Can. J. Remote Sensing, Vol. 38, No. 5, pp. 1–13, 2012

Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory

Petteri Packalén, Hailemariam Temesgen, and Matti Maltamo

Abstract. We examined the problem of selecting predictor variables for Nearest Neighbor (NN) imputation in remote sensing based forest inventory. Eighty-three variables were calculated from Airborne Laser Scanning data and aerial images, with responses being either dominant height or a set of five common stand attributes. Three different approaches were compared with select predictor variables. Analyses were repeated with three different NN imputation methods using a varying number of predictor variables. Results indicated that variable selection is justified, but it must be done properly. The most accurate method to select predictors was to minimize error using Simulated Annealing. For a single response, the most accurate imputation method was Random Forest proximity matrix-based imputation, whereas Most Similar Neighbor was the most accurate for five responses. An optimization-based distance metric also worked well. We also examined the degree to which different imputation methods are prone to overfitting as well as how to properly do cross-validation in NN imputation.

Résumé. On a examiné la problématique de la sélection des variables prédictives dans la procédure d'imputation par la méthode du plus proche voisin dans le contexte des inventaires forestiers réalisés par télédétection. Quatre-vingt trois variables ont été calculées à partir de données SLA (scanneur laser aéroporté) et d'images aériennes, les réponses étant soit la hauteur dominante ou un ensemble de cinq attributs courants de peuplement. Trois approches différentes ont été comparées pour la sélection des variables prédictives. Les analyses ont été répétées à l'aide de trois méthodes différentes d'imputation par le plus proche voisin en utilisant un nombre variable de variables prédictives. Les résultats ont montré que la sélection variable est justifiée, mais que celle-ci doit être faite correctement. La méthode la plus précise pour sélectionner les variables prédictives consistait à minimiser l'erreur à l'aide de la technique de recuit simulé. Pour une réponse unique, la méthode d'imputation la plus précise était l'imputation basée sur la matrice de proximité de type « Random Forest » (forêt aléatoire) alors que la méthode la plus précise pour les cinq réponses était la méthode d'imputation par le voisin le plus semblable « Most Similar Neighbor ». Une mesure de distance basée sur une méthode d'optimisation a également donné de bons résultats. On a aussi étudié la propension des différentes méthodes d'imputation au sur-ajustement de même que la façon d'exécuter correctement une validation croisée dans le contexte de l'imputation par le plus proche voisin.
[Traduit par la Rédaction]

Introduction

Background

A large set of predictor variables can be calculated from remote sensing data. This poses an issue of how to select the optimal set of predictor variables to be included in a model. Hocking (1976) asserted that improved computing capacity made the problem of variable selection in linear regression an active area of research in the late 70s. Although computers today are tremendously more powerful than in 1976, the problem still exists. In addition, variable selection is not just an issue in linear regression. The same problem exists when selecting predictor variables to be included in nonparametric models as well.

Stepwise variable selection methods are commonly used to select variables in linear regression. Several stepwise procedures and criteria (e.g., F-test, AIC, BIC, and Mallows' Cp) have been proposed to select variables (Efroymson, 1960; Venables and Ripley, 2002). Variable selection has been addressed in several statistical methods including parametric regression (Murtaugh, 2009), nonparametric regression (Kulasekera, 2001), and additive models (Xue, 2009). However, these procedures cannot be used in the Nearest Neighbor (NN) imputation because the NN model is fundamentally different. For instance, model accuracy in training data does not automatically improve as more predictor variables are added, and the definition of model complexity is not straightforward. This study focuses on the

Received 29 November 2011. Accepted 19 June 2012. Published on the Web at http://pubs.casi.ca/journal/cjrs on 14 November 2012.

Petteri Packalén[1] and Hailemariam Temesgen. College of Forestry, Oregon State University, 204 Peavy Hall, Corvallis, OR 97331, USA.

Petteri Packalén and Matti Maltamo. Faculty of Science and Forestry, University of Eastern Finland, P.O. Box 111, 80101, Finland.

[1]Corresponding author (e-mail: petteri.packalen@uef.fi).

Vol. 38, No. 5, October/octobre 2012

selection of predictor variables to NN imputation to predict continuous variables. There are not many studies that address this issue, but there has been some research in machine learning for selecting variables that can be used in classification model (Guyon et al., 2004). A typical example of a classification task where variable selection plays an important role is that of gene selection from microarray data (Guyon and Elisseeff, 2003). In these applications it is typical that the number of candidate variables is considerably higher ($p >> n$ problems) than the number of observations. In remote sensing based forest inventories it is unusual to have such a situation where $p >> n$, yet many ideas used here have been borrowed from the domain of machine learning.

Variable selection algorithms

Some variable selection algorithms for the NN imputation method in remote sensing based forest inventory have been published. Maltamo et al. (2006) used an algorithm that first inserts transformations from continuous variables and then deletes variables that do not significantly contribute via stepwise optimization of the relative root mean square error (RMSE). The optimization was conducted in terms of one response variable. The algorithm was used with Most Similar Neighbor (MSN) imputation (Moeur and Stage, 1995) to predict plot volume using aerial photographs, stand register, and Airborne Laser Scanning (ALS) data. Packalén and Maltamo (2007) used an algorithm that minimizes the weighted average of relative RMSEs. The optimization was conducted in terms of multiple response variables, and weights of different responses were given by the user. In this algorithm, predictors are inserted and deleted one by one in random order until the error decreases; transformations are also considered. Packalén et al. (2009) revised this algorithm such that several predictor variables can be inserted to the solution simultaneously, and the probability of excluding variables was given as an argument. Hudak et al. (2008) imputed 36 response variables including species-specific basal areas, total basal areas, and tree densities using topographic and ALS-based predictor variables. Variable selection was based upon a measure of node impurity obtained from Random Forest (RF) classification (Liaw and Wiener, 2002). At each iteration a stepwise procedure was used to iterate RF by discarding the least important predictor. This process is analogous to backwards stepwise multiple regression. Haapanen and Tuominen (2008) compared two variable selection methods: genetic algorithm (GA) and sequential forward selection. The objective was to minimize the RMSE of plot volume. Predictor variables were calculated from satellite images and aerial photographs. GA-based variable selection provided the most accurate results, sequential forward selection was the second most accurate, and no variable selection was the least accurate. Latifi et al. (2010) also used GA in variable selection to predict plot volume and biomass using remote sensing data. The GA search was implemented with a discretized response variable, i.e., variable selection was formulated as a classification task. They also selected predictor variables with stepwise regression using backward elimination. The conclusion was that the variables selected by GA were found to be superior in terms of prediction accuracy. Breidenbach et al. (2010) used stepwise forward selection to select variables to the tree crown level imputation of species-specific volumes. A predictor variable was added if the averaged RMSE over all response variables decreased by more than 1%. Breidenbach et al. concluded that operationally a reasonable approach is to execute variable selection several times and to select a model that best fulfills the objective of the inventory.

In addition to variable selection algorithms, predictors can be selected based on the correlation of X and Y or some other metric that ranks predictors. Selection of variables is often done before NN imputation, or the user may try different variable combinations in actual imputation. An assessment of different variable combinations is nevertheless very time consuming, and it is often reasonable to replace it with an automated routine.

Numerous distance metrics are used in NN imputation and many comparisons of them have been published recently (e.g., LeMay and Temesgen, 2005; Chirici et al., 2008; Hudak et al., 2008; Breidenbach et al., 2010; Latifi et al., 2010). The sudden influx of papers in this area was probably due to the release of the R package "yaImpute" (Crookston and Finley, 2008). However, there have been no studies in which different variable selection methods were tested with different distance metrics.

Overfitting

The problem of overfitting is well known in statistic and machine learning (Reunanen, 2003; Hastie et al., 2009). Overfitting means that a model adjusts to specific random features or noise of the training data but works poorly on other datasets. Therefore, it is common in machine learning to use k-fold cross validation or separate test dataset. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross validation in which k is equal to the number of observations.

Cross-validation is commonly used in NN imputation studies to evaluate the accuracy of prediction. The use of LOOCV in NN imputation means that a distance metric (e.g., weight matrix A in Equations (5–7)) is recalculated as many times as there are observations in the dataset, i.e., one observation is excluded (target observation), and a distance metric is computed with other observations (reference observations). Then the NN(s) are searched to find reference observations for a target observation. In the NN imputation studies it is common to ignore the repeated computation of distance metric, instead, distance metric is computed only once and then the NN(s) are searched by ignoring the target observation itself. This may give misleading accuracies in leave-one-out cross-validation.

Objectives

The objective of this study was to examine how to select predictor variables for NN imputation with one or more continuous response variables. Analyses were done in the context of remote sensing based inventory. Three different approaches to select predictor variables were compared in addition to the case with all predictor variables. Analyses were repeated with three different NN imputation methods using a varying number of predictor variables and one or five response variables. Evaluation was carried out by means of LOOCV. We also addressed an issue of unrealistic model accuracy caused by potential overfitting.

Material

Study area and field data

The study area of about 10 000 ha is located in eastern Finland in the municipality of Juuka. It is a typical managed Finnish boreal forest area dominated by coniferous tree species, namely Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* (L.) Karst.). In this study, all the deciduous tree species were lumped together; this species group is hence referred to as deciduous (trees). Most of the stands are fairly even aged and the data include both naturally and artificially regenerated forests. Only one tree species occurred in 12% of the plots, two in 34%, and three in 54% (**Figure 1**).

The field data, consisting of 493 sample plots, were collected during the summers of 2005 and 2006. Circular sample plots with a radius of 9 m were placed in the young, middle-aged, and mature forests. A Global Positioning System with differential correction was used to determine the position of the centre of each plot to an accuracy of about 1 m (Trimble GeoXT with external antenna elevated to 5 m, the accuracy of the positioning system was tested in a comparable forest area, unpublished data). The diameter at breast height (dbh), tree and storey class, and tree species were recorded for all trees with dbh over 5 cm, and the height of one sample tree of each species in each storey class was measured on each plot. Näslund's (1937) height model with a random constant for each plot was fitted to the data of measured heights and the model with predicted plot effects was utilized to predict heights for trees without height measurement. The volumes of individual trees were calculated as a function of dbh and predicted tree heights using the species-specific models reported by Laasasenaho (1982). Finally, tree volumes were summed up to plot level by tree species. In addition, without considering tree species, the diameter of the basal area median tree, stem number, and dominant height were calculated for each plot. Dominant height is defined here as the mean height of the 100 largest dbh trees per hectare. The characteristics of plot attributes are presented in **Table 1**.

Figure 1. The location of Juuka study area in Finland.

Remote sensing data

ALS data were collected on 13 July 2005 using an Optech ALTM 3100C laser scanning system. The test site was measured from an altitude of 2000 m above ground level (AGL) using a field of view of 30 degrees and a side overlap of about 20%. This resulted in a swath width of approximately 1050 m and a nominal sampling density of about 0.6 measurements per square metre. The Optech ALTM 3100C laser scanner captures 4 range measurements for each pulse, but in this study the measurements were reclassified to represent first and last echoes. A digital terrain model (DTM) was generated from the ALS data. First, laser points were classified as ground and nonground points using the method reported by Axelsson (2000) and then a raster DTM with a pixel size of 2.5 m was interpolated by computing the mean of the ground points within each raster cell. Values for raster cells with no data were derived using Delaunay triangulation. Finally, the raster DTM was subtracted from

Table 1. Data from 493 sample plots.

	Mean	SD	Min	Max
V Pine	89.6	68.9	0.0	351.0
V Spruce	41.6	73.2	0.0	448.8
V Deciduous	15.4	28.9	0.0	230.3
N	1289	577	158	4127
DGM (cm)	18.2	5.2	8.6	38.9
HD (m)	16.9	3.4	8.2	26.2

Note: SD, standard deviation; V, volume (m^3ha^{-1}); N, stem number; DGM, diameter of the basal area median tree; HD, dominant height.

Vol. 38, No. 5, October/octobre 2012

the ellipsoidal heights of laser points to scale the ALS data to the AGL.

Aerial photographs were provided with a Vexcel UltraCamD digital aerial camera on 1 September 2005. The images were taken at an altitude of 3000 m above ground level, which resulted a ground sample distance of 25 cm for the panchromatic band. There was plenty of overlap in images with sidelap being 65% and endlap 80%. The study area was covered by 260 images. Original color (red, green, blue) and near-infrared (NIR) bands (Vexcel refers to this processing level as Level-2) and pan-sharpened images were utilized in the analyses. For each image, external orientation was resolved by a bundle block adjustment as explained in Packalén et al. (2009).

Predictor variables

Predictor variables were calculated at the plot level from the ALS data and aerial images. Two types of variables were calculated from the aerial images: textural and spectral features. Texture features were calculated using orthorectified pan-sharpened images. The image having its nadir closest to the sample plot was always used. Grey-level co-occurrence matrices were constructed with varying re-scaling classes and lag distances as an average of all directions by bands for each plot, and texture metrics were calculated as explained in Haralick et al. (1973). Finally, 12 texture variables were selected as was done in Packalén and Maltamo (2007). Spectral features were calculated by projecting ALS points to original color and NIR bands (Level-2, no pan-sharpening) in the same manner as in Packalén et al. (2009). To avoid ground pixels, only the first echoes lying at least 0.5 m at AGL were considered. The mean and maximum pixel values were fetched to each point by iterating through all images where a point hit. A mean value was then calculated by plot from the pointwise mean and maximum values. In total, eight spectral variables were computed from aerial images.

Several height and density variables were calculated from the ALS data. All variables were computed separately with first and last echoes. The first step was to calculate height distributions for each sample plot using the heights of the AGL data. All the laser hits were considered. Weighted height percentiles 5, 10, 20, . . . , 80, 90, 95 (h_5, . . . , h_{95}) were computed, and the corresponding densities (p_5, . . . , p_{95}) were calculated for the respective percentiles. Height percentiles were calculated by summing the heights AGL. For instance, the metric $h50$ is the height at which 50% of the cumulative height has accumulated and p_{50} is the number of laser hits below h_{50} divided by all the laser hits on the plot. In addition, the mean and standard deviation of heights AGL and the proportion of vegetation hits versus ground hits using a threshold of 0.5 m were calculated. Fifty height and density variables were produced this way. Also 14 metrics were calculated from the LiDAR intensity. First, intensity was normalized for the range (Korpela et al., 2010). Then, the following intensity variables were calculated separately for the first and last echoes: percentiles 10, 30, 50, 70, 90 and both the mean and standard deviation of points 0.5 m AGL.

Eighty-three variables were calculated and used as candidate predictors in variable selection and NN imputation. Variables similar to these have been used in many studies. Therefore, it was outside the scope of this study to examine the variables that turned out to be useful in further analyses. This study focuses solely on constructing and selecting subsets of variables that are useful to build good NN-based predictors.

Response variables

Two responses were used in this study. One response is dominant height, which was selected because it can be predicted very accurately even with few predictor variables. From this point onward, in the text dominant height as a response is denoted as $1Y$. The other response is a set of five variables: volume of pine, spruce, and deciduous trees; stem number; and diameter of the basal area median tree. This set contains plot attributes that are substantially more difficult to predict than dominant height. For instance, relative RMSE is more than 10 times higher for the volume of spruce or deciduous trees than for dominant height. This set of five response variables is denoted by $5Y$.

Methods

General description of the analyses

Three different approaches to select predictors for NN imputation were evaluated. The first approach was to select predictor variables before NN imputation using factor loadings of canonical analysis (VSCC) (VS, variable selection). The second approach was a stepwise procedure in which RF importance is used as a criterion (VSRF). The third approach was to use optimization (VSSA) to minimize RMSE-%. Variable selection was implemented such that the number of predictors was fixed either to 3, 8, or 15, denoted as $3X$, $8X$, and $15X$, respectively. The performances of variable selection methods were also examined by comparing their prediction accuracies with those obtained using all predictor variables ($83X$).

Variable selection was repeated with three different distance metrics: MSN, RF proximity (NNRF), and optimized weighted Euclidean (NNSA). Detailed descriptions of these strategies for finding neighbors are found in the section "Nearest Neighbor imputation methods". The "reference set" denotes a set of observations for which both response and predictor variables are known. Similarly the "target set" refers to observations for which only the predictor variables are known. In a cross validation, each observation in the reference set is, in turn, a target observation.

Predictor variables were standardized to have a mean of 0 and standard deviation of 1 because all distance metrics were not scale invariant. The number of NN (k) was fixed to five after preliminary experiments. The estimate for the target observation i was calculated as:

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

where $\left\{ y_j^i; j = 1, \ldots, k \right\}$ is the set of NN in the reference set to the target observation i with respect to a distance metric. In the case of $5Y$, $\hat{y}_i$ was a vector of five elements, i.e., all response variables were imputed simultaneously using the same set of NN.

The accuracy of a prediction was assessed by means of RMSE-%. For the sake of simplicity, mean RMSE-% is reported in the case of $5Y$. Our focus was prediction accuracy, not reproducing variance structure evident in the observations. In the variable selection stage the distance metric was only solved once, and then NNs were searched by discarding the target observation itself as a reference observation. This approach is called TRAINCV. After variable selection the final accuracy assessment was made by means of proper LOOCV, i.e., the distance metric was re-evaluated in each iteration using reference observations only. Possible overfitting was assessed by comparing TRAINCV and LOOCV accuracies. Bias was also considered but not reported in a comprehensive manner. Because VSSA, NNSA, and NNRF are stochastic, there was some variation between runs. The number of iterations and how the stochasticity was taken into account is provided for each algorithm in the respective method sections. However, for the sake of simplicity variances between runs were not considered in this study.

All routines were implemented by C/C++. Complicated linear algebra computations were performed by the LAPACK library (Anderson et al., 1999). RF was implemented based on the code of Jaiantilal (2011), which is based on the work of Liaw and Wiener (2002), which is an R port (core by C) of original FORTRAN code by Breiman and Cutler (2011).

Variable selection methods

Preselection by Factor Loadings (VSCC)

In VSCC predictor variables are selected before conducting NN imputation using canonical analysis. The purpose of canonical correlation analysis is to determine the relationship between a set of predictor variables (X) and a set of response variables (Y). This is done by finding the linear transformations U_r and V_r for the X and Y, which maximize the correlation between them:

$$U_r = \alpha_r X \text{ and } V_r = \gamma_r Y \quad (2)$$

where α_r represents the canonical coefficients of the X variables and λ_r the canonical coefficients of the Y variables (Gittins, 1985). U_r and V_r are ordered in such a manner that the canonical correlation is largest for $r = 1$, second largest for $r = 2$, etc., and it is constrained that successively extracted canonical variates must be uncorrelated, i.e., $\mathrm{cov}(U_p, U_r) = 0$ and $\mathrm{cov}(V_p, V_r) = 0$ when $p > r$.

For variable selection purposes, we calculated the correlation between the first variate U_1 and the original predictor (X) variables. These correlations are called Factor Loadings (FL) with respect to the first canonical variate. FL summarizes how strongly the original predictor variables contribute to the first canonical variate U_1 FL is interpreted here as an importance of predictor variables in terms of a response and the most important (3, 8, or 15) predictor variables were selected accordingly. In the case of one response variable ($1Y$) this method is the same as using Pearson's correlation as an importance criterion (Rodgers and Nicewander, 1988).

Stepwise selection by Random Forest importance (VSRF)

RF is an algorithm for regression and classification developed by Leo Breiman (2001). It uses an ensemble of trees and is a modification of bagging with trees. Bagging is a technique for reducing the variance of an estimated prediction function (Hastie et al., 2009). In bagging, the idea is to fit the same regression (or classification) tree many times to a bootstrapped version of the training data and to obtain the prediction as an average. The correlation of pairs of bagged trees limits the benefit that can be obtained by averaging trees (see Equation (15.1) in Hastie et al. (2009)). The essential idea in RF is greater variance reduction in bagging by reducing the correlation between the trees without increasing variance too much. This is achieved by selecting a random subset of predictor variables at each split when trees are grown. In the model training, the prediction is an average of only those trees corresponding to bootstrap samples in which the observation itself did not appear, namely out-of-bag (OOB) samples. Detailed introduction to trees, bagging, and RF is given by Hastie et al. (2009). In this study actual RF predictions were not used; instead, RF variable importances were used in variable selection, and RF proximity matrices were used in NN imputation. In RF runs, the number of trees was set to 500, the size of the terminal nodes to 5, and the number of variables randomly sampled at each split to 1/3 of the number of variables.

RF provides two variable importance measures. Here we used the permutation importance measure which uses OOB samples (Breiman 2001). First, a tree is grown and the prediction accuracy on the OOB data is recorded. In regression the accuracy is measured by the mean squared error (Breiman and Cutler, 2011). The values for the jth variable are then randomly permuted and the prediction accuracy is recalculated. Finally, the decrease in accuracy due to permutation is averaged over all trees. This is a measure of the importance of jth predictor variable in the RF (RFIM in Algorithm 1).

Vol. 38, No. 5, October/octobre 2012

VSRF is a backward stepwise variable selection algorithm in which RF importance is used as a drop criterion. It resembles the gene selection and classification algorithm of microarray data by Díaz-Uriarte and de Andrés (2006). As in other selection procedures in this study, the desired number of predictor variables, p, is given as an argument. The core functionality of VSRF is presented in Algorithm 1. In the case of several response variables ($5Y$) an average of RF importance is taken over all responses. Because RF is stochastic, VSRF was always repeated 50 times and the predictors that occurred most often in the solutions were selected. However, the variation between RF runs was minor and repetition can be ignored in real world application.

Algorithm 1. The stepwise variable selection algorithm (VSRF) using RF importance as a drop criterion.

1. Initialize $h \leftarrow p$
2. while $h < p$
   (a) For $j = 1$ to h:
   Compute the importance criterion:
   $M_j = R^{-1} \sum_{i=1}^{R} \mathrm{RFIM}(j, i)$, where R is the number of response variables and $\mathrm{RFIM}(j, i)$ returns the RF importance of X variable j in terms of response i.
   (b) Drop the least important variable: $M_j = \min(M)$
   (c) $h \leftarrow h - 1$
3. Output the selected X variables

Minimizing RMSE by Simulated Annealing (VSSA)

In VSSA the idea is to minimize RMSE-% by solving the NN model repeatedly. The cost function (CF) to be minimized is:

$$\mathrm{CF} = R^{-1} \sum_{i=1}^{R} \frac{\sqrt{\frac{(y_i - \hat{y}_i)^2}{n}}}{\bar{y}_i} \tag{3}$$

where R is the number of response variables, y_i is the vector of observed values, $\hat{y}_i$ is the vector of predicted values, and $\bar{y}_i$ is the mean of predicted values in terms of the response i. This formulation takes into account the multivariate nature of response: it is the mean RMSE-% over all response variables. Prediction $\hat{y}$ is obtained by selected NN imputation method using a subset X of all the available predictors, XALL, thus the optimization problem can be formulated as:

$$\text{Minimize } \mathrm{CF}(\mathrm{NN}(X), Y) \text{ subject to } X \in X\mathrm{ALL} \tag{4}$$

The minimization was carried out by Simulated Annealing (SA) (Kirkpatrick et al., 1983). SA is a randomized local search method (Aarts and Lenstra, 1997). It usually gives a good approximation of the global optimum in a large search space, but it is unlikely to find the optimum solution. SA has analogy with annealing of a metal from which the name comes. The central idea is to avoid local optima by accepting probabilistically moves to worse solutions. This is controlled by the parameter called temperature which is gradually decreased according to cooling schedule. The pseudo code for SA as implemented in this study is given in Algorithm 2.

Algorithm 2. Simulated Annealing algorithm in VSSA.

1. $t \leftarrow$ set initial temperature
   $s \leftarrow$ generate random solution
   $e \leftarrow$ CalculateCost(s)
   $e_{\mathrm{best}} \leftarrow e$; $s_{\mathrm{best}} \leftarrow s$; $k \leftarrow 0$
2. while $k <$ niter
   (a) $s' \leftarrow \mathrm{PickNeighborSolution}\left(s, \frac{k}{\mathrm{niter}}\right)$
   (b) $e' \leftarrow \mathrm{CalculateCost}(s')$
   (c) if $(\exp((-(e' - e))/t) > \mathrm{Random}())$
      (i) $e \leftarrow e'$; $s \leftarrow s'$
   (d) if $(e' < e_{\mathrm{best}})$
      (i) $e_{\mathrm{best}} \leftarrow e'$; $s_{\mathrm{best}} \leftarrow s'$
   (e) $t \leftarrow \mathrm{CoolTemperature}\left(t, \frac{k}{\mathrm{niter}}\right)$
   (f) $k \leftarrow k - 1$
3. return s_{best}

PickNeighborSolution generates a new neighbor solution by altering the current solution. Following the idea of Simulated Annealing, the magnitude of change in the neighborhood is constantly decreasing while the execution proceeds. In this study, initially two-thirds of the predictors were changed simultaneously. The magnitude of change decreases linearly such that when 80% of the iterations (niter) have been completed, only one variable at a time is replaced by another. The change in neighborhood is entirely random when NN imputation is carried out by NNRF. In the case of MSN, inverses of FLs are used as weight in the random selection of variables to be replaced. The idea is that the importance of a variable in a current solution should guide how the neighborhood is modified for the next solution. In the NNSA, a similar weighting scheme is used, but instead of FLs variable weights of the current NNSA run are used. Temperature controls the probability of accepting a worse solution. It is linearly decreased until 80% of the iterations have been done and then set to zero. Thus, during the last 20% of the iterations worse solutions are no longer accepted. Initial temperatures were set to 0.2 and 1.0 for $1Y$ and $5Y$, respectively. The cost was calculated with Equation (2) and there were 500 iterations (niter) in every run.

The same optimization parameters were used with different NN imputation methods. Optimization was repeated

50 times with every combination of response, number of predictors and NN method, and the solution that gave the smallest cost was selected. Therefore, the number of optimizations runs can also be considered a parameter of optimization.

Nearest Neighbor imputation methods

NN methods use "similar" observations in p dimensional input space M of predictor variables X to infer Y. The problem of similarity can be defined as follows: given a set S of points in M and a query point $q \in M$, find the closest point(s) in S to q. Closeness, or similarity, is defined by a distance metric. A common way to define a distance metric between points X_i and X_j is:

$$d(X_i, X_j) = \sqrt{(X_i, X_j)^T A (X_i, X_j)} \tag{5}$$

where A is a square weight matrix, often required to be a positive and semi-definite (Shen et al., 2009). Matrix A can be determined either trivially (e.g., Euclidean distance), by using only X variables (e.g., Mahalanobis distance), or by using both X and Y (e.g., MSN distance based on canonical analysis). The distance metric cannot always be defined by equation 5, as in the case of NNRF. The strategies for finding NN in this study are explained in the following sections.

Most Similar Neighbor (MSN)

The MSN distance is obtained when the matrix A in Equation (5) is determined via canonical correlation analysis as follows (Moeur and Stage, 1995):

$$A = \lceil \Lambda^2 \lceil^T \tag{6}$$

where $\lceil$ is the matrix of estimated coefficients for the X variables found by canonical correlation analysis between X and Y, and Λ is the diagonal matrix of canonical correlations. Canonical variates used in Equation (6) were restricted to those which explained 97% of the variance in canonical correlation analysis.

Random Forest proximity matrix (NNRF)

In NNRF, the distance metric cannot be defined by Equation (5) as it is derived from the RF proximity matrix. The proximity matrix indicates which observations are similar. Let N denote the number of reference and M the number of target observations. Conceptually the size of the proximity matrix is $N \times M$ where reference observations are in rows and target observations are in columns (to be exact, the size of proximity matrix is $N \times (N + M)$ because proximity is solved also for training data but this complication is not considered for now). First, a proximity matrix is initialized to zero. After a tree is grown using reference observations, all of the target observations are put down the tree. If a target observation, i, ends up at the same terminal node as a reference observation, j, the corresponding element in the proximity matrix is increased by one. This is repeated for every tree, and proximities are normalized by dividing by the number of trees. In this study, the size of terminal node was five; thus, for each target observation five elements in the proximity matrix were increased by one. In the case of several responses (5 Y) the proximity matrix was built separately for each response and summed together. The actual distance metric is one minus the proportion of trees where a target observation is in the same terminal node as a reference observation. In the TRAINCV mode RF is executed with training data only. In that case the proximity matrix contains OOB proximities of training data (training data = all observations).

NN imputation using RF proximities was first described by Crookston and Finley (2008). The implementation in this study followed their approach with some minor differences. Because of stochastic nature of RF, the LOOCV accuracy was calculated with a model that produced a median cost among 50 NNRF runs.

Optimized distance metric (NNSA)

In NNSA A is the diagonal matrix:

$$A = \left| d_{i,j} \right| \quad d_{i,j} = 0 \text{ if } i \neq j \quad \forall i, j \in \{1, 2, \ldots, p\} \tag{7}$$

where the values in the diagonal are determined by optimization. This corresponds to learning a distance metric in which the different axes (predictors) are given different weights (Xing et al., 2002). A similar approach was used by Franco-Lopez et al., (2001) and Haapanen and Tuominen (2008), which minimized RMSE using the Nelder and Mead (1965) simplex search, while Tomppo and Halme (2004) used a GA. However, this type of metric is generally rare in NN imputation literature.

In this study the values in the diagonal of A were searched by minimizing the RMSE. In the case of several responses (5 Y) the mean RMSE was minimized. Minimization was carried out by SA in a similar way as in VSSA. The difference is that diagonal values of A are modified in PickNeighborSolution instead of the subset of predictor variables. Thus, here SA is used with a continuous solution space, although usually it is used to solve combinatorial optimization problems. Otherwise the algorithm follows the SA as presented in Algorithm 2. Initially, the diagonal is set to one corresponding to Euclidean distance. Diagonal values are then altered by PickNeighborSolution as described in Algorithm 3. The central idea is to decrease the magnitude of change while the execution proceeds. It is assumed that every predictor is meaningful (nonzero weight); therefore, weight is restricted to 0.05 before scaling the diagonal to have a mean of one. Initial temperatures were set to 0.2 for 1 Y and 1.0 for 5 Y, and cooled as in VSSA. The LOOCV accuracy was calculated with a model that

Vol. 38, No. 5, October/octobre 2012

produced a median cost among 50 NNSA runs because of the stochasticity of NNSA.

Algorithm 3. PickNeighborSolution with continuous solution space in NNSA. Otherwise NNSA follows the principle presented in Algorithm 2.

1. $m \leftarrow$SetModifyProbability // here fixed to 0.5
2. $g \leftarrow$SetMagnitude$\left(\frac{k}{\text{niter}}\right)$// decreases linearly from 2 to 1.1
3. for each w in Diag(A)
   (a) $r \leftarrow$Random(−1,1)
   (b) if ($r < (1-m)$)
      (i) $w \leftarrow w \times g$ // increase weight
   else if ($r > (1-m)$)
      (i) $w \leftarrow \frac{w}{g}$ // decrease weight
4. Set w to 0.05 if it is below 0.05
5. Scale Diag(A) to have a mean of one.

Results

LOOCV accuracies

Accuracies obtained by LOOCV with respect to 1Y are presented in **Table 2**. VSSA was the most accurate variable selection method. Variables selected by VSRF provided systematically less accurate predictions, but in some cases the difference was minor. VSCC was clearly the least accurate method to select variables. The most accurate prediction of 6.97% was obtained by NNRF with 8X variables. In general, accuracies improved when the number of predictor variables was increased, but with VSSA this trend was not apparent. If variables were selected by VSSA, already 3X provided about the same accuracy as 15X. Accuracies obtained by VSSA were always better, also with 3X, than what was achieved by using all 83X variables. This indicates that predictor variables that have predictive power for dominant height are correlated, and VSSA was capable of selecting a subset of three variables that really cannot be improved any more by adding more variables. The most accurate NN imputation method in terms of 1Y was NNRF. NNSA was almost as accurate as NNRF; whereas, MSN was always clearly the poorest choice.

Accuracies (mean RMSE-%) with respect to 5Y are presented in **Table 3**. VSSA was clearly the most accurate method to select variables to NN imputation when there were five response variables. VSRF was the second best, but it was already less accurate than an option where variable selection is ignored entirely (83X). VSCC was distinctly the least accurate variable selection method. Accuracies improved considerably when the number of variables was increased from 3X to 8X, but increasing the number of variables from 8X to 15X did not improve the accuracy in every case.

The smallest mean RMSE, 48.28%, was obtained with MSN using VSSA and 8X (**Table 3**). MSN also provided the second smallest RMSE at 48.40%. Hence, MSN seems to work better with 5Y than with 1Y. The differences between NN imputation methods were quite minor, especially with VSSA and all 83X variables.

LOOCV vs. TRAINCV

The decrease in accuracy between TRAINCV and LOOCV is summarized in **Figure 2**. Decrease was calculated in a relative manner ($[(\text{RMSE}_{\text{LOOCV}} - \text{RMSE}_{\text{TRAINCV}})/(\text{RMSE}_{\text{LOOCV}})] \times 100$. We use optimism as a synonym for the decrease of accuracy: a decrease in accuracy means an

Table 2. RMSE-% for 1Y (dominant height) using LOOCV by variable selection method, number of predictor variables, and NN imputation method.

	VSCC			VSRF			VSSA			All
NN method	3X	8X	15X	3X	8X	15X	3X	8X	15X	83X
MSN	9.16	8.67	8.66	8.68	8.41	8.73	7.80	7.72	8.00	8.41
NNRF	8.43	8.17	7.69	8.16	7.35	7.33	7.21	6.97	7.05	7.53
NNSA	8.36	8.09	7.42	7.96	7.55	7.23	7.26	7.23	7.21	7.56

Table 3. Mean of RMSE-% for 5Y (volume of pine, spruce, and deciduous trees, stem number, and diameter of the basal area median tree) using LOOCV by variable selection method, number of predictor variables, and NN imputation method.

	VSCC			VSRF			VSSA			All
NN method	3X	8X	15X	3X	8X	15X	3X	8X	15X	83X
MSN	74.48	73.41	66.18	79.43	55.67	56.54	59.73	48.28	48.40	52.05
NNRF	77.03	76.70	72.88	71.02	54.71	53.23	59.54	49.42	48.97	53.02
NNSA	80.12	76.55	76.29	69.72	53.15	53.47	60.00	50.17	48.89	52.47

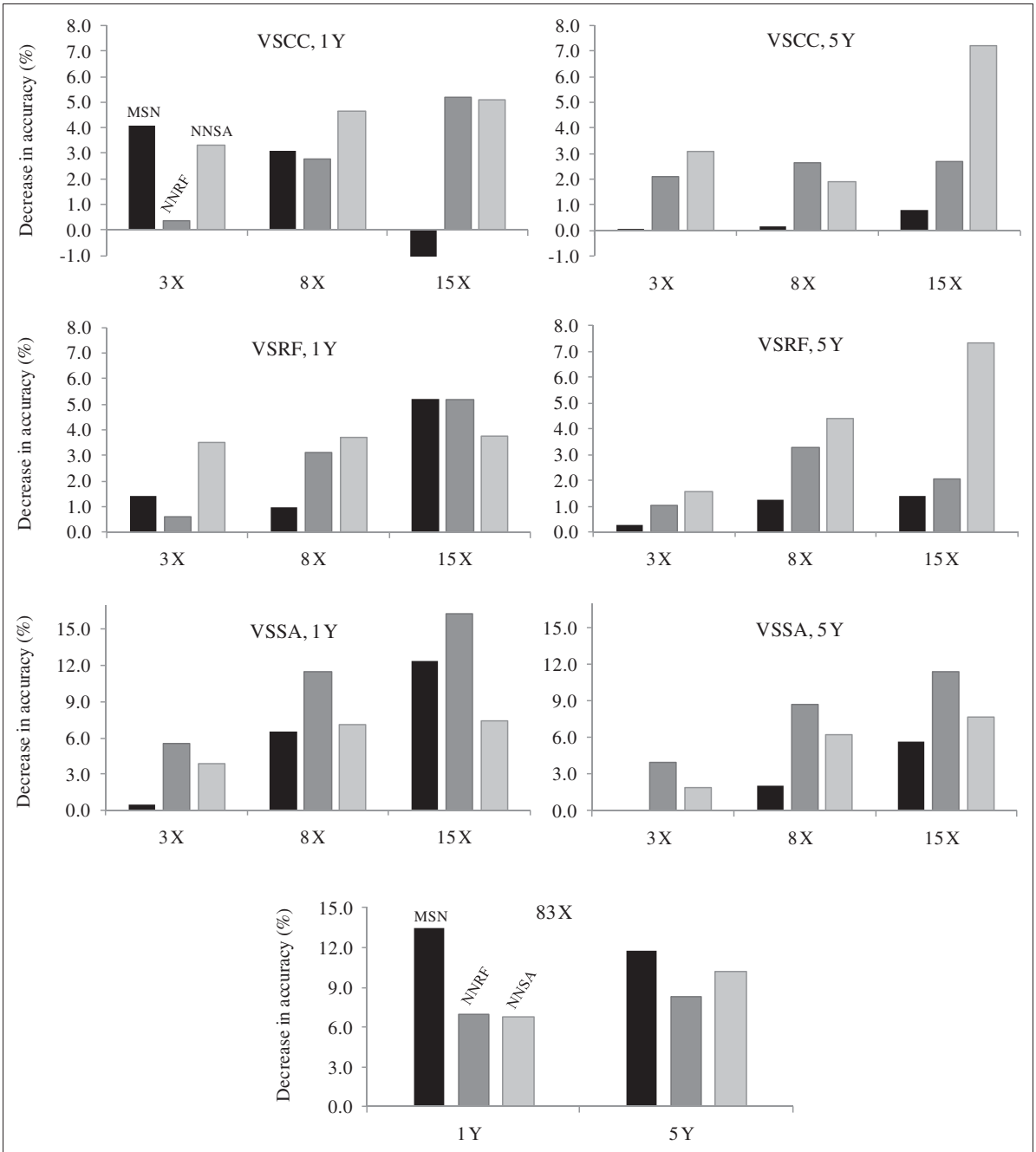

Figure 2. Decrease in accuracy between TRAINCV and LOOCV by variable selection method, number of predictor and response variables, and NN imputation method.

increase in optimism. It indicates how overoptimistic training set error is compared with the generalization error.

There was a trend that the optimism increases when the number of variables increases. This is a logical outcome, as it is apparent that a more complex model is more prone to overfitting (Hastie et al., 2009). However, in the case of VSCC, 1*Y*, MSN had a reverse trend and optimism was negative with 15*X* variables. Optimism varied between −1.03% and 16.34%. The two highest optimisms were in NNRF and MSN in the case of VSSA in terms of 1*Y*. There was also a trend that optimism was higher in the case of 1*Y* than in the case of 5*Y*. Variable selection by VSSA and the use of all variables (83*X*) were more optimistic than VSCC and VSRF.

There was not any clear trend in which the NN imputation method would be more prone to optimism. NNRF had the highest optimism in VSSA, 1*Y* and 5*Y*. MSN clearly had the lowest optimism in VSSA, 5*Y*, which is related to the reason why MSN was the most accurate method in terms of 5*Y*. However, when variable selection was ignored and all the variables were used in imputation, MSN had the highest optimism also with respect to 5*Y*.

Vol. 38, No. 5, October/octobre 2012

BIAS

In the case of 1 *Y* bias was negligible. Even the maximum bias, obtained with MSN and 8*X*, was only 0.21%. In the case of 5 *Y*, bias increased when the number of predictor variables was increased. Generally, NNRF was the least biased imputation method but differences between methods were small. Of the response variables, *N* was normally slightly overestimated and DMG was underestimated. Some response variables were clearly more biased than others; the volume of deciduous trees seemed to be especially prone to bias, which might be attributed to few plots of deciduous trees over 25 m^3 ha^{-1} (92 plots out of 493). A typical example of bias (VSSA, 15*X*, LOOCV) in the case of 5 *Y* (separate for each response) by the NN imputation method is shown in **Table 4**.

Discussion

Accuracies obtained by VSSA were better than what was achieved by using all variables. However, in many cases the use of all variables provided better accuracy than variables selected by VSCC and VSRF. In the case of several responses (5 *Y*) the use of all variables (83*X*) always provided better accuracy than what was obtained by variables selected by VSCC or VSRF. Thus, the variable selection seems to be well justified, but it must be done properly. VSSA was clearly the best variable selection method, and variable selection was in many cases more important than the NN imputation method used. This study was conducted in a single test area in boreal forests. An interesting question is to what extent would our conclusions be generalized to the broader range of forest types? We believe that our results are quite general and not restricted to data we examined. However, we assert that our comparison will lead to similar or follow-up studies in broader range of forest conditions.

VSSA is computationally demanding because its cost (or loss) function requires that the NN model be solved at each iteration. This is an issue with all optimization approaches in which RMSE or equivalent loss function is minimized. RMSE is fundamentally the same as the squared error loss used commonly for penalizing errors in regression. Unfortunately, at least based on observations made in this study, computationally less expensive methods such as VSCC and VSRF are not as efficient as methods that minimize the squared error loss by optimization (i.e., VSSA).

Table 4. Bias-% obtained with VSSA, 15*X* and LOOCV by NN imputation method.

NN method	V Pine	V Spruce	V Decid	N	DGM
MSN	−0.45	2.45	4.12	−0.18	0.40
NNRF	0.89	−1.95	−1.22	−1.02	0.99
NNSA	0.86	0.99	4.73	−1.09	1.56

Note: Negative bias means overestimation.

The number of predictor variables, *p*, was fixed in this study to 3, 8, and 15. This design choice was made for simplicity in variable selection algorithms and comparability of methods using equally complex models. Another option would be to treat *p* as an optimized parameter; this is viable but computationally more demanding. Let's think about the case where *p* is a free parameter. We know that the LOOCV accuracy does not improve as a function of model complexity (here *p*) if cost is calculated using TRAINCV. Therefore, it does not help to minimize RMSE of TRAINCV predictions. There are two solutions to overcome this issue: cost is calculated using proper cross-validation (LOOCV) or model complexity is incorporated in the cost function. Using LOOCV to calculate the cost would be computationally very demanding, however, *k*-fold cross-validation (e.g., $k = 5$ or 10) could be used instead. Incorporating complexity into model selection is a common practice in stepwise regression. AIC and BIC are probably the most common criteria of this type; however, they cannot be used in NN imputation. Vapnik–Chervonenkis theory provides a general measure of model complexity (VC dimension) but it has not been used in this context (Vapnik, 1996).

This study focused on selecting subsets of variables that are suitable to improve the accuracy of NN imputation. This contrasts with the problem of finding or ranking all potentially relevant variables (e.g., Guyon and Elisseeff, 2003). The RF variable importance score, for example, does not measure the "prediction strength" of a predictor variable when the variable is excluded and a model is refitted without that particular variable (Hastie et al., 2009). On the other hand, factor loadings used in VSCC may be better suited for finding potentially relevant variables instead of being used as a criterion in variable selection.

NNRF was the most accurate NN imputation method in the case of one response variable (1 *Y*). NNSA was nearly as accurate as NNRF; whereas, MSN was clearly the least accurate imputation method in every case against one response. Against several response variables (5 *Y*), however, MSN was the most accurate imputation method with NNRF and NNSA being slightly less accurate. NNRF was the least biased imputation method but the differences were minor. In general, bias was not a serious issue in this study.

An ideal NN imputation method would have an embedded variable selection mechanism. In this study, VSSA variable selection always improved the accuracy; thus, the imputation methods examined in this study were not optimal without explicit variable selection. NN imputation method with an embedded variable selection mechanism could be based on the discrete exclusion of variables (basically as variable selection works in this study) or the effect of unimportant variables approaching zero in a continuous manner. The latter would be similar to the idea of shrinkage methods in regression, such as ridge or Lasso regression (e.g., Hastie et al., 2009). The number of NN *k* was fixed here to 5. Alternatively, *k* could be treated as an optimized parameter, likewise *p* discussed earlier.

The RF-based imputation (NNRF) has recently gained popularity in remote sensing studies. Its distance metric is based on RF proximities, which differs substantially from most distance metrics. The RF proximity matrix is always built for one response variable because there can only be one response in RF. In NNRF with several response variables, a proximity matrix is first built separately for each response, and then these matrices are summed together to obtain a final proximity matrix. This distance metric could be considered artificial or just seemingly multivariate in terms of several responses, although it seems to work well.

"Generalization error" is an error rate when new observations are drawn from the joint distribution of the XY data, a model $f(X)$ is used to make a prediction and an error is computed by a (loss) function $L(Y,f(X))$. In this study, L computes RMSE. Note that in the generalization error, the training set is fixed. Our ultimate goal is to approximate the generalization error by LOOCV. However, according to Hastie et al. (2009) cross-validation typically estimates the "expected prediction error". The expected prediction error is an average generalization error over all training sets drawn from the same joint distribution, i.e., the training set is not fixed. Nevertheless, here we ignore this discrepancy and assume that the generalization error is obtained by LOOCV.

In this study we used LOOCV to denote that a distance metric is recalculated as many times as there are observations in the dataset, i.e., one observation is excluded (target observation), and a distance metric is solved with other observations (reference observations); whereas, TRAINCV denotes that the distance metric is solved only once and then searched for NN. The difference of TRAINCV and LOOCV varies by distance metric. In the case of Euclidean distance the result is almost the same. The difference originates from different means and standard deviations used in standardization. In the case of Mahalanobis distance, the difference originates from different covariance matrices and standardization. In this study we used distance metrics in which both predictors (X) and responses (Y) are employed. It is the most prone setup to overfitting. Although in TRAINCV the observation itself cannot belong to the set of its NN, it has a positive effect on the search for NN because it is included to the distance metric. Yet in many NN imputation studies only TRAINCV error is reported. In this study, the optimism was in many cases more than 10%. The TRAINCV error rate cannot be used to rank different methods because different methods typically adapt differentially to the training data.

It would be interesting to make a comparison of a simple Euclidean distance metric and the ones tested here. However, because the optimism originates from different sources it would exaggerate the accuracy of Euclidean distance metric. Reasonable comparison would require that LOOCV is used in variable selection as well but for computational reasons it was not possible.

In general, cross-validation should be applied to the entire sequence of modeling steps. Here we selected variables using TRAINCV and then validated the selected model by LOOCV. Therefore, we have some "selection bias" in our LOOCV accuracies. The misuse of cross-validation was intentionally made for the sake of computational efficiency. Performing variable selection repeatedly for each omitted observation would be exceedingly demanding computationally. k-fold cross-validation (e.g., $k = 5$ or 10) could be used instead, but it also has drawbacks such as randomness caused by the selection of samples to folds. But how important is this misuse of cross-validation? The selection bias is significant in genomic studies where $p >> n$ (Ambroise and McLachan, 2002). In most use cases, like in this study, p is less than n. Therefore, we do not believe that selection bias is a very important factor here. The difference between LOOCV and TRAINCV – called optimism in this study – is a far more important factor than the selection bias.

Conclusions

The optimization-based VSSA was the most accurate method to select predictors. It always provided better accuracy than what was achieved by using all variables. However, in many cases the use of all variables provided better accuracy than pre-selection by factor loadings (VSCC) or stepwise VSRF. This indicates that variable selection is an important component of imputation, and hence it must be done properly.

For a single response variable, NNRF was the most accurate distance metric while MSN was the least accurate metric. Thus, this study indicates that the MSN distance metric is not the best choice for one response. For five response variables, however, MSN was the most accurate distance metric. The optimization-based NNSA worked particularly well when the number of predictors was low.

We also demonstrated the difference between TRAINCV and LOOCV procedures. Because of overfitting TRAINCV may give accuracies that are too optimistic. Analysts should be aware of this when conducting cross-validation.

Acknowledgements

We acknowledge Prof. Timo Pukkala and Dr. Tero Heinonen for their insights while implementing optimization routines. We also thank Prof. Jukka Tuomela and Mr. Pekka Savolainen for mathematical support.

References

Aarts, E., and Lenstra, J.K. 1997. Introduction. In *Local search in combinatorial optimization*. Edited by Aarts, E., and Lenstra, J.K. John Wiley & Sons, New York, pp. 1–16.

Ambroise, C., and McLachlan, G. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, Vol. 99, No. 10, pp. 6562–6566. doi: 10.1073/pnas.102102699.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. 1999. *LAPACK Users' Guide*, 3rd. ed. 407 p.

Axelsson, P. 2000. DEM Generation from laser scanner data using adaptive TIN models. In *Proceedings of XIXth ISPRS Conference*, IAPRS, Vol. XXXIII, Amsterdam, The Netherlands. pp. 110–117.

Breidenbach, J., Næsset, E., Lien, V., Gobakken, T., and Solberg, S. 2010. Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. *Remote Sensing of Environment*, Vol. 114, No. 4, pp. 911–924. doi: 10.1016/j.rse.2009.12.004.

Breiman, L. 2001. Random forests. *Machine Learning*, Vol. 45, No. 1, pp. 5–32. doi: 10.1023/A:1010933404324.

Breiman, L., and A. Cutler. 2011. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. [accessed 5 December 2011].

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F., and Bertini, R. 2008. Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sensing of Environment*, Vol. 112, No. 5, pp. 2686–2700. doi: 10.1016/j.rse.2008.01.002.

Crookston, N.L., and Finley, A. 2008. yaImpute: An R Package for kNN Imputation. *Journal of Statistical Software*, Vol. 23, No. 10. Available from http://www.jstatsoft.org/.

Díaz-Uriarte, R., and de Andrés, S.A. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7: 3. Available from http://www.biomedcentral.com/bmcbioinformatics.

Efroymson, M.A. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*. Edited by Ralston, A., and Wilf, H.S. New York: John Wiley & Sons. pp. 191–203.

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 neighbor method. *Remote Sensing of Environment*, Vol. 77, No. 3, pp. 251–274. doi: 10.1016/S0034-4257(01)00209-7.

Gittins, R. 1985. *Canonical Analysis: A Review with Applications in Ecology*. Springer-Verlag, Berlin. 351 p.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182.

Guyon, I.M., Gunn, S.R., Ben Hur, A., and Dror, G. 2004. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*. Edited by Saul, L.K., Weiss, Y., and Bottou, L. MIT Press, Cambridge, MA, pp. 545–552.

Haapanen, R., and Tuominen, S. 2008. Data combination and feature selection for multi source forest inventory. *Photogrammetric Engineering and Remote Sensing*, Vol. 74, No. 7, pp. 869–880.

Haralick, R.M., Shanmugam, K., and Dinsteein, J. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 3, No.6, pp. 610–621. doi: 10.1109/TSMC.1973.4309314.

Hastie, T., Tibshirani, R., and Friedman, J.H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Springer-Verlag, New York. 745 p.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. *Biometrics*, Vol. 32, No. 1, pp. 1–49. doi: 10.2307/2529336.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, Vol. 112, No. 5, pp. 2232–2245. doi: 10.1016/j.rse.2007.10.009. Corrigendum: Remote Sensing of Environment 2009. 113(1): 289–290. doi: 10.1016/j.rse.2008.08.006.

Jaiantilal, A. 2011. http://code.google.com/p/randomforest-matlab. [accessed 10 September 2011].

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. 1983. Optimization by Simulated Annealing. *Science*, Vol. 220, No. 4598, pp. 671–680. doi: 10.1126/science.220.4598.671.

Korpela, I., Ørka, H.O., Maltamo, M., Tokola, T., and Hyyppä, J. 2010. Tree species classification using airborne LiDAR – effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. *Silva Fennica*, Vol. 44, No. 2, pp. 319–339.

Kulasekera, K.B. 2001. Variable selection by stepwise slicing in nonparametric regression. *Statistics & Probability Letters*, Vol. 51, pp. 327–336. doi: 10.1016/S0167-7152(00)00167-X.

Laasasenaho, J. 1982. Taper curve and volume function for pine, spruce and birch. *Communicationes Instituti Forestalis Fenniae*, Vol. 108, pp. 1–74.

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*, Vol. 83, No. 4, pp. 395–407. doi: 10.1093/forestry/cpq022.

LeMay, V., and Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science*, Vol. 51, No. 2, pp. 109–119.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. *R News*, Vol. 2, No. 3, pp. 18–22. Available from http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., and Kangas, J. 2006. Nonparametric estimation of plot volume using laser scanning, aerial photography and stand register data. *Canadian Journal of Forest Research*, Vol. 36, No. 2, pp. 426–436. doi: 10.1139/x05-246.

Moeur, M., and Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science*, Vol. 41, No. 2, pp. 337–359.

Murtaugh, P.A. 2009. Performance of several variable-selection methods applied to real ecological data. *Ecology Letters*, Vol. 12, No. 10, pp. 1061–1068. doi: 10.1111/j.1461-0248.2009.01361.x.

Näslund, M. 1937. *Skogsförsöksanstaltens gallringsförsök i tallskog*. Meddelanden från Statens Skogsförsöksanstalt, 29. 169 p. (In Swedish).

Nelder, J.A., and Mead, R. 1965. A simplex method for function minimization. *Computer Journal*, Vol. 7, No. 4, pp. 391–398.

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 Sensing of Environment*, Vol. 109, No. 3, pp. 328–341. doi: 10.1016/j.rse.2007.01.005.

Packalén, P., Suvanto, A., and Maltamo, M. 2009. A two stage method to estimate species-specific growing stock by combining ALS data and aerial photographs of known orientation parameters. *Photogrammetric Engineering and Remote Sensing*, Vol. 75, No. 12, pp. 1451–1460.

Reunanen, J. 2003. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, Vol. 3, No. 7–8, pp. 1371–1382.

Rodgers, J.L., and Nicewander, W.A. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, Vol. 42, No. 1, pp. 59–66. doi: 10.2307/2685263.

Shen, C., Kim, J., Wang, L., and Van den Hengel, A. 2009. Positive semidefinite metric learning with Boosting. *Proceedings of Advances in Neural Information Processing Systems*, Vol. 22, pp. 1651–1659.

Tomppo, E., and Halme, M. 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. *Remote Sensing of Environment*, Vol. 92, No. 1, pp. 1–20. doi: 10.1016/j.rse.2004.04.003.

Vapnik, V. 1996. *The nature of statistical learning theory.* Springer, New York. 314 p.

Venables, W.N., and Ripley, B.D. 2002. *Modern Applied Statistics with S-plus*, 4th edition. Springer-Verlag, New York. 495 p.

Xing, E.P., Ng, A.Y., Jordan, M.I., and Russell, S. 2002. Distance metric learning, with application to clustering with side-information. *Proceedings of Advances in Neural Information Processing Systems*, Vol. 15, pp. 505–512.

Xue, L. 2009. Consistent variable selection in additive models. *Statistica Sinica*, Vol. 19, pp. 1281–1296.