighbours (*k*-NN) technique calculates predictions as linear combinations of observations for sample units that are nearest in a space of auxiliary variables to population units for which predictions are desired. Implementations of *k*-NN require four choices: a distance or similarity metric, the specific auxiliary variables to be used with the metric, the number of nearest neighbours, and a scheme for weighting the nearest neighbours. The study objective was to compare optimized *k*-NN configurations with respect to confidence intervals for airborne laser scanning-assisted estimates of mean volume or biomass per unit area for study areas in Norway, Italy, and the USA. Novel features of the study include a new neighbour weighting scheme, a statistically rigorous method for selecting feature variables, simultaneous optimization with respect to all four *k*-NN implementation choices and comparisons based on confidence intervals for population means. The primary conclusions were that optimization greatly increased the precision of estimates and that the results of optimization were similar for the *k*-NN configurations considered. Together, these two conclusions suggest that optimization itself is more important than the particular *k*-NN configuration that is optimized.

## Introduction

For forest-related spatial problems, the ultimate objective is often an inference in the form of a confidence interval for a population parameter. Prime examples are strategic forest resource assessments conducted by national forest inventories (NFI) that require confidence intervals for large suites of variables including most prominently forest area and growing stock volume (Tomppo *et al.*, 2010). Such inferences can be enhanced using spatial predictions in the form of remote sensing-based maps as auxiliary information. Nearest neighbours techniques are non-parametric, multivariate approaches to spatial prediction. Population unit predictions are calculated as linear combinations of observations for sample units that are nearest or most similar in a space of auxiliary variables to units for which predictions are desired. Nearest neighbours techniques have received considerable attention for mapping and areal estimation of forest attributes, particularly when used with forest inventory and remotely sensed data. Chirici *et al.* (2016b) documented more than 250 forestry applications for more than 25 countries on 6 continents.

McRoberts (2012) reviewed multiple factors that contribute to the appeal of nearest neighbours techniques including use with both continuous and categorical response variables, use for both univariate and multivariate prediction, and lack of distributional assumptions. In addition, recent advances have included both guidance and examples for the use of these techniques for inference rather than just prediction.

Most efforts to optimize nearest neighbour algorithms have focused on prediction accuracy and the distance metric. For predicting forest stand attributes using auxiliary variables obtained from aerial photography, LeMay and Temesgen (2005) reported that a metric based on canonical correlation analysis was superior with respect to prediction accuracy than the Euclidean and Manhattan distance metrics. For predicting forest attributes from Landsat-based variables, Chirici *et al.* (2008) reported that distance metrics giving greater weights to reference units whose response variable observations were closer to the mean of the observations were superior to Euclidean, Mahalanobis and two other metrics. Latifi *et al.* (2010) compared Euclidean, Mahalanobis, canonical correlation analysis and Random Forest

distance metrics for predicting forest volume and biomass using lidar, Landsat and aerial image data. The results were mixed with different metrics producing more optimal results for different response variables. Gagliasso *et al.* (2014) compared distance metrics based on canonical correlation analysis and canonical correspondence analysis (Ohmann and Gregory, 2002), and reported that the former metric produced smaller values of root mean square error when predicting biomass and basal area from lidar data. Packalén *et al.* (2012) reported the most comprehensive comparative analysis of *k*-nearest neighbours (*k*-NN) distance metrics. Their analyses included comparisons of results for distance metrics based on canonical correlation analysis, Random Forest and simulated annealing using all feature variables and, in addition, selected subsets of feature variables. Their primary conclusions were that the metric based on simulated annealing with a selected subset of feature variables produced the greatest accuracy and specifically that use of selected subsets produced greater accuracy than use of all feature variables. The only general conclusion that can be drawn from all these studies is that metrics that are optimized using observations of the response variable produce the most accurate predictions.

The number of nearest neighbours, *k*, is often arbitrarily selected as $k = 1$ or $k = 5$ but may also be selected to optimize a criterion such as root mean square error. Neighbours are often equally weighted although they are also often weighted inversely to the distances in feature space between units requiring predictions and sample units. The number of literature reports on optimizing *k* and neighbour weighting schemes is so small that no generalizations are possible.

Implementation of a nearest neighbours algorithm requires four choices: selection of a distance or similarity metric, selection of the particular auxiliary variables to be used with the metric, selection of the number of neighbours and selection of a scheme for weighting the neighbours. No reports are known of comprehensive efforts to optimize a nearest neighbours configuration by simultaneously considering all four choices. Furthermore, although the accuracies of predictions corresponding to sample observations are useful for optimization purposes, they are generally only an intermediate step enroute to an inference in the form of a confidence interval for a large area inventory parameter such as mean volume or biomass per unit area. Few authors other than Baffetta *et al.* (2009, 2011), McRoberts (2012) and McRoberts *et al.* (2002, 2007, 2015) have reported such inferences based on *k*-NN predictions. The primary study objective was to compare *k*-NN configurations consisting of combinations of the four *k*-NN choices with respect to the widths of confidence intervals for airborne laser scanning (ALS)-assisted estimates of mean forest volume or biomass per unit area for large area populations such as are reported by NFIs. Data were used for three study areas, one in Norway, one in Italy and one in the USA.

The novel features of the study are fourfold: (1) the paper introduces the Dudani neighbour weighting scheme, which has not previously been reported for forestry applications, (2) the paper introduces a statistically rigorous method for selecting feature variables, (3) the *k*-NN technique is optimized simultaneously with respect to all four choices and (4) *k*-NN configurations are statistically rigorously compared with respect to the ultimate estimation objective, an inferences in the form of confidence intervals for large area means rather than simply prediction accuracies.

## Data

### *Hedmark County, Norway*

The 1259-km$^2$ study area in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway, is dominated by Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.). Field measurements were acquired for 250-m$^2$ Norwegian NFI field plots located at the intersections of a 3-km × 3-km grid (Tomter *et al.*, 2010). Data for 145 plots measured within 1 year of the ALS acquisition dates were used for this study. Thus, the study area was defined as the geographic area represented by the portion of the Latin Square sampling design inventoried by the Norwegian NFI between 2005 and 2007 (Figure 1). Plot locations were determined using global positioning system (GPS) receivers with accuracies on the order of 0.05 m. All plot trees with diameters at-breast-height (dbh, 1.3 m) of at least 5 cm were callipered. For each tree, stem volume to the top of the tree including bark was predicted using species-specific volume models with dbh and either measured or predicted height (ht) as predictor variables (Braastad, 1966; Brantseg, 1967; Vestjordet, 1967). Volume predictions for individual trees were added to produce plot-level predictions, which were then scaled to a per unit area basis (m$^3$ ha$^{-1}$).

Wall-to-wall ALS data were acquired between 15 July 2006 and 12 September 2006 with average point density of 0.7 pulses per m$^2$. Data for only single echoes or the first of multiple echoes were used. For each plot and population unit, ALS height distributions were estimated for first echoes with heights greater than 2 m. Echoes with heights less than 2 m were considered to be from non-tree objects such as shrubs, grass or the ground. For each plot and population unit, heights corresponding to the 10$^{th}$, 20$^{th}$, …, 100$^{th}$ percentiles of the distributions were calculated and denoted as $h_1$, $h_2$, …, $h_{10}$, respectively. Canopy densities were calculated as the proportions of echoes with heights greater than 0, 10, …, 90 per cent of the range between 2 m above ground and the 95$^{th}$ height percentile and were denoted as $d_0$, $d_1$, …, $d_9$, respectively (Gobakken and Næsset, 2008). McRoberts *et al.* (2013) provide additional details for the data set.

### *Molise Region, Italy*

The 363.6-km$^2$ study area is in the southwestern part of the Molise Region in central Italy (Figure 2). Approximately 56 per cent of the area, or 20 518 ha, is covered by forests of which ~60 per cent is dominated by deciduous oaks (*Quercus cerris* L., *Quercus pubescens* Willd), ~18 per cent is dominated by hop hornbeam (*Ostrya carpinifolia* Scop.) and ~9 per cent is dominated by unmanaged beech (*Fagus sylvatica* L.) forests with structures approaching natural, old-growth forest status. The study area was tessellated into 437 hexagons, each with area of 1 km$^2$. A point was randomly selected in each hexagon and classified as 'forest' or 'non-forest' using high-resolution aerial ortho-photography. From the 197 points classified as forest, 62 were randomly selected and served as centres for 13-m radius field plots. Plot centres were determined using GPS receivers with sub-metre accuracy. For each plot, dbh (1.3 m) was measured for all trees with dbh of at least 5.0 cm. Height was measured for a sub-sample of ~10 plot trees and predicted for the remaining trees. Above ground biomass was predicted for all

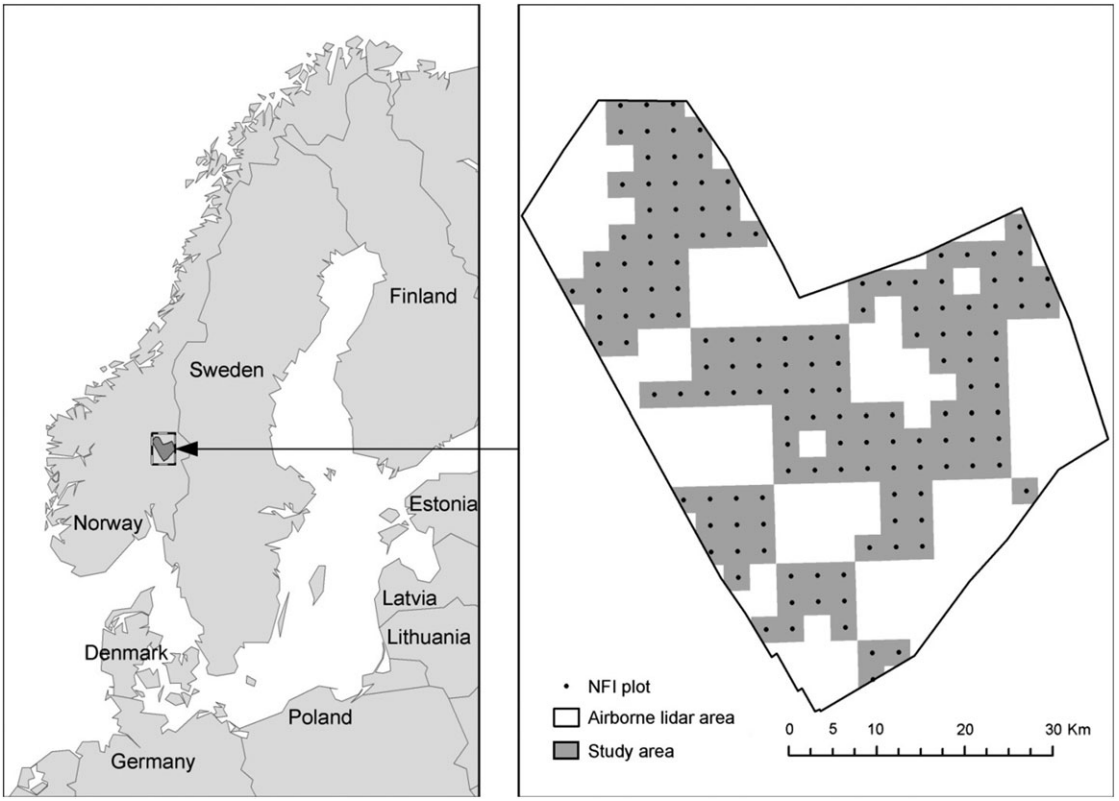


**Figure 1** Hedmark study area in Norway.

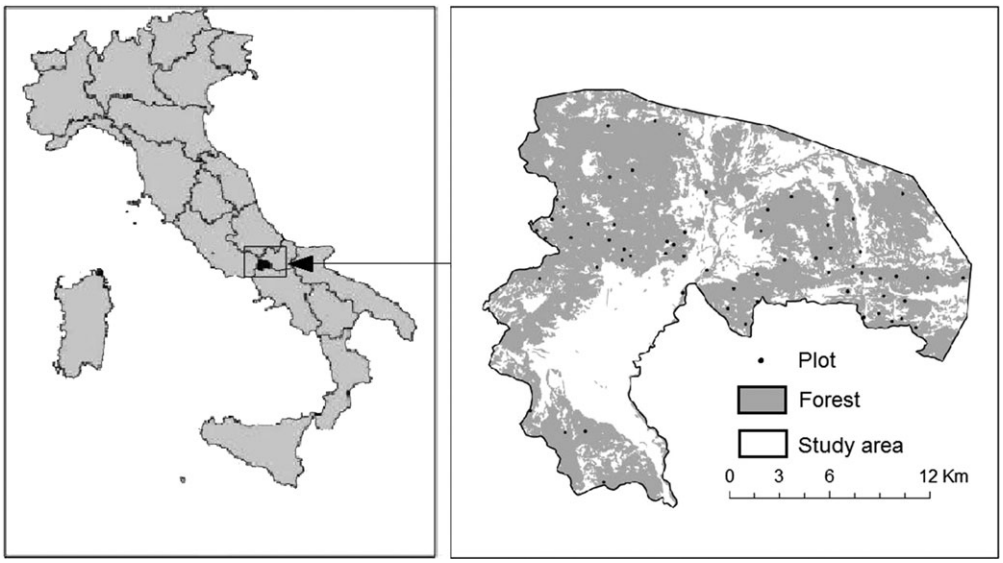


**Figure 2** Molise study area in Italy.

trees using national models (Tabacchi *et al.*, 2011), added to produce plot-level predictions and scaled to a per unit area basis (Mg $ha^{-1}$).

Wall-to-wall ALS data with mean pulse density of 1.5 pulses per $m^2$ were acquired in June 2010. For first or single echoes, the ALS metrics included heights corresponding to the $10^{th}$, $20^{th}$, …, $90^{th}$, $99^{th}$ percentiles of the height canopy distribution and the maximum, average, standard deviation, coefficient of variability, skewness and kurtosis of the distribution of echo heights. All metrics were calculated for 23-m × 23-m cells that mimicked the plot area of ~531 $m^2$ and that served as population units. Chirici *et al.* (2016a) provide more details for the data set.

### *Itasca County, Minnesota, USA*

The 7583-km$^2$ study area was located in north central Minnesota in the USA (Figure 3), and was characterized as ~80 per cent forest land. Species compositions include upland deciduous mixtures, pines (*Pinus* spp.) spruce (*Picea* spp.) and balsam fir (*Abies balsamea* (L.) Mill.) and lowlands with spruce (*Picea* spp.), tamarack (*Larix laricina* (Du Roi) K. Koch), white cedar (*Thuja occidentalis*) and black ash (*Fraxinus nigra* Marshall). Data were obtained for 115 field plots established by the Forest Inventory and Analysis (FIA) programme of the U.S. Forest Service (McRoberts *et al.*, 2010). Data were restricted to the central subplot of the four 7.32-m (24-ft) radius circular subplots to avoid issues of spatial correlation among subplot observations. Data were further restricted to plots measured in 2014, the only year for which GPS receivers with sub-metre accuracy were available. Field crews measure dbh (1.37 m, 4.5 ft) and ht for all trees with dbh of at least 12.7 cm (5 in). These data were used with statistical models to predict individual tree stem volumes, which were aggregated at subplot-level and scaled to a per unit area basis (m$^3$ ha$^{-1}$).

Wall-to-wall ALS data were acquired in April 2012 with a nominal pulse spacing of 1.5 m using Leica ALS 60 or ALS 70 sensors. The average flying height above ground was 2100–2300 m, the field of view was 40°, and the vertical accuracy was 11–15 cm. Ground returns were classified by the provider and used to construct a digital terrain model via interpolation using the Tiffs (Toolbox for Lidar Data Filtering and Forest Studies) software (Chen, 2007). Distributions of all first and single echo heights were constructed for the 168.3-m$^2$ plots and 169-m$^2$ square cells that tessellated the study area. For each plot and cell, the mean, standard deviation, skewness and kurtosis of the distributions were calculated as was quadratic mean height (Lefsky *et al.*, 1999; Chen *et al.*, 2012). In addition, heights corresponding to the 10$^{th}$, 20$^{th}$, …, 100$^{th}$ percentiles of the distributions were calculated, and canopy densities were calculated as the proportions of echoes with heights greater than 0, 10, …, 90 per cent of the range between 1.3 m above ground and the 95$^{th}$ height percentile (Gobakken and Næsset, 2008).

## Nearest neighbours techniques

### *Terminology and notation*

For notational purposes, **Y** denotes a possibly multivariate vector of response variables observed for a sample, and **X** denotes a vector of auxiliary variables with observations for the entire population. In the terminology of nearest neighbours techniques, the auxiliary variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the sample of population units for which observations of both response and feature variables are available is designated the *reference set* with size denoted *n*; and the set of population units for which predictions of response variables are desired is designated the *target set* with size denoted *N*.

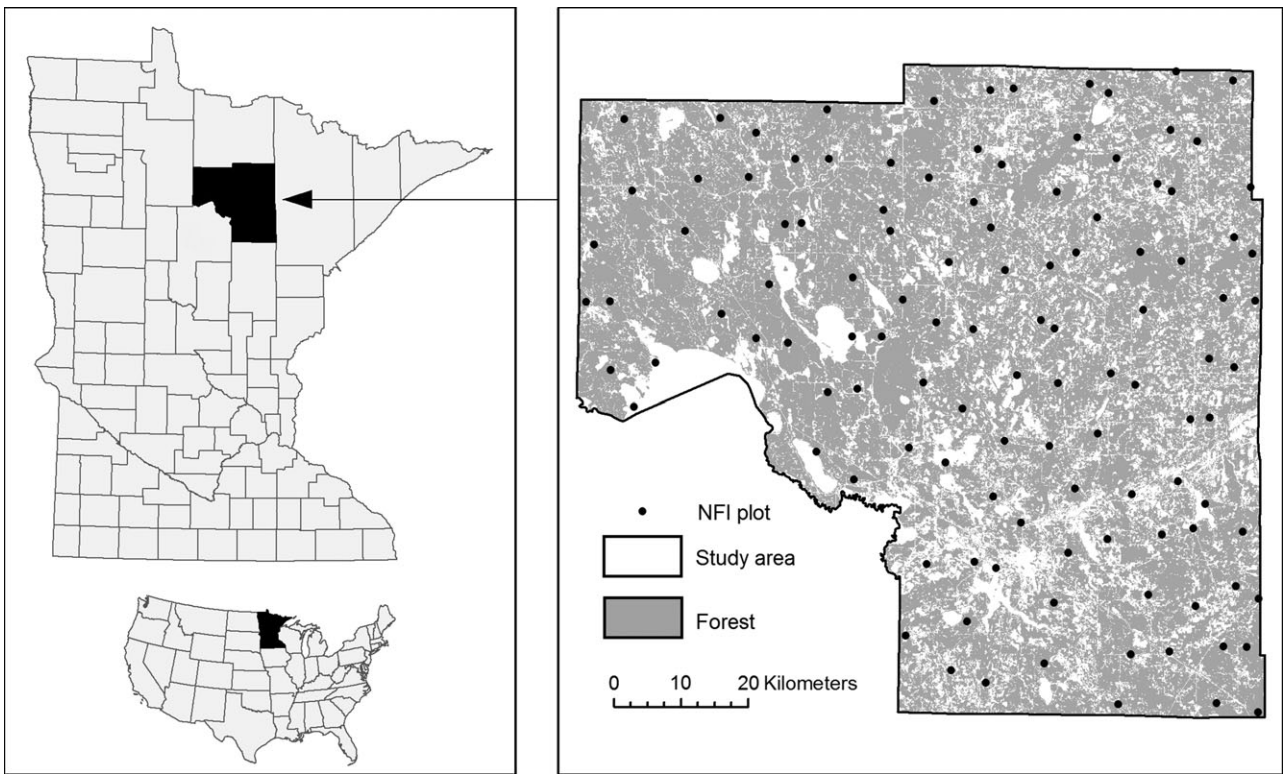


**Figure 3** Itasca study area in Minnesota, USA.

For continuous response variables such as forest volume and biomass, the nearest neighbours prediction, $\hat{y}_i$, for the $i^{th}$ target unit is calculated as,

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

where $\{y_j^i, j = 1, 2, .., k\}$ is the set of response variable observations for the $k$ reference units that are most similar or nearest to the $i^{th}$ target unit in feature space with respect to a distance metric, $d$, and $w_{ij}$ is the weight assigned to the $j^{th}$ nearest neighbour with $\sum_{j=1}^{k} w_{ij} = 1$.

## Distance metrics

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

$$d_{ij} = \sqrt{\left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{M} \left(\mathbf{X}_i - \mathbf{X}_j\right)}, \quad (2)$$

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 $\mathbf{M}$ is a square, positive semi-definite matrix. A recent study conducted under the auspices of Action FP1001 of the European programme Cooperation in Science in Technology (COST, 2015) identified the Euclidean metric, the Mahalanobis metric and the canonical correlation analysis metric (Section 3.2.4) as the most frequently used metrics (Chirici *et al.*, 2016b). These three distance metrics, including a weighted variation of the Euclidean metric, were investigated for this study.

### *Euclidean distance metric*

With the Euclidean distance metric (EUCL), the $\mathbf{M}$ matrix from equation (2) is the identity matrix, $\mathbf{I}$, and distance is expressed as

$$d_{ij} = \sqrt{\left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{I} \left(\mathbf{X}_i - \mathbf{X}_j\right)}. \quad (3)$$

The EUCL metric is the simplest, most intuitive, and probably the most frequently used metric.

### *Weighted Euclidean distance metric*

The weighted Euclidean distance metric (WEUCL) is similar to the EUCL metric, except $\mathbf{M}$ from equation (2) is a diagonal, non-identity matrix, $\mathbf{D}$, and distance is expressed as

$$d_{ij} = \sqrt{\left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{D} \left(\mathbf{X}_i - \mathbf{X}_j\right)}. \quad (4)$$

Optimization of the metric entails selection of optimal values for the matrix diagonal elements and can be computationally intensive, even for relatively small numbers of feature variables.

Genetic algorithms (GAs) have emerged as an increasingly useful technique for optimizing selection of the diagonal elements of the $\mathbf{D}$ matrix. GAs are search heuristics that mimic natural selection to solve optimization problems (Holland, 1975). The process is iterative and starts from a population of randomly generated individuals with the population in each iteration called a generation. In each generation, an individual in the population consists of a set of diagonal elements for the $\mathbf{D}$ matrix. Individuals are evaluated with respect to their fitness, which, for $k$-NN applications, is typically a criterion related to the sum of squared errors for continuous response variables. Each subsequent generation consists of the individuals from the previous generation that are characterized as more fit relative to a selected criterion. These individuals are modified by combining and randomly mutating to produce new individuals. The new generation of individuals is then used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced or a satisfactory fitness level has been reached for at least one individual in the population. For this study, GAs were used to optimize only the WEUCL metric. Tomppo and Halme (2004), Tomppo *et al.* (2009), Holopainen *et al.* (2010) and McRoberts (2012, 2015) all used GAs to select diagonal elements as a means of optimizing the WEUCL distance metric, and McRoberts (2008) and Latifi *et al.* (2010) used GAs to select feature variables.

### *Mahalanobis distance metric*

With the Mahalanobis distance metric (MAHA), $\mathbf{M}$, from equation (2) is the inverse of the feature variable covariance matrix, $\mathbf{V}$, and distance is expressed as

$$d_{ij} = \sqrt{\left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{V}^{-1} \left(\mathbf{X}_i - \mathbf{X}_j\right)}, \quad (5)$$

(Mahalanobis, 1936). The MAHA metric is often used for comparison purposes, but seldom is selected as the optimal metric for forestry applications (Maltamo *et al.*, 2003; Latifi *et al.*, 2010; Ver Hoef and Temesgen, 2013).

### *Canonical correlation analysis distance metric*

With the canonical correlation analysis distance metric (CCA), a system of linear models is solved to obtain estimates of coefficient vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, that maximize the correlation between $U = \alpha_1 \cdot Y_1 + \cdots + \alpha_p \cdot Y_p$ and $V = \beta_1 \cdot X_1 + \cdots + \beta_q \cdot X_q$ where $Y_j$ denotes the $j^{th}$ response variable, $X_j$ denotes the $j^{th}$ feature variable, and $p$ and $q$ are the numbers of response and feature variables, respectively. The solutions are obtained using canonical decompositions for which the eigenvectors, also designated canonical correlation coefficients, are denoted $\boldsymbol{\Gamma}$, and the corresponding eigenvalues, also designated canonical correlations, are denoted $\lambda$. Feature space distances with this metric are expressed as

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \Gamma \Lambda^2 \Gamma' \left(X_i - X_j\right)}, \quad (6)$$

where the elements of the diagonal matrix, $\boldsymbol{\Lambda}$, are the squares, $\lambda^2$, of the canonical correlations.

The CCA metric assumes a linear relationship between each of the response variables and the feature variables. The user has no control over the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors, and although the $\boldsymbol{\beta}$ vector is of little consequence, the $\boldsymbol{\alpha}$ vector may not combine the response variables in a relevant manner. The metric was first proposed for nearest neighbours applications by Moeur and Stage (1995) who used only a single neighbour and characterized the combination of the CCA metric and $k = 1$ as the 'Most Similar Neighbour' technique (LeMay and Temesgen, 2005). However,

the metric has also been used with multiple neighbours (Maltamo *et al.*, 2003, 2009; Packalén and Maltamo, 2007).

### *Feature variable selection*

When considering nearest neighbours 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. Three important detrimental consequences follow: (1) nearest neighbours are at greater distances from target units (Schaal *et al.*, 1998); (2) the distance to the nearest neighbour approaches the distance to the farthest neighbour (Beyer *et al.*, 1998) and (3) extrapolations beyond the ranges of the feature variables in the reference set are more probable (McRoberts, 2009). McRoberts *et al.* (2015) showed that when using ALS data to predict above ground biomass, and presumably also related response variables such as forest volume, the effects of the first two consequences are not severe. 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) characterized such feature variables as 'irrelevant'. Irrelevant feature variables introduce randomness into distance calculations and contribute to selection of spurious neighbours and less accurate predictions. Thus, optimization of nearest neighbours distance metrics should focus on simultaneously eliminating irrelevant feature variables and weighting feature variables in proportion to their relevance.

### *Neighbour weighting*

#### *Inverse distance weighting*

The most common approach to weighting neighbours when calculating $k$-NN predictions is to weight neighbours inversely proportionally to a power of the distance, $d_{ij}$, between the $i^{th}$ target unit and the $j^{th}$ reference unit,

$$w_{ij} = \frac{d_{ij}^{-t}}{W}, \tag{7}$$

where $W = \sum_{j=1}^{k} d_{ij}^{-t}$ and $t \geq 0$. Commonly, $t = 0$, $t = 1$ or $t = 2$ is arbitrarily selected where $t = 0$ corresponds to weighting all neighbours equally. For this study, the special case of $t = 0$ is characterized as '$c$-weighting', whereas weighting schemes corresponding to $t > 0$ are characterized as '$t$-weighting'. Other than McRoberts (2012), McRoberts *et al.* (2015) and Wilson *et al.* (2012), no reports of attempts to optimize the selection of $t$ are known. For small numbers of feature variables and/or large reference sets, $d_{ij} = 0$ may occur in which case equation (7) leads to computational errors. For this study, if $d_{ij} = 0$ for $j = 1, \ldots, k$, then all distances are arbitrarily reset to 1, i.e., $d_{ij} = 1$ for all neighbours. If $d_{ij} = 0$ for $j = 1, \ldots, k'$ where $k' < k$, then all 0-distances are arbitrarily reset to half of the smallest non-zero distance, i.e.,

$$d_{ij} = \frac{d_{ik'+1}}{2}, \tag{8}$$

for $j = 1, \ldots, k'$.

#### *Dudani weighting*

Dudani (1976) proposed a weighting scheme that bases the weight for the $j^{th}$ neighbour on the ratio of two distances, the distance between the $j^{th}$ and $k^{th}$ neighbours and the distance between the first and the $k^{th}$ neighbours. Because this scheme gives 0 weight to the $k^{th}$ neighbour, it was modified for this study by using the $k + 1^{st}$ neighbour instead of the $k^{th}$ neighbour to,

$$w_{ij} = \frac{\left(d_{ik+1} - d_{ij}\right)/\left(d_{ik+1} - d_{i1}\right)}{W}, \tag{9}$$

where $W = \sum_{j=1}^{k} ((d_{ik+1} - d_{ij})/(d_{ik+1} - d_{i1}))$. For notational purposes, this metric is characterized as '$d$-weighting'. With the formulation of equation (9), calculation of weights for $k$ neighbours requires distances for $k + 1$ neighbours. For small numbers of feature variables and/or large reference sets, $d_{ik'} = d_{ik+1}$ for $k' < k + 1$ may occur in which case, $w_{ij} = 0$ for $j = k', \ldots, k$. For this study, if $d_{i1} = d_{ik+1}$, then all distances are reset to 1, i.e., $d_{ij} = 1$ for all neighbours. If $d_{ik'} = d_{ik+1}$ for $j = k', \ldots, k$ where $1 < k' < k + 1$, then all such distances are reset to the mean of the $k + 1^{st}$ distance and the greatest distance that differs from the $k + 1^{st}$ distance,

$$d_{ij} = \frac{d_{ik'-1} + d_{ik+1}}{2}, \tag{10}$$

for $j = k', \ldots, k$. For example, suppose $k = 5$ and the six smallest distances for the $i^{th}$ target unit are $d_{i1} = 1$, $d_{i2} = 2$, $d_{i3} = 3$, $d_{i4} = 5$, $d_{i5} = 5$ and $d_{i6} = 5$. Using only equation (9), the weights for the fourth and fifth neighbours would be $w_{i4} = w_{i5} = 0$, which would exclude the observations from the fourth and fifth neighbours when calculating the $k$-NN prediction. However, by applying equation (10), the fourth and fifth distances are reset to $d_{i4} = d_{i5} = 4.0$ with the result that $W = 2.75$ and $w_{i1} = 0.364$, $w_{i2} = 0.272$, $w_{i3} = 0.182$, $w_{i4} = 0.091$ and $w_{i5} = 0.091$.

### *Number of nearest neighbours,* k

The value of $k$ may be selected to optimize multiple criteria either individually or in combination. The relevant factors in selecting $k$ are threefold: (1) small values of $k$ are generally preferred as a means of reducing complexity and computational intensity, (2) small values may yield root mean square errors that are greater than the standard deviations of the response variable observations, meaning that the response variable mean over all reference set observations when used as a prediction for every target unit would better maximize accuracy than the $k$-NN predictions and (3) large values of $k$, tend to produce overestimation for small observations and underestimation for large observations. McRoberts (2012) reviews additional issues related to selection of $k$. For many reported applications, the value of $k$ has been arbitrarily selected as $k = 1$ or $k = 5$ based on values reported elsewhere in the literature. The rationale for such decisions is uncertain, if not confusing, at least without assessing the consequences; for example, if a regression model was used, parameter estimates reported elsewhere in the literature would certainly not be arbitrarily selected for a new application that used the same model form and predictor variables.

## Analyses

### *Optimization*

For each distance metric, the value of $k$, and the value of $t$ for the $t$-weighting scheme that minimized the sum of squared residuals, $SS_{res}$, were determined using leave-one-out cross validation (Elisseeff and Pontil, 2002) for each combination of each number of feature variables, $m$, beginning with $m = 1$. For each value of $m$, the combination of feature variables with the smallest $SS_{res}$ was selected. In addition, beginning with $m = 3$, three $F$-tests were conducted: (1) $SS_{res}^{m}$ was compared with $SS_{res}^{m-1}$, (2) $SS_{res}^{m-1}$ was compared with $SS_{res}^{m-2}$ and (3) $SS_{res}^{m}$ was compared with $SS_{res}^{m-2}$. For each test, the statistic was calculated as

$$F = \frac{(SS_{res}^{m_1} - SS_{res}^{m_2})/(m_2 - m_1)}{SS_{res}^{m_2}/(n - m_2)}, \tag{11}$$

where $m_1$ and $m_2$ are the smaller and larger number of feature variables, respectively. The second and third tests were conducted because for several combinations of data sets, metrics and weighting schemes, the first test corresponding to inclusion of a single new feature variable indicated no decrease in $SS_{res}$, but inclusion of a second new feature variable did produce a decrease in $SS_{res}$. If all three tests failed to exceed the critical value corresponding to $\alpha = 0.01$, the combination with $m - 2$ feature variables was selected, and no combinations with greater numbers of feature variables were considered. Issues related to significance level and the $\alpha = 0.01$ significance level in particular are addressed in the Appendix 1. Although feature variable selection was based on $SS_{res}$, results are reported using a pseudo-$R^2$ denoted and calculated as

$$R^{2*} = \frac{SS_{mean} - SS_{res}}{SS_{mean}}. \tag{12}$$

A stepwise variable selection procedure is an alternative to considering all possible combinations of each number of feature variables. However, stepwise algorithms are known to perform poorly when the feature variables are strongly correlated (Harrell, 2001, pp. 64–65). For this study, preliminary investigations produced decidedly sub-optimal results when using stepwise procedures, presumably because of the strong correlations among the ALS height and density metrics used as feature variables.

### *Arbitrary selections*

Many implementations of nearest neighbours techniques feature arbitrary selections of $k$ and $t$, often using all feature variables. For example, the popular Most Similar Neighbour variation of $k$-NN uses the CCA metric and $k = 1$ with no attempt to optimize other than the optimization that is inherent in the metric (Moeur and Stage, 1995). A reasonable question pertains to the degree of sub-optimality that results from such arbitrary selections. As a second example, metrics that incorporate feature variable weighting such as the WEUCL and CCA metrics theoretically circumvent the necessity of selecting feature variables because they assign negligible weights to irrelevant variables. Thus, a second reasonable question pertains to the degree to which this potential is realized when using these metrics. To address these issues, estimates of means and standard errors for common arbitrary selections of $k$ and $t$ using the WEUCL and CCA metrics were compared with estimates and standard errors for optimized configurations.

### *Multiple neighbours at the k$^{th}$ distance*

An issue related to both the value of $k$ and the number of feature variables pertains to multiple neighbours at the same distance as the $k^{th}$ neighbour. For example, if $k = 5$ and the fifth and sixth neighbours are at the same distance for a particular prediction, then a decision must be made as to which of the two neighbours to select. Often the first neighbour in the order is selected, and any detrimental consequences are ignored. Alternatively, both neighbours could be used or additional criteria such as distance in geographical space could be introduced (Franco-Lopez *et al.*, 2001, Section 3.1.4; Baffetta *et al.*, 2009, Section 2.1). Multiple neighbours at the $k^{th}$ distance are more probable with larger reference sets and smaller numbers of feature variables, and the consequences are more detrimental with smaller values of $k$. For larger values of $k$, the effects are attenuated because predictions are calculated as means over more neighbours and because the $k^{th}$ neighbour receives a smaller weight relative to the other neighbours at smaller distances. Finally, the phenomenon is independent of the particular neighbour weighting scheme used. For each data set and for selected optimized combinations of distance metrics and neighbour weighting schemes, the numbers of neighbours at the $k^{th}$ distance were determined.

### *Inference*

For each data set, consideration was given to inferences for the mean per unit area of the response variable expressed as,

$$\hat{\mu} \pm t \cdot \mathrm{SE}(\hat{\mu}), \tag{13}$$

where $\hat{\mu}$ is the estimate of the mean, $\mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})}$ is the standard error of $\hat{\mu}$ and $t$ depends on the desired significance level and the distribution of the response variable. Because $t = 2$ produces an ~95 per cent confidence interval for most distributions and applications, two-SE confidence intervals were used for this study. Any uncertainty resulting from using a model to predict individual tree volume or biomass was considered negligible relative to the variances of estimators of the large area population mean (McRoberts and Westfall, 2014).

#### *Simple random sampling estimators*

For equal probability samples, the simple random sampling (SRS) estimator of the mean is,

$$\hat{\mu}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{14}$$

where $n$ is the reference set size, $i$ indexes the reference units (plots) and $y_i$ is the reference unit observation. The estimator of the variance of $\hat{\mu}_{SRS}$ is,

$$\hat{\mathrm{Var}}(\hat{\mu}_{\mathrm{SRS}}) = \frac{1}{n\cdot(n-1)} \sum_{i=1}^{n} \left(y_i - \hat{\mu}_{\mathrm{SRS}}\right)^2. \quad (15)$$

Of importance the SRS estimators use no auxiliary information. For systematic samples, as used for this study, the SRS variance estimator may have a slight positive bias (Särndal *et al.*, 1992, p. 83). However, Aune-Lundberg and Strand (2014) concluded that the SRS estimators are still a safe and conservative choice. The primary advantages of the SRS estimators are that they are intuitive and unbiased, but the disadvantage is that variances may be large, particularly for small sample sizes and/or large within-population variability.

*Model-assisted regression estimators*

Model-assisted regression estimators use models based on auxiliary data to enhance inferences but rely on the probability sample for validity (Särndal *et al.*, 1992; Särndal, 2011). A 'synthetic' estimator of the mean is formulated as

$$\hat{\mu}_{\mathrm{Syn}} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i, \quad (16)$$

where $N$ is the target set (population) size and $\hat{y}_i$ is the $k$-NN prediction for the $i^{\mathrm{th}}$ target unit. Any systematic prediction errors induce bias into this estimator, which for equal probability samples can be estimated as

$$\hat{\mathrm{Bias}}\left(\hat{\mu}_{\mathrm{Syn}}\right) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i, \quad (17)$$

where $\varepsilon_i = \hat{y}_i - y_i$. The model-assisted, generalized regression (GREG) estimator is then defined as

$$\hat{\mu}_{\mathrm{GREG}} = \hat{\mu}_{\mathrm{Syn}} - \hat{\mathrm{Bias}}\left(\hat{\mu}_{\mathrm{Syn}}\right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i - \frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right), \quad (18)$$

with variance estimator,

$$\hat{\mathrm{Var}}(\hat{\mu}_{\mathrm{GREG}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2, \quad (19)$$

where $\bar{\varepsilon} = (1/n) \sum_{i=1}^{n} \varepsilon_i$ (Särndal *et al.*, 1992). Despite the label characterizing the estimator, prediction techniques other than regression can be used (Breidt and Opsomer, 2000, 2009; Zheng and Little, 2004; Lehtonen *et al.*, 2005; Särndal, 2007). The degree to which the auxiliary information increases precision and thereby shortens the confidence interval is often calculated using relative efficiency,

$$\mathrm{RE} = \frac{\hat{\mathrm{Var}}(\hat{\mu}_{\mathrm{SRS}})}{\hat{\mathrm{Var}}(\hat{\mu}_{\mathrm{GREG}})}. \quad (20)$$

The primary advantage of the GREG estimators is that they capitalize on the relationship between the reference set observations and their corresponding predictions to reduce the variance of the estimate of the population mean.

*Comparisons*

Estimates of population means and their SEs and corresponding two-SE confidence intervals were compared for $k$-NN configurations consisting of all combinations of the three data sets, the WEUL and CCA metrics, and the $t$- and $d$-weighting schemes. For these analyses, optimization consisted of selecting optimal subsets of feature variables, optimal values of $k$ and for the $t$-weighting scheme, optimal values of $t$. Neither the EUCL and MAHA metrics nor the c-weighting scheme were used for these analyses because the WEUCL and CCA metrics were deemed preferable for reasons noted in the following sections. For all three data sets, estimated means and SEs for the optimized WEUCL and $t$-weighting configuration were compared with estimates obtained using the EUCL metric with arbitrary choices for $t$ and $k$. Similarly, for all three data sets, estimated means and SEs for the optimized CCA and $d$-weighting configuration were compared with estimates obtained using the same metric and neighbour weighting scheme with arbitrary choices for $k$. For the latter two sets of analyses, arbitrary choices for $k$ were $k = 1$ and $k = 5$ and arbitrary choices for $t$ when using $t$-weighting were $t = 0$, $t = 1$ and $t = 2$.

## Results and discussion

*Software*

Appendix 2 briefly describes the software constructed for implementation of the $k$-NN technique for this study.

*Feature variable selection*

Methods for selecting feature variables entailed considering all combinations of all numbers of feature variables until reductions in $SS_{res}$ failed to exceed selected threshold values. For the MAHA metric, and to a lesser degree for the CCA metric, multiple combinations of larger numbers of feature variables produced matrices that were not positive definite and, therefore, could not be inverted as is required for these metrics. This result is attributed to large correlations between some pairs of the ALS metrics. This phenomenon suggests that the MAHA metric, and to some extent the CCA metric, may not be appropriate when feature variables are highly correlated.

Selection of numbers and combinations of feature variables also entailed using successive $F$-tests to accommodate the possibility of gradual and significant reductions in $SS_{res}$ that could not be detected with single $F$-tests for consecutive numbers of feature variables. The technique generally performed as anticipated.

*Prediction accuracy*

*Distance metrics*

Overall, the primary results of optimization with respect to prediction accuracy were twofold. First, optimization for the WEUCL, MAHA and CCA metrics produced slightly greater accuracies as assessed by $R^{2^*}$ than the EUCL metric (Table 1). This result was as expected because the three former metrics all feature greater potential for optimization. Second, optimization for

the WEUCL and CCA metrics produced accuracies that were generally similar to each other.

Optimization for the WEUCL, MAHA and CCA metrics produced $R^{2*}$ values in the range 0.79–0.85 for all three data sets. Relative to $R^{2*}$ values for the EUCL metric, $R^{2*}$ values for the three former metrics represented increase in the range 0.00–0.04 for the Molise and Itasca data sets but 0.08–0.12 for the Hedmark data set. These small ranges indicate that optimization of the WEUCL, MAHA and CCA metrics via selection of feature variables, *k*, and weighting scheme produced comparable results.

Although optimization produced similar $R^{2*}$ values, each of the WEUCL, MAHA and CCA metrics has disadvantages. For the WEUCL metric, optimization can be computationally intensive, particularly with large reference sets and large numbers of feature variables. In particular, GA optimization is unrealistic, if not impossible, for very large numbers of feature variables. As noted in the literature review, supervised metrics that are optimized using observations of the response variable are generally more accurate than unsupervised metrics. In this context, the WEUCL and CCA metrics are characterized as supervised, whereas the MAHA metric is characterized as unsupervised. As previously noted, many combinations of feature variables could not be considered for the MAHA metric because the associated matrix was not positive definite. In addition, the MAHA metric makes no provision for minimizing the influence of irrelevant feature variables as do the WEUCL and CCA metrics. The CCA metric, as with the MAHA metric, was occasionally hampered by matrices that were not positive definite, but not to the same degree as for the MAHA metric and mostly for larger numbers of feature variables than for the MAHA metric. Also, larger numbers of feature variables were selected when optimizing the CCA metric, a result that can likely be attributed to near negligible weights associated with mostly irrelevant feature variables that otherwise would not have been selected. The disadvantage is that larger numbers of feature variables mean that optimization requires consideration of more combinations of feature variables, thereby increasing computational intensity.

*Number of neighbours*

Optimal values of *k* tended to be small, never greater than $k = 10$ for all combinations of data set, metric and neighbour weighting scheme. In the event of optimal values of $k > 10$, smaller values would likely be used as a means of reducing computational intensity. As graphs of optimization criteria versus *k* typically reveal, considerably smaller values of *k* often produce only slight changes in the optimization criterion (e.g. Figure 2 in McRoberts *et al.*, 2002; Figure 2 in McRoberts, 2012).

*Neighbour weighting*

The *t*- and *d*-weighting schemes produced little increase in $R^{2*}$ relative to *c*-weighting. This result can likely be partially attributed to the relatively small values of *k*. Among the three neighbour weighting schemes, *c*- and *d*-weighting schemes require no optimization and are simple to implement, whereas optimization of *t*-weighting can be computationally intensive.

*Summary*

Overall, based on the potential for optimization, the WEUCL and CCA metrics are preferable to the EUCL and MAHA metrics. In addition, based on the potential for optimization *t*- and *d*-weighting are preferable to c-weighting. For purposes of assessing the effects on inferences of optimizing *k*-NN choices versus using all feature variables in combination with arbitrary selections of *k* and neighbour weighting, only the WEUCL and CCA metrics and only the *t*- and *d*-weighting schemes were considered.

## *Inference*

Estimates of population means per unit area obtained using the SRS estimators and the GREG estimators with the optimized WEUCL and CCA metrics were generally similar; in particular, all GREG estimates were within two SRS SEs of the SRS estimates

**Table 1** Prediction accuracies

| Data set | Distance metric[1] | *c*-Weighting | | | *t*-Weighting | | | | *d*-Weighting | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | No. feature variables | *k* | $R^{2*}$ | No. feature variables | *k* | *t* | $R^{2*}$ | No. feature variables | *k* | $R^{2*}$ |
| Hedmark | EUCL | 3 | 6 | 0.74 | 3 | 6 | 0.0 | 0.74 | 3 | 10 | 0.73 |
| | WEUCL | 3 | 3 | 0.84 | 3 | 3 | 0.6 | 0.84 | 3 | 4 | 0.84 |
| | MAHA | 4 | 2 | 0.82 | 6 | 3 | 0.7 | 0.83 | 4 | 8 | 0.83 |
| | CCA | 7 | 3 | 0.85 | 7 | 4 | 0.7 | 0.85 | 7 | 3 | 0.85 |
| Molise | EUCL | 3 | 2 | 0.78 | 3 | 4 | 1.6 | 0.82 | 2 | 3 | 0.78 |
| | WEUCL | 3 | 2 | 0.80 | 3 | 4 | 1.6 | 0.82 | 2 | 4 | 0.78 |
| | MAHA | 3 | 3 | 0.79 | 3 | 3 | 1.1 | 0.80 | 3 | 4 | 0.79 |
| | CCA | 3 | 2 | 0.79 | 4 | 6 | 1.2 | 0.80 | 4 | 2 | 0.80 |
| Itasca | EUCL | 5 | 2 | 0.84 | 4 | 2 | 0.6 | 0.84 | 4 | 2 | 0.83 |
| | WEUCL | 5 | 2 | 0.85 | 4 | 2 | 0.4 | 0.85 | 4 | 2 | 0.84 |
| | MAHA | 4 | 1 | 0.86 | 6 | 9 | 3.6 | 0.88 | 5 | 2 | 0.87 |
| | CCA | 7 | 2 | 0.87 | 7 | 2 | 1.8 | 0.87 | 7 | 3 | 0.87 |

[1]EUCL, Euclidean; WEUCL, weighted Euclidean; MAHA, Mahalanobis; CCA, canonical correlation analysis.

(Table 2). Despite similarity in estimates of means, the GREG SEs were less than half of the SRS SEs, which indicate the utility of the ALS auxiliary information for increasing the precision of estimates of the population means. The differences in GREG and SRS SEs are also reflected in REs, which can be interpreted as the factor by which the sample sizes would have to be increased to achieve the same SEs using the SRS estimators without the ALS data as were achieved using the GREG estimators, the optimized *k*-NN technique, and the ALS data. In particular, optimization of the *k*-NN technique with both the WEUCL and CCA metrics produced REs in the range 4.5–9.5. In addition, ratios of squares of SEs for optimized and arbitrary choices for the WEUCL and CCA metrics indicate that effects of optimization are equivalent to increasing sample sizes by factors as great as 5.5.

### *Arbitrary selections*

Overall, the effects of optimization versus arbitrary choices were twofold. First, for each data set, estimates of means per unit area for the response variables were generally similar, although the estimates obtained using the optimal *k*-NN choices were generally smaller than the estimates obtained using the arbitrary choices (Table 2). Second, optimization produced substantially smaller SEs than arbitrary choices for *k*-NN selections. The consequences of common arbitrary choices of $k = 1$ or $k = 5$; $t = 0$, $t = 1$ or $t = 2$; and all feature variables for use with the WEUCL and CCA metrics were always detrimental relative to optimized choices. For the Hedmark data set, standard errors for the arbitrary choices ranged from 3.74 to 5.53, but for optimized configurations were 2.89 for both the WEUCL and CCA metrics. For the Molise data set, standard errors for arbitrary choices ranged from 6.04 to 11.61, but for optimized configurations were 5.08 for the WEUCL metric and 4.88 for the CCA metric. For the Itasca data set, standard errors ranged from 2.77 to 3.52 for the arbitrary choices, but for optimized configurations were 2.41 for the WEUCL metric and 2.15 for the CCA metric.

Two overall results relative to arbitrary choices versus optimization are important. First, optimized selection of *k*, *t* and feature variables contributed to substantially more accurate predictions and enhanced inferences in the form of shorter confidence intervals. This result suggests that arbitrary choices and/or failure to optimize will be increasingly difficult to justify. While perhaps placing an additional burden on researchers, this result represents continuing maturation of the *k*-NN technique. Second, minimization of the detrimental effects of irrelevant feature variables theoretically possible with the WEUCL and CCA metrics was at best only partially realized.

### *Multiple neighbours at the k$^{th}$ distance*

Among the 36 combinations of data set, distance metric and neighbour weighting scheme, the combination of the CCA metric and *d*-weighting produced as large or nearly as large $R^{2*}$ than any other combination of metric and weighting scheme (Table 1). In addition, the combination of the CCA metric and *d*-weighting was computationally easiest to implement. For all three data sets, there were no instances of multiple neighbours at the $k^{th}$ distance for forest predictions. However, for the Hedmark data set, 11.4 per cent of non-forest predictions had multiple neighbours at the $k^{th}$ distance, but in each instance reference set observations for both the $k^{th}$ and $k + 1^{st}$ neighbours were identically zero and, therefore, the predictions were not affected. For this study, all three reference sets were relatively small and the numbers of optimally selected feature variables

**Table 2** Confidence intervals for population means per unit area for optimal configurations

| Estimator | Distance metric | Neighbour weighting | $\hat{\mu}$ | SE($\hat{\mu}$) | Confidence interval[1] | RE[2] |
|---|---|---|---|---|---|---|
| *Hedmark* | | | | | | |
| SRS | – | – | 74.26 | 7.45 | [59.36, 89.16] | – |
| GREG | WEUCL | *t*-weighting | 79.55 | 2.47 | [74.61, 84.49] | 9.36 |
| | | *d*-weighting | 80.82 | 2.53 | [75.76, 85.88] | 8.90 |
| | CCA | *t*-weighting | 77.55 | 2.51 | [72.53, 82.57] | 9.03 |
| | | *d*-weighting | 83.23 | 2.89 | [77.45, 89.01] | 7.45 |
| *Molise* | | | | | | |
| SRS | – | – | 108.23 | 10.94 | [86.35, 130.11] | – |
| GREG | WEUCL | *t*-weighting | 111.56 | 4.60 | [102.36, 120.76] | 5.66 |
| | | *d*-weighting | 105.97 | 5.09 | [95.79, 116.15] | 4.62 |
| | CCA | *t*-weighting | 106.01 | 4.87 | [96.27, 115.75] | 5.05 |
| | | *d*-weighting | 107.58 | 4.88 | [97.82, 117.34] | 5.04 |
| *Itasca* | | | | | | |
| SRS | – | – | 50.63 | 5.96 | [38.71, 62.55] | – |
| GREG | WEUCL | *t*-weighting | 50.51 | 2.35 | [45.81, 55.21] | 6.43 |
| | | *d*-weighting | 51.16 | 2.40 | [46.36, 55.95] | 6.16 |
| | CCA | *t*-weighting | 50.38 | 2.16 | [46.06, 54.70] | 7.64 |
| | | *d*-weighting | 54.04 | 2.15 | [49.74, 58.34] | 7.68 |

[1]Two-standard error confidence interval.
[2]RE = Relative efficiency.

were relatively large, both of which minimize the probability of multiple neighbours at the $k^{th}$ distance; conversely, the relatively small optimal values of $k$ would tend to exacerbate the effects of the phenomenon. For other studies with reference set sizes on the order of 1000s, small numbers of feature variables such as often are derived from remotely sensed spectral data, and $k = 1$ (e.g. Ohmann *et al.*, 2014), results may be quite different.

## Conclusions

Four conclusions were drawn from the study. First, regardless of whether the weighted Euclidean, Mahalanobis or the canonical correlation analysis metric was used, optimization via selection of feature variables, $k$, and neighbour weighting produced generally comparable prediction accuracies and, more importantly, comparably precise estimates of mean volume or mean biomass per unit area. Furthermore, regardless of the metric, optimization produced considerably greater precision than arbitrary choices. Together, these two results suggest that the crucial issue is optimization versus no optimization rather than the particular $k$-NN configuration that is optimized.

Second, despite comparable results, the weighted Euclidean and canonical correlation analysis metrics have greater potential for optimization and are therefore preferable to the Euclidean and Mahalanobis metrics. Among the neighbour weighting schemes, $c$-weighting has little potential for optimization relative to $t$- and $d$-weighting, but optimization of $t$-weighting is computationally intensive, whereas no optimization is necessary for $d$-weighting. Given the consistency of the results and the considerable variety among the forest conditions represented by the three data sets, a reasonable degree of generalization is warranted for these findings. Thus, the second conclusion is that the combination of the canonical correlation metric and $d$-weighting, when optimized by selecting optimal subsets of feature variables and optimal values of $k$, merits serious consideration when estimating parameters related to forest volume and biomass using ALS data as auxiliary information.

Third, despite the potential of the weighted Euclidean and canonical correlation analysis metrics to circumvent selection of feature variables by assigning negligible weights to irrelevant feature variables, selections of smaller numbers of feature variables produced greater precision for estimates of population means per unit area than use of all feature variables. This conclusion confirms a conclusion previously reported by Packalén *et al.* (2012).

Fourth, optimization of $k$-NN configurations via selection of feature variables, $k$, and neighbour weighting, regardless of the distance metric, produced considerably more precise estimates of population means per unit area than use of all feature variables and arbitrary selections of $k$ and $t$. Although arbitrary selections may be warranted under unique situations, authors should justify any decisions not to optimize and report assessments of the degree to which arbitrary selections produce suboptimal prediction accuracies and inferences.

For forestry applications, particularly inventory applications, the greater prediction accuracies and corresponding smaller standard errors achieved using optimized $k$-NN configurations represent substantial inferential enhancements. Relative efficiencies associated with smaller standard errors (Table 2) are theoretically the factors by which sample sizes can be reduced with no loss of precision when using the combination of the ALS auxiliary data and the $k$-NN technique relative to using SRS estimators with no auxiliary data.

## Conflict of interest statement

None declared.

## References


Aune-Lundberg, L. and Strand, G.H. 2014 Comparison of variance estimation methods for use with two dimensional systematic sampling of land use/land cover data. *Environ. Modell. Softw.* **61**, 87–97.

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

Baffetta, F., Corona, P. and Fattorini, L. 2011 Design-based diagnostics for k-NN estimators of forest resources. *Can. J. For. Res.* **41**, 59–72.

Bellman, R. 1961 *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

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

Blum, A.L. and Langley, P. 1997 Selection of relevant features and examples in machine learning. *Artif. Intell.* **97**, 245–271.

Braastad, H. 1966 Volume tables for birch. *Meddelser norske SkogforsVes* **21**, 23–78 (In Norwegian with English summary).

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

Breidt, F.J. and Opsomer, J.D. 2000 Local polynomial regression estimators in survey sampling. *Ann. Stat.* **28** (4), 1026–1053.

Breidt, F.J., Opsomer, J.D. 2009 Nonparametric and semiparametric estimation in complex surveys. In *Handbook of Statistics - Sample Surveys: Inference and Analysis* **Vol. 29B**. Pfeffermann D. and Rao C.R. (eds). North-Holland, pp. 103–120.

Chen, Q. 2007 Airborne lidar data processing and information extraction. *Photogramm. Eng. Remote Sens.* **73** (12), 1355–1365.

Chen, Q., Vaglio Laurin, G., Battles, J.J. and Saah, D. 2012 Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. *Remote Sens. Environ.* **121**, 108–117.

Chirici, G., McRoberts, R.E., Fattorini, L., Mura, M. and Marchetti, M. 2016a Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sens. Environ.* **174**, 1–9.

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O. and Waser, L.T., *et al.* 2016b A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sens. Environ.* **176**, 282–294.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D. and Maselli, F., *et al.* 2008 Non-parametric and parametric methods using satellite imagery for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sens. Environ.* **112**, 2686–2700.

COST. (2015) *COST Action FP1001 – USEWOOD: Improving Data and Information on the Potential Supply of Wood Resources*. Available at: https://sites.google.com/site/costactionfp1001/. Last accessed: February 2016.

Crookston, N.L., Finley, A.O., Coulston, J. 2015 yaImpute. Available at: https://cran.r-project.org/web/packages/yaImpute/index.html. Last accessed: June 2016.

Dudani, S.A. 1976 The distance-weighted k-Nearest-Neighbor rule. *IEEE Trans. Syst. Man Cybern B Cybern* **6** (4), 325–327.

Elisseeff, A. and Pontil, M. 2002. Leave-one-out and stability of learning algorithms with applications. Chapter 6 in Suykens, J., Horváth, G. Basu, S., Micchelli, C., and Vandewalle, J. (eds). Advances in learning theory: methods, models, and applications. IOS Press. pp. 111–1130.

Franco-Lopez, H., Ek, A.R. and Bauer, M.E. 2001 Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sens. Environ.* **77**, 251–274.

Gagliasso, D., Hummel, S. and Temesgen, H. 2014 Selected parametric and non-parametric imputation methods for estimating forest biomass and basal area. *Open J. For.* **4** (1), 42–48.

Gobakken, T. and 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. *Can. J. For. Res.* **38**, 1095–1109.

Harrell, F.E. 2001 *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag.

Holland, J.H. 1975 Adaptation in natural and artificial system. University of Michigan Press.

Holopainen, M., Haapanen, R., Karjalainen, M., Vastaranta, M., Hyyppä, J. and 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 Sens.* **2**, 432–445.

Langley, P. and Iba, W. 1993 Average-case analysis of a nearest neighbor algorithm. In: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*. Chambery, France, 28 Aug – 3 Sep 1993, pp. 889–894.

Latifi, H., Nothdurft, A. and Koch, B. 2010 Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. *Forestry* **83** (4), 395–407.

Lefsky, M.A., Harding, D., Cohen, W.B., Parker, G. and Shugart, H.H. 1999 Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. *Remote Sens. Environ.* **67**, 83–98.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. 2005 Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Stat. Trans.* **7** (3), 649–673.

LeMay, V. and Temesgen, H. 2005 Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Sci.* **51** (2), 109–119.

Mahalanobis, P.C. 1936 On the generalized distance in statistics. *Proc. Nat. Instit. Sci. India* **2**, 49–55.

Maltamo, J., Malinen, J., Kangas, A., Härkönen, S. and Pasanen, A.M. 2003 Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. *Forestry* **76**, 449–463.

Maltamo, M., Peuhkurinen, J., Malinen, J., Vauhkonen, J., Packalén, P. and Tokola, T. 2009 Predicting tree attributes and quality characteristics of Scots pine using airborne laser scanning data. *Silva Fenn.* **43** (3), 507–521.

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

McRoberts, R.E. 2009 Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sens. Environ.* **113**, 489–499.

McRoberts, R.E. 2012 Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manage.* **272**, 3–12.

McRoberts, R.E. and Westfall, J.A. 2014 The effects of uncertainty in model predictions of individual tree volume on large area volume estimates. *Forest Sci.* **60**, 34–43.

McRoberts, R.E., Nelson, M.D. and Wendt, D.G. 2002 Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. *Remote Sens. Environ.* **82**, 457–468.

McRoberts, R.E., Tomppo, E.O., Finley, A.O. and Heikkinen, J. 2007 Estimating areal means and variances using the k-nearest neighbors technique and satellite imagery. *Remote Sens. Environ.* **111**, 466–480.

McRoberts, R.E., Hansen, M.H., Smith, W.B. 2010 Development of the national forest inventory of the United States of America. In *National Forest Inventories: Pathways to Common Reporting*. Tomppo E., Gschwantner T., Lawrence M. and McRoberts R.E. (eds). Springer, 612 pp.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2013 Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sens. Environ.* **128**, 268–275.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2015 Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. *Remote Sens. Environ.* **163**, 13–22.

Miller, R. 1981 *Simultaneous Statistical Inference*. 2nd edn. Springer-Verlag, 299 pp.

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

Ohmann, J.L. and 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. *Can. J. For. Res.* **32**, 725–741.

Ohmann, J.L., Gregory, M.J. and Roberts, H.M. 2014 Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. *Remote Sens. Environ.* **151**, 3–15.

Packalén, P. and Maltamo, M. 2007 The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sens. Environ.* **109**, 328–341.

Packalén, P., Temesgen, J. and Maltamo, M. 2012 Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Can. J. Remote Sens.* **38** (5), 1–13.

Särndal, C.E. 2007 The calibration approach in survey theory and practice. *Surv. Methodol.* **33** (2), 99–119.

Särndal, C.-E. 2011 Combined inference in survey sampling. *Pakistan J. Statist.* **27** (4), 359–370.

Särndal, C.-E., Swensson, B. and Wretman, J. 1992 *Model Assisted Survey Sampling*. Springer.

Schaal, S., Vijayakumar, S., Atkeson, C.G. 1998 Local dimension reduction. In *Advances in Neural Information Processing Systems 10*. Jordan M.I., Kearns J.J. and Solla S.A. (eds). MIT Press, pp. 1–7.

Tabacchi, G., Di Cosmo, L., Gasparini, P. and Morelli, S. 2011 *Stima del volume e della fitomassa delle principali specie forestali italiane. Equazioni di previsione, tavole del volume e tavole della fitomassa arborea epigea*. Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di Ricerca per il Monitoraggio e la Pianificazione Forestale, Trento, 412 pp.

Tomppo, E. and Halme, M. 2004 Using coarse scale forest variables as ancillary information and weighting of k-NN estimation: a genetic algorithm approach. *Remote Sens. Environ.* **92**, 1–20.

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

Tomppo, E., Gschwantner, T., Lawrence, M. and McRoberts, R.E. (eds). 2010 National Forest Inventories: Pathways to Common Reporting, Springer, 612 pp.

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

Ver Hoef, J.M. and Temesgen, H. 2013 A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. *PLoS ONE* **8** (3), e59129.

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

Wilson, B.T., Lister, A.J. and Riemann, R.I. 2012 A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. *For. Ecol. Manage.* **271**, 182–198.

Zheng, H. and Little, R.J.A. 2004 Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Surv. Methodol.* **30**, 209–218.

# Appendix

## A1. Significance level for *F*-test

The significance level for the *F*-test used to select feature variables is only approximate. First, for linear regression models, the test assumes that the model with $m$–1 predictor variables is nested within the model with $m$ predictor variables in the sense that the set of $m$–1 predictor variables is completely included in the set of $m$ predictor variables. For $k$-NN applications, this criterion is not always satisfied. For example, for the Molise data set, none of the feature variables selected for $m = 3$ were included among those selected for $m = 4$. Second, the leave-one-out technique does not necessarily produce the same distributions of $SS_{res}$ as would a regression model. Third, the multiple applications of the test may require adjustment of the significance level as is done for statistical multiple comparisons analyses (Miller, 1981, Section 3.1). Nevertheless, the approach is an automated and objective technique for selecting feature variables.

The relatively small $\alpha = 0.01$ significance level was selected rather than larger values for three reasons: (1) to partially compensate for the stringent three-test selection criterion, (2) to partially compensate for multiple applications of the test and (3) to err on the side of selecting fewer feature variables because, as shown by McRoberts *et al.* (2015), with larger numbers of feature variables less of the optimization achieved in the reference set is actually realized in the target set.

## A2. *k*-NN software

Software for implementation of the $k$-NN algorithm was constructed specifically for this study to run in a generic Microsoft Windows environment. The algorithm was coded in Fortran for three reasons: (1) experience and familiarity, (2) availability of necessary modules and subroutines and (3) ease of optimization. The primary components are an input/output module, an optimization module, a distance calculation and neighbour selection subroutine, and a criterion (sum of squared errors) evaluation subroutine. Additional subroutines for matrix inversion and canonical correspondence analysis were obtained from free Internet-based Fortran software libraries. Because multiple aspects of the overall algorithm including neighbour weighting schemes and distance metrics have yet to reach sufficiently mature stages, the software is not available for distribution. However, some of the features of the algorithm including the canonical correspondence analysis distance metric, albeit not the Dudani neighbour weighing scheme, are already available in an R-based software package titled yaImpute (Crookston *et al.*, 2015).