Remote Sensing of Environment 184 (2016) 387–395

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Using genetic algorithms to optimize k-Nearest Neighbors configurations for use with airborne laser scanning data

Ronald E. McRoberts [a,*], Grant M. Domke [a], Qi Chen [b], Erik Næsset [c], Terje Gobakken [c]

[a] Northern Research Station, U.S. Forest Service, Saint Paul, MN, USA
[b] Department of Geography, University of Hawai`i at Mānoa, Honolulu, HI
[c] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form 2 June 2016
Accepted 7 July 2016
Available online 29 July 2016

Keywords:
Inference
Spatial estimation
National forest inventory

ABSTRACT

The relatively small sampling intensities used by national forest inventories are often insufficient to produce the desired precision for estimates of population parameters unless the estimation process is augmented with auxiliary information, usually in the form of remotely sensed data. The k-Nearest Neighbors (k-NN) technique is a non-parametric, multivariate approach to prediction that has emerged as particularly popular for use with forest inventory and remotely sensed data and has been shown to contribute substantially to increasing precision. k-NN predictions are calculated as linear combinations of observations for sample units that are nearest in a space of auxiliary variables to the population unit for which a prediction is desired. Implementation of a nearest neighbors algorithm requires four choices: (i) a distance metric, (ii) specific auxiliary variables to be used with the distance metric, (iii) the number of nearest neighbors, and a (iv) scheme for weighting the nearest neighbors. Regardless of the choices for a distance metric and weighting scheme, emerging evidence suggests that optimization of the technique, including selection of an optimal subset of auxiliary variables, greatly enhances prediction. However, optimization can be computationally intensive and time-consuming. A promising approach that is gaining favor is based on genetic algorithms, a technique that uses search heuristics that mimic natural selection to solve optimization problems.

The objective of the study was to compare optimized k-NN configurations with respect to inferences for mean volume per unit area using airborne laser scanning variables as auxiliary information. For two study areas, one in Norway and one in Minnesota, USA, the analyses focused on optimizing k-NN configurations that used the weighted Euclidean and canonical correlation distance metrics and two neighbor weighting schemes. Novel features of the study include introduction of a neighbor weighting scheme that has not previously been used for forestry applications, simultaneous optimization of all four k-NN choices, and basing comparisons on confidence intervals, rather than intermediate products such as prediction accuracies. Two conclusions were primary: (1) optimized selection of feature variables produced greater precision than using all feature variables, and (2) computational intensity necessary to optimize the weighted Euclidean metric was considerably greater than for the canonical correlation analysis metric. Specific findings were that optimization produced pseudo-R^2 as large as 0.87 for the Norwegian dataset and as large as 0.89 for the Minnesota dataset. For the optimized canonical correlation distance metric, widths of approximate 95% confidence intervals as proportions of the estimated means were as small as 0.13 for the Norwegian dataset and as small as 0.15 for the Minnesota dataset.

Published by Elsevier Inc.

1. Introduction

Among the parameters estimated by national forest inventories (NFI), the most common are related to forest area and growing stock volume (Lawrence et al., 2010). Because NFI sampling intensities are relatively small, precision requirements for these parameters often cannot be achieved apart from augmenting the estimation process with auxiliary information. For estimation of parameters related to forest area, auxiliary information in the form of remotely sensed spectral data from sensors such as Landsat have been demonstrated to be particularly effective (e.g., McRoberts, 2010). However, for parameters related to growing stock volume, spectral data are less effective which should be expected because spectral sensors respond primarily to light reflected from the top of the canopy. For the latter parameters, remotely sensed data from active sensors such as lidar have been demonstrated to be much more effective (Næsset et al., 2011; Strunk et al., 2012; d'Oliveira et al., 2012; McRoberts et al., 2013).

For estimation of both area and growing stock volume, the k-Nearest Neighbors (k-NN) technique has emerged as particularly popular for

* Corresponding author.
E-mail address: rmcroberts@fs.fed.us (R.E. McRoberts).

http://dx.doi.org/10.1016/j.rse.2016.07.007
0034-4257/Published by Elsevier Inc.

use with NFI and remotely sensed data. Chirici et al. (2016) conducted a meta-analysis and review of the k-NN literature and reported more than 250 forestry applications using remotely sensed data for 26 countries on six continents. The most commonly reported response variables were growing stock volume and closely related variables such as biomass and carbon.

The k-NN technique is a non-parametric approach to prediction. Population unit predictions are calculated as linear combinations of observations for sample units designated neighbors that are nearest or most similar in a space of auxiliary variables to units for which predictions are desired. Implementation of a nearest neighbors algorithm requires four choices: (i) a distance metric, (ii) specific auxiliary or feature variables to be used with the distance metric, (iii) the number of nearest neighbors, and (iv) a scheme for weighting the nearest neighbors. Chirici et al. (2016) found that no particular k-NN configuration could be considered optimal for all the cases, and therefore recommended an optimization phase for each application. Similarly, Packalén et al. (2012) reported that selection of optimal subsets of feature variables was beneficial.

Selection of optimal subsets of feature variables can be extremely computationally intensive when the number of such variables is large. Further, when the feature variables are highly correlated such as is the case with many sets of remotely sensed feature variables, stepwise selection methods are known to perform poorly (Harrell, 2001). For such applications, Tomppo and Halme (2004) introduced genetic algorithms (GA) as a technique for optimizing k-NN configurations, and subsequent investigations have shown them to have considerable potential for this purpose (McRoberts, 2012; McRoberts et al., 2015; Tomppo et al., 2009; Latifi et al., 2010).

GAs are iterative search heuristics that mimic natural selection to solve optimization problems (Holland, 1975). These algorithms start with a population of randomly generated individuals, compare them with respect to an optimization criterion, and combine the more optimal individuals to produce additional individuals. For use with k-NN, individuals are components of k-NN configurations including particular combinations of feature variables and/or elements of distance metrics. Despite their increasing popularity, there have been few reports of broad investigations of GAs for forestry applications.

The objective of the study was to compare optimized k-NN configurations obtained using GAs with respect to inferences in the form of confidence intervals for mean forest volume per unit area. Two study areas were used, one in Norway and one in Minnesota, USA, both with wall-to-wall airborne laser scanning (ALS) data used as k-NN feature variables. In addition to being among the first broad investigations of GAs for forestry applications, the novel components of the study include introduction of a neighbor weighting scheme not previously used for forestry applications, simultaneous optimization of all four k-NN choices, and basing comparisons on the final NFI product, confidence intervals, rather than intermediate products such as prediction accuracies.

2. Data

2.1. Overview

Two datasets with distinct features were used to permit a degree of generalization of the results. The important dataset distinctions are the Norwegian boreal biome and the Minnesota temperate biome, the species compositions, the minimum tree diameter threshold of 5 cm for the Norwegian dataset but the considerably larger threshold of 12.7 cm for the Minnesota dataset, the ground sampling designs, and the 250-m^2 Norwegian plots but the considerably smaller 168.3-m^2 Minnesota plots. Comparable, but yet still different features include the allometric volume models, ALS pulse densities, and ALS metrics. Details for the two datasets follow.

2.2. Hedmark, Norway

The 1259-km^2 study area was mostly in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway (Fig. 1) and was completely forested. Dominant tree species are 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), but restricted to the 145 plots inventoried by the Norwegian NFI between 2005 and 2007 and the geographic area represented by the corresponding portion of the Latin Square sampling design (Fig. 1). The volume of each sample tree with diameter of at least 5 cm was predicted using species-specific volume models with diameter at breast-height (dbh, 1.3 m) and either measured height or predicted height as independent variables (Braastad, 1966; Brantseg, 1967; Vestjordet, 1967). Volume predictions for individual trees were added to produce plot-level totals which were then scaled to a per unit area basis (m^3/ha) and considered to be observations without error (McRoberts and Westfall, 2014).

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, heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions of echoes with heights above 2 m were calculated and denoted $h_1, h_2, \ldots, h_{10}$, respectively. Canopy densities were calculated as the proportions of echoes with heights >0%, 10%, …, 90% of the range between 2 m above ground and the 95th height percentile and were denoted $d_0, d_1, \ldots, d_9$, respectively (Gobakken and Næsset, 2008). Næsset (2002) provides additional details for the dataset.

2.3. Itasca County, Minnesota, USA

The 7583-km^2 study area was located in north central Minnesota in the USA (Fig. 2) and was characterized as approximately 80% forest land. Land cover includes water, wetlands and forest consisting of upland deciduous mixtures of 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* (L.)), and black ash (*Fraxinus nigra* Marsh.). Data were obtained for plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the NFI of the USA. Field crews observe species and measure dbh (1.37 m, 4.5 ft) and height for all trees with dbh of at least 12.7 cm (5 in). Model predictions of individual tree volumes were aggregated to obtain plot-level volume predictions which were scaled to a per unit area basis (m^3/ha) and considered to be observations without error (McRoberts and Westfall, 2014). Data were used for 115 plots measured in 2014 because this was the only year for which GPS receivers with sub-meter accuracy were available. Further, data for only the 168.3-m^2 central subplot of the four subplot cluster were used to avoid issues of spatial correlation among subplot observations.

Wall-to-wall ALS data were acquired in April 2012 with a nominal pulse density of 0.67 pulses/m^2. Ground returns were classified by the provider and were 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 echo heights were constructed for the 168.3-m^2 plots and 169-m^2 square cells that tessellated the study area. Metrics for each plot and cell included the mean, standard deviation, skewness, kurtosis, quadratic mean height of the distributions of heights for all echoes (Lefsky et al., 1999; Chen et al., 2012). In addition, heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions were calculated, and canopy densities were calculated as the proportions of echoes with heights >0%, 10%, …, 90% of the range between 1.3 m above ground and the 95th height percentile (Gobakken and Næsset, 2008).

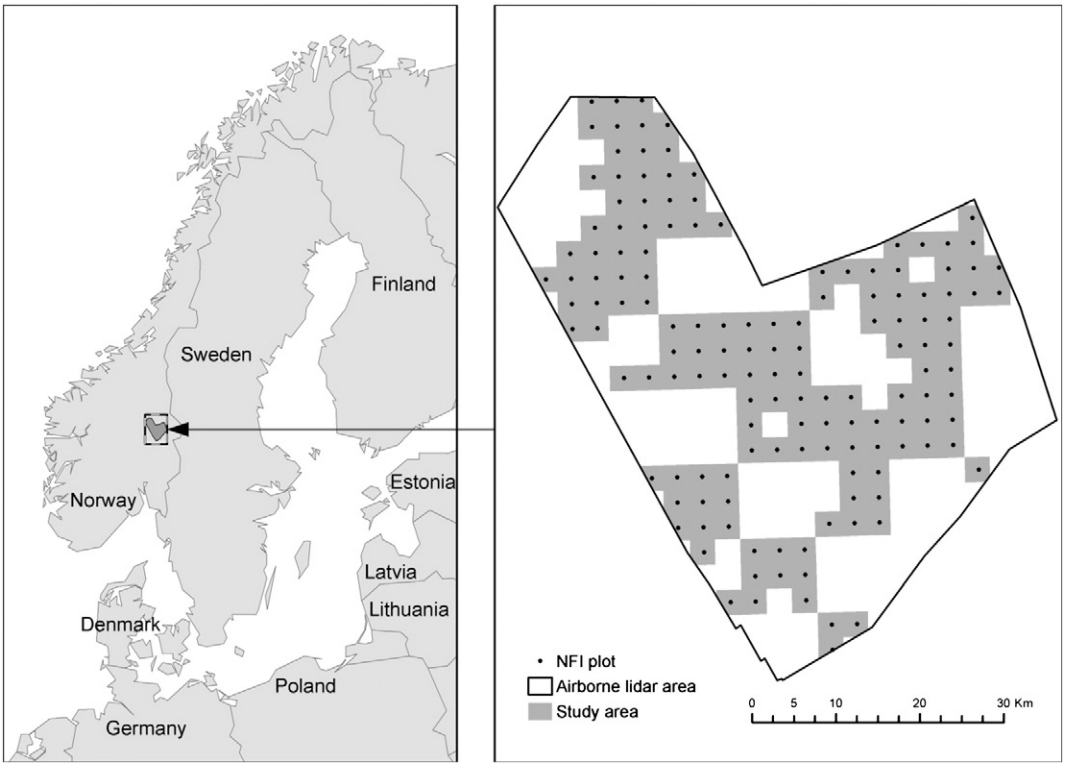

Fig. 1. Study area in Hedmark County, Norway. The shaded areas are the portion of the Norwegian NFI's Latin Square sampling design corresponding to plots measured in 2005–2007.

3. Nearest neighbors techniques

3.1. Terminology and notation

For notational purposes, **Y** commonly 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 neighbors techniques, the auxiliary variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the sample of sample 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. All population units for both the reference and target set are assumed to have a complete set of observations for all feature variables.

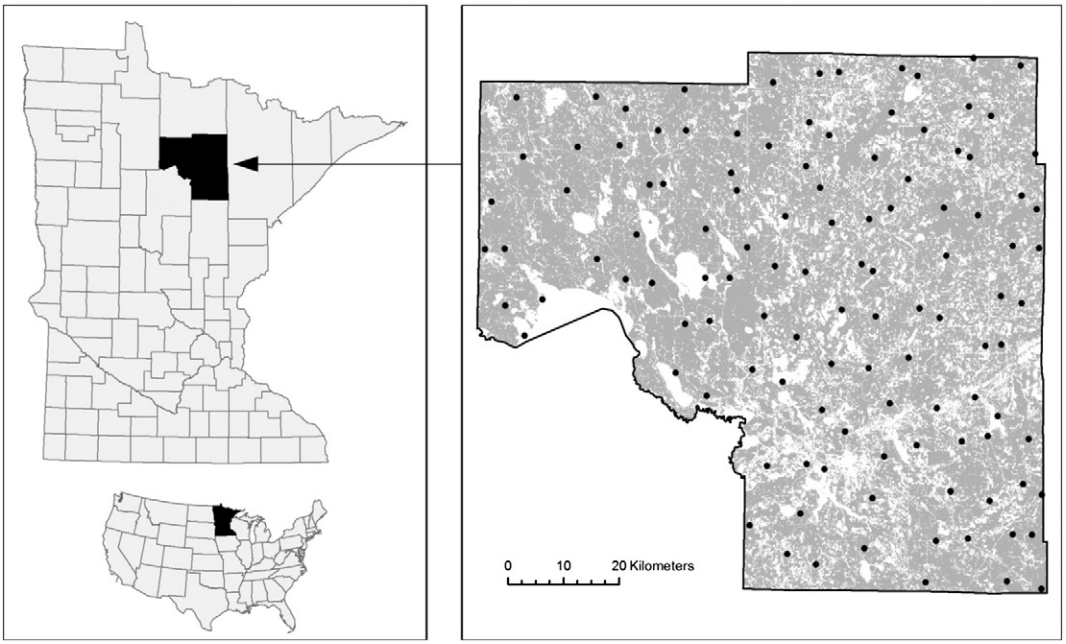

Fig. 2. Study area in Itasca County, Minnesota, USA.

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

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

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 unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the jth nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1$.

3.2. Neighbor weighting

3.2.1. t-weighting

One common neighbor weighting scheme designated *t-weighting* is to weight neighbors inversely to a power of the distance, d_{ij}, between the jth reference unit and the ith target unit,

$$w_{ij} = \frac{d_{ij}^{-t}}{W} \quad (2)$$

where $W = \sum_{j=1}^{k} d_{ij}^{-t}$ and $t \geq 0$. For small numbers of feature variables and/or large reference sets, $d_{ij} = 0$ may occur in which case Eq. (2) leads to computational errors. For this study, if $d_{ij} = 0$ for j = 1, …, k, then all distances are arbitrarily reset to 1, i.e., $d_{ij} = 1$ for all neighbors. If $d_{ij} = 0$ for j = 1, …, k' where k' < k, then all 0-distances are arbitrarily reset to half the smallest non-zero distance, i.e., for j = 1, …, k',

$$d_{ij} = \frac{d_{ik'+1}}{2}. \quad (3)$$

For many applications, the default value of $t = 0$, meaning equal weighting of neighbors, is selected; otherwise, $t = 1$or $t = 2$ is usually selected, although there is no reason integer values must be used. Only a few reports of efforts to optimize selection of t are known (Wilson et al., 2012; McRoberts, 2012; McRoberts et al., 2015).

3.2.2. d-Weighting

Dudani (1976) proposed a weighting scheme that bases the weight for the jth neighbor on the ratio of two distances, the distance between the jth and kth neighbors and the distance between the first and the kth neighbors. Because the scheme assigns weight of 0 to the kth neighbor, the scheme was modified for this study to,

$$w_{ij} = \frac{\frac{d_{ik+1} - d_{ij}}{d_{ik+1} - d_{i1}}}{W} \quad (4)$$

where $W = \sum_{j=1}^{k} \frac{d_{ik+1} - d_{ij}}{d_{ik+1} - d_{i1}}$. For notational purposes, this scheme is characterized as *d-weighting*. Note that with the formulation of Eq. (4), calculation of weights for k neighbors requires distances for k + 1 neighbors. 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', …, k. For this study, if $d_{i1} = d_{ik+1}$, then all distances are reset to 1, i.e., $d_{ij} = 1$ for all neighbors. If $d_{ij} = d_{ik+1}$ for j = k', …, k where 1 < k' < k + 1, then all such distances are reset to the mean of the k + 1st distance and the greatest distance that differs from the k + 1st distance,

$$d_{ij} = \frac{d_{ik'-1} + d_{ik+1}}{2} \quad (5)$$

for j = k', …, k. For example, suppose k = 5 and the six smallest distances for the ith target unit are $d_{i1} = 1$, $d_{i2} = 2$, $d_{i3} = 3$ $d_{i4} = 5$,$d_{i5} = 5$,$d_{i6} = 5$. Using only Eq. (4), the weights for the fourth and fifth neighbors would be $w_{i4} = w_{i5} = 0$ which would exclude the observations from the fourth and fifth neighbors when calculating the k-NN prediction. However, by applying Eq. (5), the fourth and fifth distances are reset to $d_{i4} = d_{i5} = ½ \cdot (d_{i3} + d_{i4}) = 4.0$ with the result that W = 2.75 and $w_{i1} = 0.364$, $w_{12} = 0.272$, $w_{i3} = 0.182$, $w_{i4} = 0.091$, and $w_{i5} = 0.091$. No previous reports of the use of d-weighting for forestry applications are known.

3.3. Distance metrics

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

$$d_{ij} = \sqrt{(X_i - X_j)' M (X_i - X_j)} \quad (6)$$

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; when **M** is the inverse of the covariance matrix for the feature variables, Mahalonobis distance results. Other popular metrics such as those based on canonical correlation analysis (Moeur and Stage, 1995; Maltamo et al., 2003; LeMay and Temesgen, 2005) and canonical correspondence analysis (Wilson et al., 2012; Ohmann et al., 2014) can also be expressed in matrix form.

For this study, only metrics that can be readily expressed in matrix form were considered. Further, based on a comprehensive review of nearest neighbors configurations (Chirici et al., 2016), distance metrics were further restricted to the weighted Euclidean and canonical correlation analysis metrics.

3.3.1. Weighted Euclidean metric

With the *Weighted Euclidean distance metric* (WEUCL), the matrix **M** from Eq. (6) is a non-identity diagonal matrix, **D**, and distance is expressed as,

$$d_{ij} = \sqrt{(X_i - X_j)' D (X_i - X_j)}. \quad (7)$$

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

3.3.2. Canonical correlation analysis metric

With the *canonical correlation analysis distance metric* (CCA), a system of linear models is solved to obtain estimates of coefficient vectors, **α** and **β**, that maximize the correlation between $U = \alpha_1 \cdot Y_1 + \ldots + \alpha_p \cdot Y_p$ and $V = \beta_1 \cdot X_1 + \ldots + \beta_q \cdot X_q$ where Y_j denotes the jth response variable, X_j denotes the jth 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 **Γ**, and the corresponding eigenvalues, also designated canonical correlations, are denoted λ. Feature space distances with this metric are expressed as,

$$d_{ij} = \sqrt{(X_i - X_j)' \Gamma \Lambda^2 \Gamma' (X_i - X_j)} \quad (8)$$

where the elements of the diagonal matrix, **Λ**, are the squares, λ^2, of the canonical correlations. The CCA metric was first proposed by Moeur and Stage (1995) who used only a single neighbor, but it has also been used with multiple neighbors (Maltamo et al., 2003, 2009; Packalén and Maltamo, 2007).

3.4. Feature variables

Feature variables that are unrelated to the response variables introduce randomness into distance calculations and thereby contribute to selection of spurious neighbors and less accurate predictions. Langley and Iba (1993) and Blum and Langley (1997) characterize such feature variables as *irrelevant*. Metrics such as WEUCL and CCA theoretically select negligible weights for irrelevant feature variables and thereby minimize their adverse effects. Nevertheless, Packalén et al. (2012) reported detrimental effects of not eliminating such variables altogether. Thus, one approach to optimization is to identify and then eliminate irrelevant feature variables. Multiple methods may be considered for accomplishing this task. Stepwise selection techniques have been used successfully for this purpose for some problems. However, as noted by Harrell (2001, pp. 64–65) stepwise methods are considerably less effective when the feature variables are strongly correlated such as is the case with large numbers of ALS metrics. McRoberts (2012) and McRoberts et al. (2015) evaluated all combinations of all numbers of feature variables and selected the combination that optimized a criterion of interest or the combination with the smallest number of feature variables beyond which little or no statistically significant optimization was realized. Although effective, these approaches can be extremely computationally intensive, particularly for large numbers of feature variables and/or large reference sets. Packalén et al. (2012) compared simulated annealing and random forests for selecting feature variables and reported that the most effective technique depended on the particular response variable.

3.5. Genetic algorithms

GAs are search heuristics that mimic natural selection to solve optimization problems (Holland, 1975). GAs have multiple advantages over other optimization algorithms: first, they can deal with very large search spaces such as the large numbers of combinations of feature variables or large dimension continuous spaces of weights for feature variables; second, GAs do not remain trapped in local sub-optimal solutions; third, GAs do not require starting values; and fourth, GAs can deal with objective functions such as k-NN sum of squared errors surfaces that are neither smooth nor continuous with respect to variable weights and thereby preclude gradient and derivative methods. The only serious disadvantage is that optimal solutions are not guaranteed. Tomppo and Halme (2004) first proposed GAs to select feature variable weights for use with the WEUCL metric. Subsequent research has shown the approach to have considerable promise (McRoberts, 2012; McRoberts et al., 2015; Tomppo et al., 2009; Latifi et al., 2010). However, GAs are not known to have been used in two-step procedures for which the first step is to select feature variables and the second step is to optimize weighting of the selected variables using metrics such as WEUCL or CCA.

GAs are iterative and start from a population of randomly generated individuals, each consisting of multiple genomes, with the population in each iteration called a generation. For this k-NN application, a genome is a value for one element of a diagonal distance matrix, and a full set of genomes or diagonal values constitutes an individual. In each generation, each individual in the population is evaluated with respect to its fitness which, for k-NN applications, is typically a criterion related to the sum of squared errors for continuous response variables or classification accuracy for categorical response variables. Each subsequent generation consists of the most fit individuals from the previous generation, modifications of them, and a small number of new randomly generated individuals representing immigration into the population. Modifications take multiple forms: (i) *cross-over* consisting of combinations of pairs of individuals obtained by randomly selecting one genome from each pair to construct a new individual, (ii) *mutation* consisting of small random perturbations of randomly selected genomes for individuals, and (iii) *immigration* consisting of randomly generated new individuals. The new generation of individuals is then used in the next iteration of the algorithm. The algorithm terminates following a maximum number of generations or when a satisfactory fitness level has been reached for at least one individual in the population (Fig. 3).

Implementation of a GA requires choices for multiple parameters. First, because GAs ensure only near optimal solutions, the algorithm should be independently replicated NREP times with entirely new randomly generated initial populations. Within each replication, the number of individuals, designated the population size and denoted NPOP, must be selected. For a given starting population, the algorithm iterates through multiple generations denoted NGEN. The number of most fit individuals from a previous generation selected to enter the next population is designated NFIT. For this study, subsequent populations consisted of four components: (i) the NFIT individuals from the previous generation, (ii) $\text{NCROSS} = \text{NFIT} \cdot (\text{NFIT} - 1)$ crossovers obtained by combining each of the NFIT individuals with the NFIT-1 other individuals, (iii) NMUT mutations obtained by randomly increasing or decreasing the value for one randomly selected genome for randomly selected individuals from among the NFIT most fit individuals from the previous generation; and (iv) NIMM immigrants constructed by randomly selecting values for each genome for new individuals. Values selected for these parameters are discussed in Section 4.1.

GAs can be used in two modes. The objective of the first mode is simply to identify and eliminate irrelevant feature variables without regard to the metric with which the remaining variables will be used. The objective of the second mode is to select positive values for the diagonal matrix for use with the WEUCL metric.

3.6. Assessments

3.6.1. Prediction accuracy

Prediction techniques can be evaluated at either the individual population unit level or at the overall population level. At the individual population unit level, common measures of prediction accuracy for continuous response variables are typically in terms of squared deviations between sample observations and their corresponding predictions calculated as,

$$SS_{dev} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

or equivalently as a pseudo-R^2,

$$R^{2^*} = \frac{SS_{mean} - SS_{dev}}{SS_{mean}} \tag{10}$$

where SS_{mean} is the sum of squared deviations of reference set observations from their mean. For this study, the accuracy of individual population unit predictions was assessed using the popular leave-one-out method whereby the prediction for each reference set unit was calculated using only the observations for the other reference set units (Elisseeff and Pontil, 2002).

3.6.2. Population parameter inference

Although maps and prediction accuracies for sample units are of inherent interest, increasingly they are recognized as only intermediate steps enroute to an inference in the form of a confidence interval for a population parameter such as mean volume per unit area. Confidence intervals can be expressed as,

$$\hat{\mu} \pm t_{1-\alpha/2} \cdot SE(\hat{\mu}) \tag{11}$$

where $\hat{\mu}$ is the estimate of the mean per unit area, $SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})}$ is the standard error of $\hat{\mu}$, $t_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of Student's t-distribution, and α is the significance level. For this study, the ultimate focus of the analyses was estimation of the population mean and the

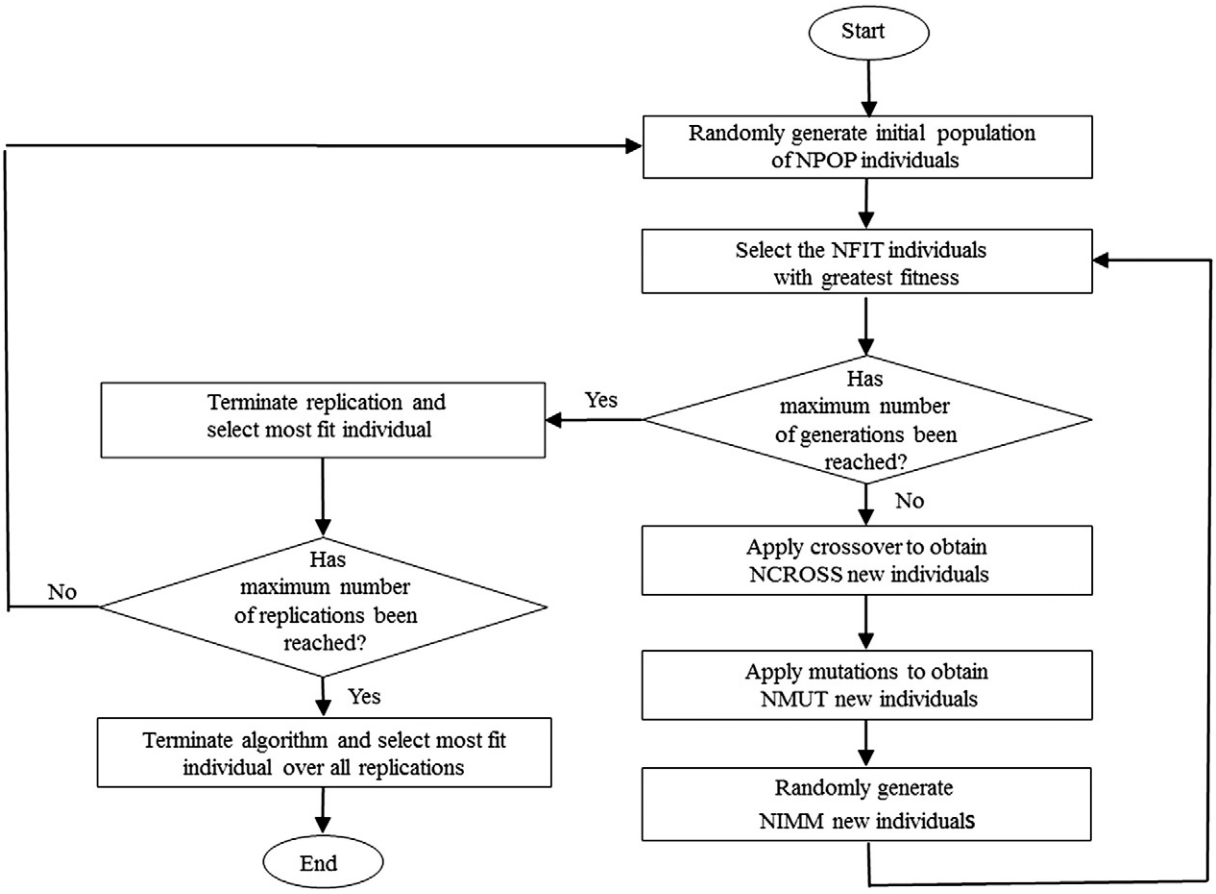

Fig. 3. Flow chart for genetic algorithm.

SE of the estimated mean using simple random sampling and model-assisted regression estimators.

The *simple random sampling* (SRS) estimator of the population mean is,

$$\hat{\mu}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{12}$$

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,

$$\hat{Var}(\hat{\mu}_{SRS}) = \frac{1}{n\cdot(n-1)}\sum_{i=1}^{n}(y_i - \hat{\mu}_{SRS})^2. \tag{13}$$

For systematic samples, as used for this study, variances may be slightly overestimated relative to estimates based on a simple random sample (Särndal et al., 1992, p. 83). 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 use models based on auxiliary data to enhance inferences but rely on the probability sample for validity (Särndal et al., 1992). An initial estimator of the population mean is formulated as,

$$\hat{\mu}_{Init} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i \tag{14}$$

where N is the target set (population) size and $\hat{y}_i$ is the k-NN prediction for the ith population unit. Systematic prediction errors induce bias into this estimator which can be estimated as,

$$\hat{Bias}(\hat{\mu}_{Init}) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \tag{15}$$

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

$$\begin{aligned}\hat{\mu}_{GREG} &= \hat{\mu}_{Init} - \hat{Bias}(\hat{\mu}_{Init}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\end{aligned} \tag{16}$$

with variance estimator,

$$\hat{Var}(\hat{\mu}_{GREG}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2, \tag{17}$$

where $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$ (Särndal et al., 1992; Särndal, 2011). Despite use of the term *regression* in the label characterizing the GREG estimators, multiple prediction techniques other than linear models have been used with the GREG estimators (Breidt and Opsomer, 2000, 2009; Lehtonen et al., 2005; Särndal, 2011; Zheng and Little, 2004).

The utility of auxiliary information such as the ALS metrics can be evaluated using relative efficiency (RE) calculated as,

$$RE = \frac{\hat{Var}(\hat{\mu}_{SRS})}{\hat{Var}(\hat{\mu}_{GREG})} \tag{18}$$

and interpreted as the factor by which the sample size would have to be

increased to achieve the same precision using the SRS estimators without the auxiliary information as was achieved using the GREG estimators with the auxiliary information. For each combination of dataset, distance metric, and weighting scheme, $\hat{\mu}_{SRS}$, $SE(\hat{\mu}_{SRS})$, $\hat{\mu}_{GREG}$, $SE(\hat{\mu}_{GREG})$, 2-SE confidence intervals, and RE were calculated.

3.7. Analyses

The analyses were conducted in three steps. In the first step, GAs were used in the first mode to identify and eliminate irrelevant feature variables for each combination of dataset, distance metric and weighting scheme. For this application, GA individuals consisted of genomes, one for each feature variable, with values of either 0 or 1 where 0 indicated the variable was not selected and 1 indicated it was selected. For the t-weighting scheme, the leave-one-out cross-validation technique was used to evaluate the fitness for each individual, including selecting the values of k and t that minimized SS_{dev}, and for the d-weighting scheme, the leave-one-out cross-validation technique was also used to evaluate fitness including selecting the value of k that minimized SS_{dev}. Values of k were permitted to range from 1 to 35, and values of t were permitted to range from 0 to 6.0.

The second step was used only for the WEUCL metric and entailed using the GA in the second mode to select weights for each of 50 combinations of feature variables with the smallest values of SS_{dev} from the first step. A 51st combination that included all feature variables was also evaluated in the second step as a means of assessing the detrimental effects of not optimizing the selection of feature variables. GA individuals consisted of genomes corresponding to the weights in the WEUCL metric, one for each feature variable, with values in the interval [0,1]. As for the first step, the leave-one-out cross-validation technique was used to evaluate fitness including selecting the values of k and t. When using all feature variables for the 51st combination, successful application of the GA in the second step required that genomes for one of the individuals in the initial population be set to the proportions of times the respective feature variables were selected among the NREP replications in the first step.

In the third step, confidence intervals were constructed using both the SRS and GREG estimators. For the SRS estimators, the estimates were based only on the reference data for the two datasets. Confidence intervals were also constructed using the GREG estimators with selected feature variable combinations for both datasets, both metrics and both neighbor weighting schemes. For the CCA metric, confidence intervals were constructed for the feature variable combination from the first step with the smallest SS_{dev} and also for the combination that included all feature variables. For the WEUCL metric, confidence intervals were constructed for three feature variable combinations with their respective weights from the second step: (i) the combination with the smallest SS_{dev} from the second step, (ii) the combination with the smallest SS_{dev} from the first step and with weights selected in the second step, and (iii) the combination with all feature variables and with the weights from the second step.

4. Results and discussion

4.1. Inventory consequences

For both datasets, estimates of the population mean of volume per unit area, $\hat{\mu}_{MA}$, were generally similar; precision as represented by standard errors, $SE(\hat{\mu}_{MA})$, as proportions of $\hat{\mu}_{MA}$ were less than 0.05; and maximum REs for optimized k-NN configurations were between 7.53 and 9.00. For comparison purposes, these RE values are substantially greater than RE = 5.36 and RE = 6.38 for two Norwegian datasets (Næsset et al., 2011; McRoberts et al., 2013), RE = 3.66 for a Brazilian dataset (d'Oliveira et al., 2012), and RE = 3.76 for a Washington, USA dataset (Strunk et al., 2012), all of which were obtained using regression models with similar ALS metrics.

Importantly for inventory applications, RE can be interpreted as the factor by which the sample size for use with the SRS estimators would have to be increased to achieve the same precision as was achieved using the GREG estimators with the current sample size. Given the current considerable expense associated with any kind of ground sampling, REs as large as 7.53 are substantial and have the potential to greatly enhance NFI estimation.

4.2. Genetic algorithm performance

The GA approach to first-step identification and elimination of feature variables for both the WEUCL and CCA metric worked well, as did the approach for the second-step selection of weights for the WEUCL metric. GA parameter values were subjectively selected, albeit following experimentation: NFIT = NMUT = NIMM = 10, which produces $NCROSS = NFIT \cdot (NFIT - 1) = 90$ and $NPOP = NFIT + NCROSS + NMUT + NIMM = 120$. Graphs of maximum R^{2*} versus both numbers of replications and numbers of generations were constructed to ensure that both NREP = 50 and NGEN = 50 were sufficiently large.

4.3. Prediction accuracy

For the Hedmark dataset, the greatest R^{2*} values for the WEUCL metric resulting from the first-step of application of the GAs were 0.75 for the t-weighting scheme and 0.74 for the d-weighting scheme (Table 1). R^{2*} values when using all features variables were smaller by factors of approximately 0.13. For the CCA metric, the greatest R^{2*} values for both weighting schemes were approximately 0.87 which were only slightly larger than when using all feature variables.

For the Itasca dataset, the greatest first-step R^{2*} values for the WEUCL metric were 0.85 for the t-weighting scheme and 0.84 for the d-weighting scheme, both of which were smaller by factors of approximately 0.07 than when using all feature variables (Table 2). For the CCA metric, first step R^{2*} values were 0.89 for both neighbor weighting schemes which was larger by factors of approximately 0.05 than when using all feature variables.

Two patterns were apparent in these results. First, fewer variables were selected for the WEUCL metric, regardless of the neighbor weighting scheme, and second, R^{2*} values were larger for the CCA metric, again regardless of the neighbor weighting scheme. Both patterns can be attributed to the feature variable weighting inherent in the CCA metric which permits marginally irrelevant feature variables to have nearly negligible weights.

The second-step application of the GAs, which selected weights for the variables selected in the first-step, was used only for the WEUCL metric. The second-step produced R^{2*} values that were greater than first-step values by factors of 0.10 to 0.13 for the Hedmark dataset but only by factors less than 0.04 for the Itasca dataset. For both datasets and neighbor weighting schemes, the combinations of feature variables that produced the greatest first-step R^{2*} values were not the combinations that produced the greatest second-step R^{2*} values, although the R^{2*} differences were small. In addition, the greatest R^{2*} values obtained in the second-step following elimination of irrelevant feature variables in the first-step were greater by factors of 0.12 to 0.16 for the Hedmark dataset but by factors less than 0.04 for the Itasca dataset.

Overall, the greatest R^{2*} values, regardless of metric or neighbor weighting scheme, were generally similar. For these similar values, the CCA metric produced slightly greater values as did the t-weighting scheme, although the differences for the two weighting schemes were quite small. An additional advantage of the CCA metric is that no second-step application of the GAs was necessary. First-step identification and elimination of irrelevant features variables followed by second-step selection of variable weights produced substantially greater R^{2*} values for the WEUCL metric than when using all feature variables, even following second-step selection of variable weights. For the CCA

Table 1
Results for Hedmark study area.

Selection of feature variables				Selection of feature variable weights			Mean volume per unit area population estimates (m^3/ha)			
Rank	No. variables	R^{2*}	Generation	Rank	R^{2*}	Generation	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$	CI[a]	RE
WEUCL/d-weighting										
1	8	0.74	7	7[b]	0.82	17	77.13	3.07	70.99–83.27	5.87
10[c]	3	0.72	39	1	0.83	16	74.64	3.05	68.54–80.74	5.97
–	All 20	0.64	–	51	0.69	1	83.69	4.16	69.63–92.01	3.21
WEUCL/t-weighting										
1	10	0.75	39	26[b]	0.82	33	78.79	3.18	72.43–85.15	5.50
6[c]	3	0.74	5	1	0.84	48	76.18	2.95	70.28–82.08	6.40
–	All 20	0.65	–	51	0.74	37	82.60	3.80	75.00–90.20	3.85
CCA/d-weighting										
1	18	0.87	7	–	–	–	84.27	2.72	78.83–89.71	7.53
–	All 20	0.84	–	–	–	–	84.99	3.02	78.95–91.03	6.08
CCA/t-weighting										
1	18	0.87	1	–	–	–	84.32	2.66	79.00–89.64	7.85
–	All 20	0.84	–	–	–	–	85.63	2.97	79.69–91.57	6.31

[a] 2-SE confidence interval.
[b] Second-phase results for variable combination with greatest first-phase R^{2*}.
[c] First-phase results for feature variable combination with subsequent greatest second-phase R^{2*}.

metric, the first-step application of GAs to select feature variables produced greater R^{2*} values, but not to the same degree as for the WEUCL metric.

4.4. Population inference

For each dataset, the four combinations of metrics and neighbor weighting schemes produced similar estimates for the population mean, $\hat{\mu}_{GREG}$, with the possible exception of the WEUCL metric for the Hedmark dataset (Tables 1, 2) for which no reason is apparent. Bias estimates, $\hat{Bias}(\hat{\mu}_{Init})$, as proportions of $\hat{\mu}_{GREG}$ were small, less 0.04 for all combinations; standard errors, $SE(\hat{\mu}_{GREG})$, as proportions of $\hat{\mu}_{GREG}$ were also small, less than 0.05; and all confidence intervals overlapped.

Relative efficiencies, RE, for optimized first-step selection of feature variables and second-step selection of variable weights for the WEUCL metric were in the range of 5.97 to 6.40 for the Hedmark dataset and in the range 6.71 to 7.40 for the Itasca dataset. For the CCA metric, REs for the optimized first-step selection of feature variables were 7.53 to 7.85 for the Hedmark dataset and 8.94 to 9.00 for the Itasca dataset. Overall, for the same dataset and neighbor weighing scheme, the CCA metric produced greater RE values than the WEUCL metric by factors of 0.18 to 0.25. For the optimized configurations, the t-weighting scheme produced greater RE values than the d-weighting scheme by factors of 0.07 and 0.09 for the WEUCL metric and by factors of 0.01 and 0.04 for the CCA metric.

For all combinations of datasets, metrics, and neighbor weighting schemes, use of all feature variables, even with second-step weighting for the WEUCL metric and the inherent weighting with the CCA metric, produced RE values that were smaller than the values for optimized configurations by factors in the range 0.18 to 0.46. This result confirms the recommendation of Chirici et al. (2016) that an optimization phase should be used when applying the k-NN technique and the particular finding of Packalén et al. (2012) that optimization should include selection of a subset of feature variables.

5. Conclusions

Six conclusions were drawn from the study. First, optimized k-NN configurations have the potential to greatly increase the precision of

Table 2
Results for Itasca study area.

Selection of feature variables				Selection of feature variable weights			Mean volume per unit area population estimates (m^3/ha)			
Rank	No. variables	R^{2*}	Generation	Rank	R^{2*}	Generation	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$	CI[a]	RE
WEUCL/d-weighting				*WEUCL/d-weighting*						
1	10	0.84	23	21[b]	0.84	1	50.34	2.38	44.58–56.10	6.28
4[c]	11	0.83	15	1	0.85	16	51.60	2.30	47.00–56.20	6.71
–	All 25	0.78	–	51	0.82	8	53.30	2.56	48.18–58.42	5.43
WEUCL/t-weighting				*WEUCL/t-weighting*						
1	11	0.85	18	9[b]	0.86	10	52.34	2.25	47.84–56.84	6.99
32[c]	14	0.83	8	1	0.86	14	53.89	2.19	49.51–58.27	7.40
–	All 25	0.78	–	51	0.83	12	51.05	2.42	46.21–55.89	6.04
CCA/d-weighting										
1	20	0.89	6	–	–	–	52.88	1.99	48.90–57.72	8.94
–	All 25	0.85	–	–	–	–	53.74	2.30	49.14–58.34	6.73
CCA/t-weighting										
1	18	0.89	1	–	–	–	53.10	1.99	49.12–57.08	9.00
–	All 25	0.84	–	–	–	–	53.82	2.36	49.06–58.58	6.38

[a] 2-SE confidence interval.
[b] Second-phase results for variable combination with greatest first-phase R^{2*}.
[c] First-phase results for feature variable combination with subsequent greatest second-phase R^{2*}.

population estimates of NFI population parameters. Second, genetic algorithms are an effective and computationally effective approach for identifying and eliminating irrelevant k-NN feature variables, regardless of the distance metric and for selecting feature variable weights for the weighted Euclidean distance metric. Third, identifying and eliminating irrelevant feature variables substantially improved both the accuracy of predictions and the precision of estimates of population means for both metrics, although more so for the weighted Euclidean metric than for the canonical correlation analysis metric. Fourth, the canonical correlation analysis distance metric produced greater prediction accuracies and greater precision for population estimates than the weighted Euclidean metric, and with considerably less computational effort. However, caution should be exercised when generalizing this finding for other datasets and other response variables. Fifth, the t-weighting scheme for weighting neighbors produced slightly greater prediction accuracies and precision for population estimates than the Dudani (d-weighting) scheme. However, because optimization of the t-weighting scheme entails considerably more computational effort, the Dudani scheme merits additional consideration. Sixth, considerable gain may be realized by simultaneously optimizing all four k-NN selections: the distance metric, the feature variables used with the distance metric, the number of nearest neighbors, and the neighbor weighting scheme. The overall conclusion is that the k-NN technique is now sufficiently mature and optimization techniques have been sufficiently demonstrated that users who do not optimize must justify their decision.

Several aspects of the methods demonstrated for this study merit additional consideration. First, because genetic algorithms do not guarantee optimal solutions, a comparison of genetic algorithm and exhaustive search solutions would be appropriate. For this study, however, genetic algorithm solutions produced solutions that were comparable to regression solutions. Second, demonstration of the utility of the combination of genetic algorithms and nearest neighbors techniques for more tropical and more complex temperate forest ecosystems would be welcome. Finally, one of most appealing features of nearest neighbors techniques is their multivariate capability. Thus, a multivariate application for a suite of forest attributes such as volume, stem density, and mean height would be a useful extension of this study.

References

Blum, A.L., 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, pp. 265–365 (In Norwegian with English summary).

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

Breidt, F.J., 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: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of Statistics - Sample Surveys: Inference and Analysis Vol. 29B. North-Holland, The Netherlands, 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., 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., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T., Travaglini, D., McRoberts, R.E., 2016. 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.

d'Oliveira, M.V.N., Reutebuch, S.E., McGaughey, R.J., Andersen, H.-E., 2012. Estimating forest biomass and identifying low-intensity logging areas using airborne scanning lidar in Antimary State Forest, Acre State, Western Brazilian Amazon. Remote Sens. Environ. 124, 479–491.

Dudani, S.A., 1976. The distance-weighted k-Nearest-Neighbor rule. IEEE Transactions on SMC-6(4), pp. 325–327.

Elisseeff, A., Pontil, M., 2002. Leave-one-out error and stability of learning algorithms with applications. *Advances in Learning Theory: Methods, Models, and Applications*. NATO Advanced Study Institute on Learning Theory and Practice, pp. 111–1130.

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. 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, New York.

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

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

Latifi, H., Nothdurft, A., 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.

Lawrence, M., McRoberts, R.E., Tomppo, E., Gschwantner, T., Gabler, K., 2010. Comparisons of national forest inventories. In: Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (Eds.), National Forest Inventories: Pathways to Common Reporting. Springer, Dordrecht, pp. 19–32.

Lefsky, M.A., Harding, D., Cohen, W.B., Parker, G., 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., Veijanen, A., 2005. Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. Stat. Transition 7 (3), 649–673.

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

Maltamo, J., Malinen, J., Kangas, A., Härkönen, 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., Tokola, T., 2009. Predicting tree attributes and quality characteristics of Scots pine using airborne laser scanning data. Silva Fennica 43 (3), 507–521.

McRoberts, R.E., 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sens. Environ. 114, 1017–1025.

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

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

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

McRoberts, R.E., Næsset, E., 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.

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

Næsset, E., 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sens. Environ. 80, 88–99.

Næsset, E., Gobakken, T., Solberg, S., Gregoire, T.G., Nelson, R., Ståhl, G., Weydahl, D., 2011. Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: a case study from a boreal forest area. Remote Sens. Environ. 115, 3599–3614.

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 Sens. Environ. 151, 3–15.

Packalén, P., 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., 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., 2011. Combined inference in survey sampling. Pak. J. Stat. 27, 359–370.

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

Strunk, J.L., Reutebuch, S.E., Andersen, H.-E., Gould, P.J., McGaughey, R.J., 2012. Model-assisted forest yield estimation with light detection and ranging. West. J. Appl. For. 27 (2), 53–59.

Tomppo, E., 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., McRoberts, R.E., 2009. Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. Remote Sens. Environ. 113, 500–517.

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

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

Wilson, B.T., Lister, A.J., 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. Manag. 271, 182–198.

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