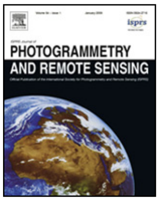
Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



Predicting individual tree attributes from airborne laser point clouds based on the random forests technique

Xiaowei Yu[a,*], Juha Hyyppä[a], Mikko Vastaranta[b], Markus Holopainen[b], Risto Viitala[c]

[a] *Finnish Geodetic Institute, Geodeetinrinne 2, PL 15, 02431 Masala, Finland*

[b] *Department of Forest Resource Management, University of Helsinki, Latokartanonkaari 7, 00014 Helsinki, Finland*

[c] *Hämeen ammattikorkeakoulu HAMK, P.O. Box 230, 13101 Hämeenlinna, Finland*

ARTICLE INFO

Article history:
Received 13 November 2009
Received in revised form
12 August 2010
Accepted 13 August 2010
Available online 21 September 2010

Keywords:
Laser scanning
Forestry
Prediction
Feature
Detection

ABSTRACT

This paper depicts an approach for predicting individual tree attributes, i.e., tree height, diameter at breast height (DBH) and stem volume, based on both physical and statistical features derived from airborne laser-scanning data utilizing a new detection method for finding individual trees together with random forests as an estimation method. The random forests (also called regression forests) technique is a nonparametric regression method consisting of a set of individual regression trees. Tests of the method were performed, using 1476 trees in a boreal forest area in southern Finland and laser data with a density of 2.6 points per m^2. Correlation coefficients (R) between the observed and predicted values of 0.93, 0.79 and 0.87 for individual tree height, DBH and stem volume, respectively, were achieved, based on 26 laser-derived features. The corresponding relative root-mean-squared errors (RMSEs) were 10.03%, 21.35% and 45.77% (38% in best cases), which are similar to those obtained with the linear regression method, with maximum laser heights, laser-estimated DBH or crown diameters as predictors. With random forests, however, the forest models currently used for deriving the tree attributes are not needed. Based on the results, we conclude that the method is capable of providing a stable and consistent solution for determining individual tree attributes using small-footprint laser data.

1. Introduction

Accurate estimation of forest characteristics is essential for forest management and planning, and there is a growing demand for accurate, precise and timely information on forest resources to support decision making. Conventionally, forest information has been acquired through expensive and time-consuming field inventories. Developments in remote sensing technologies have resulted in breakthroughs in the performance of forest resource inventories in terms of efficiency and scales (Hudak et al., 2008; Tomppo et al., 2002; Tomppo and Halme, 2004; Zhao et al., 2009). Accordingly, a variety of image processing and analysing techniques have been developed for the estimation of forest inventories and the retrieval of biophysical attributes from remotely sensed data.

Recent developments of remote sensing technologies, in particular laser scanning techniques, have taken forest assessment to a new height. With the capability of directly measuring forest structure (including canopy height and crown dimensions), laser scanning is increasingly used for forest inventories at different levels. Previous studies have shown that airborne laser scanning (ALS) data can be used to estimate a variety of forest inventory attributes, including tree-, plot- and stand-level estimates for tree height (Falkowski et al., 2006; Hyyppä and Inkinen, 1999; Magnussen et al., 1999; Maltamo et al., 2004), biomass (Bortolot and Wynne, 2005; Lefsky et al., 1999; van Aardt et al., 2008), volume (Hyyppä et al., 2001; Næsset, 1997; Wallerman and Holmgren, 2007), basal area (Lefsky et al., 1999; Means et al., 2000; Næsset, 2002) and tree species (Brandtberg, 2007; Holmgren and Persson, 2004; van Aardt et al., 2008).

Approaches to deriving forest information from ALS data can be mainly divided into two groups, one based on the statistical canopy height distribution (e.g. Holmgren, 2004; Lim et al., 2002, 2003; Næsset, 2002) and the other based on individual tree detection (e.g. Hyyppä and Inkinen, 1999). These categories relate to the need for scale and accuracy of the forestry information and available point density of the ALS data. Both approaches use canopy height models or canopy height-corrected point clouds to derive a set of features.

* Corresponding author. Tel.: +358 9 29555213; fax: +358 9 29555200.
E-mail addresses: yu.xiaowei@fgi.fi (X. Yu), juha.hyyppa@fgi.fi (J. Hyyppä), mikko.vastaranta@helsinki.fi (M. Vastaranta), markus.holopainen@helsinki.fi (M. Holopainen), risto.viitala@hamk.fi (R. Viitala).

0924-2716/$ – see front matter © 2010 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2010.08.003

In distribution-based techniques, features and predictors are assessed from the laser-derived surface models and canopy height point clouds and directly used for forest attribute estimation, typically using parametric or nonparametric regression. In individual tree-based approaches, neighbourhood information on canopy height point clouds and pixels of canopy height models are used to derive physical features and measures from each tree, such as crown size, individual tree height and location. Other important stand characteristics are estimated from these data, using existing models or regression analyses. The attributes are then aggregated at the required level (groups of trees, plots, stands) (Hyyppä et al., 2008).

Since the laser-derived features in most cases are not directly related to forest attributes, the relationships between features and the spatially coincident field measurements are established. Parametric or nonparametric regression approaches are normally used to construct the regression models (Holmgren, 2004; Means et al., 2000; Packalén and Maltamo, 2006; Wallerman and Holmgren, 2007). In parametric regression, a priori assumptions are made about the nature of the relationships among the response and predictor variables. Parametric regressions have been successful at predicting forest attributes. For example, Lim et al. (2003), Means et al. (2000) and Næsset (2002) used multiple linear regression to determine equation forms for predicting the mean height, basal area and timber volume, using laser-derived metrics at the stand level. Similarly, in Hall et al. (2005), 39 metrics were derived from the ALS data and used to estimate the stand height, canopy base height, tree density, basal area, crown bulk density, and total aboveground and foliage biomass in low-density forests by parametric regression modelling. Bortolot and Wynne (2005) used stepwise multiple linear regression to determine an equation form for predicting the biomass, using laser-derived tree counts and heights at the individual tree level.

The nonparametric regression methods commonly used for forest inventory include nearest-neighbours (NN) techniques. Depending on how the nearest neighbour in the feature space is defined, different variants of the techniques are available. These techniques have been used in multisource and multivariate forest inventories (e.g. Maltamo et al., 2006; Tomppo and Halme, 2004). Wallerman and Holmgren (2007) used the most-similar-neighbours (MSN) technique to estimate the stem density and volume for stands dominated by Norway spruce, Scots pine and birch in Sweden from ALS and optical satellite image data. Packalén and Maltamo (2006) used *k*-MSN techniques to predict species-specific plot-level volumes with a combination of ALS data and aerial images. The predictor variables were selected manually using insertions and deletions iteratively, and the best combination of variables was selected based on root-mean-squared errors (RMSEs).

More recently, the random forests (RF) approach has attracted wide publicity, due to its robustness and flexibility in modelling the relationship from samples and in predicting/imputing the values of new and unknown samples. These techniques already exist and have applications in some fields, e.g., in language modelling for speech recognition (Xu and Jelinek, 2009), species distribution modelling (Cutler et al., 2007), diagnostic and prognostic classification tasks based on microarray gene expression data (Statnikov et al., 2008) and ecosystem modelling of the distribution of large numbers of tree species under current and future climate scenarios (Prasad et al., 2006). Several studies have been conducted for applications in forest research using ALS data. In Hudak et al. (2008), the RF technique was applied for imputing plot-level basal area and tree density for managed mixed-conifer forests in northcentral Idaho, USA, using ALS data and compared with several other nonparametric regression methods. They concluded that the RF technique was the most robust and flexible method among those tested. Ørka et al. (2009) used the RF method to classify tree species based on laser-derived features or a subset of features in Norway and compared the results with those of support vector machines (SVMs) and linear discriminant analysis (LDA). They obtained fairly similar accuracies, using all features and RF or SVM methods, compared with LDA and a feature-selection strategy.

At individual tree-level estimation, it was originally assumed that by directly measuring the physical features of trees, such as the height, crown size and species, and using existing individual tree models, it could be possible to estimate other individual tree attributes, such as the DBH and stem volume. Previous work, such as Villikka et al. (2007), has shown that, even at the individual tree level, it is preferable to establish a relationship between the features derived from the laser measurements and the spatially coincident field measurements, because there are many problems in using existing forest models: (1) the use of existing national models is difficult, for example, for companies working globally, and the lack of national models is sometimes a problem, (2) the use of national models leads to errors that are caused by local variability in the forest. With the RF technique, the need for national models is limited to the field reference collection, but the use of terrestrial laser scanning in field data collection will also replace this need in the future.

In this paper, we demonstrate that the RF technique together with a new tree detection algorithm can be used efficiently for forest inventories at the individual tree level, even with relatively low point densities. Its performance in predicting three important tree attributes (tree height, DBH and volume) was investigated based on the features derived from ALS data for 1476 correctly detected trees.

2. Materials

2.1. Study area

The 5 km × 5 km study area, located in Evo, southern Finland, is part of the southern Boreal Forest Zone. It contains approximately 2000 ha of managed boreal forest, having an average stand size of slightly less than 1 ha. The elevation of the area varies from 125 to 185 m above sea level. Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*) are the dominant tree species in the study area, and they contribute 40% and 35% of the total volume, respectively, whereas the share of deciduous trees is 24% of the total volume.

2.2. Field measurements

Field measurements were undertaken in summer 2007 and 2008 on 69 circular plots (10 m fixed radius). Sampling of the field plots was based on prestratification of existing stand inventory data. The plots were selected so that they would represent different forest densities, ages, site types and tree species. Fig. 1 shows six selected plots with different characteristics. All trees with a DBH of over 5 cm were tallied and the tree height, DBH, lower limit of living crown, crown width and species were recorded. The stem volumes were calculated with standard Finnish models (Laasasenaho, 1982).

The tree locations were calculated using the geographic coordinates of the plot centres and the direction and distance of trees relative to the plot centre. The plot centres were measured with a Trimble GEOXM 2005 global positioning system (GPS) device (Trimble Navigation Ltd., Sunnyvale, CA, USA), and the locations were postprocessed with local base station data, resulting in an average error of approximately 0.6 m (Holopainen et al., 2008). The tree heights were measured using Vertex clinometers. The DBHs were measured with steel callipers. The descriptive statistics of the trees are summarized in Table 1.

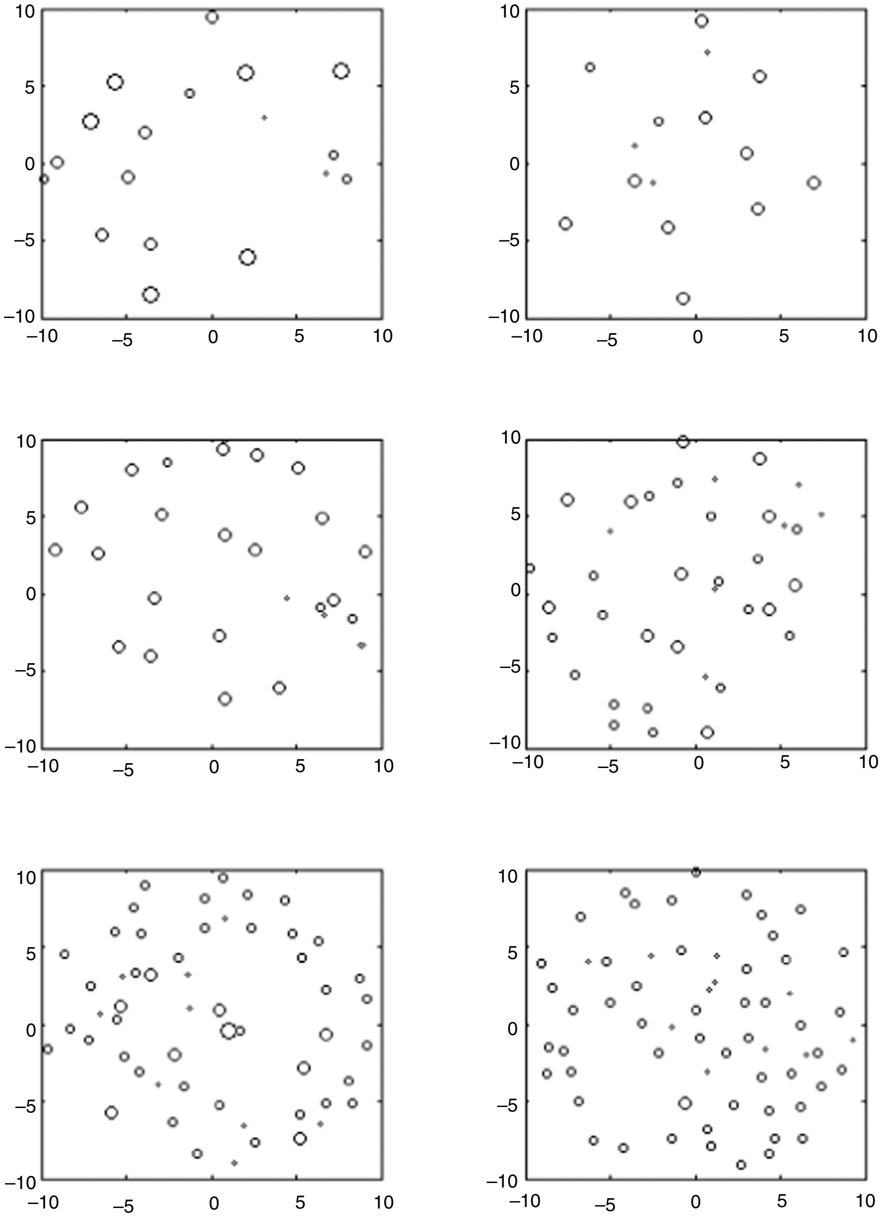


Fig. 1. Stem maps for plots of varying stem density and volume. The radius of the plots is 10 m. The size of the circle indicates the size of the stems.

Table 1
Statistical summary of tree variables.

	Min	Max	Mean	Standard deviation
Tree height (m)	4.4	34.6	16.53	4.53
DBH (cm)	1.8	56.2	18.23	6.62
Volume (m^3)	0.002	2.85	0.27	0.28

2.3. Airborne laser data

Airborne laser scanning data were collected in midsummer 2006, using an Optech ALTM3100C-EA system operating at a pulse rate of 100 kHz. The system was configured to record up to three echoes per pulse, i.e., first or only, last and intermediate. The data were acquired at a flight altitude of 800 m, resulting in an average pulse density of 2.6 (ranging from 1.8 to 3.4) laser hits per square metre in nonoverlapping areas and a footprint size that was 70 cm in diameter.

The ALS data were first classified into ground or nonground points using the standard approach of the TerraScan based on the method explained by Axelsson (2000). A digital terrain model was then created using classified ground points. The laser heights above ground (normalized height or canopy height) were calculated by subtracting the ground elevation from the corresponding laser measurements. Canopy heights close to zero were considered as the returns from ground and those greater than 2 m from vegetation. Only the returns from vegetation were used for tree feature extraction.

3. Methodology

3.1. Tree detection and feature extraction

A raster canopy height model (CHM) was created from normalized data for all plots inside the coverage of ALS data for individual tree detection and crown segmentation. Single-tree segmentations were performed on the CHM images (Fig. 2(a)) using a minimum curvature-based region detector. During the segmentation processes, the tree crown shape and location of

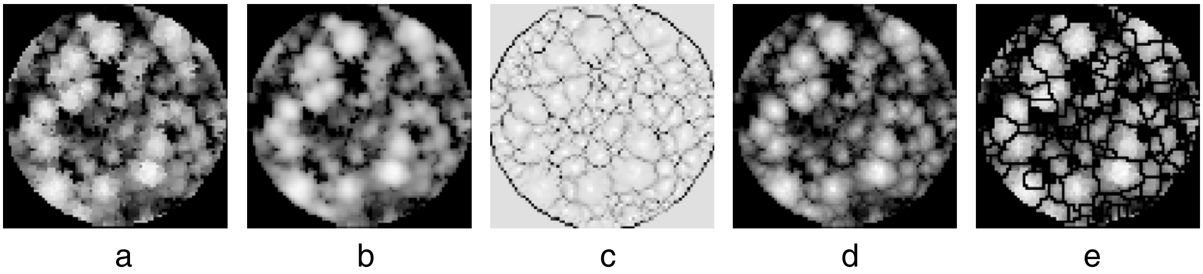


Fig. 2. Individual tree detection for one sample plot. (a) Original CHM, (b) smoothed image, (c) minimum curvature image, (d) image stretched by minimum curvature, (e) watershed segmentation.

individual trees were determined. The procedure consisted of the following steps.

1. The CHM was smoothed with a Gaussian filter to remove small variations on the crown surface. The degree of smoothness is determined by the value of the standard deviation (Gaussian scale) and kernel size of the filter (Fig. 2(b)).
2. Minimum curvatures were calculated (Fig. 2(c)). Minimum curvature is one of the principal curvatures. For a surface such as a CHM, a higher value of minimum curvature describes the tree top.
3. The smoothed CHM image was then scaled based on the computed minimum curvature, resulting in a smoothed yet contrast-stretched image (Fig. 2(d)).
4. Local maxima were then searched in a given neighbourhood. They were considered as tree tops and used as markers in the following marker-controlled watershed transformation for tree crown delineations (Fig. 2(e)).

Each segment was considered as presenting a single tree crown. The laser returns falling within each individual tree segment were extracted and the canopy heights of these returns were used for deriving the tree features. In total, 26 features were generated (Table 2).

After the segmentation, field-measured trees were matched with laser-detected trees using a method based on the Hausdorff distance (Yu et al., 2006). The Hausdorff distance is the maximum distance of a set to the nearest point in the other set. More formally, the Hausdorff distance from set A to set B is a maximum function, defined as

$$h(A, B) = \max_{a \in A} \left\{ \min_{b \in B} \{d(a, b)\} \right\}, \quad (1)$$

where a and b are points of sets A and B, respectively, and $d(a, b)$ is any metric between these points in any dimension of space, e.g. the Euclidian distance between a and b. When the method is applied for tree matching, the aim is to determine the tree pairs, one from each dataset (field-measured trees and laser-detected trees), that are closest to each other.

3.2. *Random forests concept*

The RF algorithm developed by Breiman (2001) is a nonparametric regression approach. It is composed of a set of regression trees that are constructed from bootstrapped training data, which in general are sets of samples taken randomly with replacement from the original training set. A regression tree is built for each of the bootstrap sets, which together create the RF. A regression tree is a sequence of rules that split the feature space into partitions that have similar values for the response variable. Fig. 3 illustrates the workflow for random forests.

The performance of the RF is dependent on the prediction accuracy of the individual regression trees and the correlation between the regression trees (Breiman, 2001). To reduce the correlation, two types of randomness are used: first, a random sample of training sets for growing each regression tree, and second, in growing any given regression tree, a random selection of predictor features at each node in choosing the best split. Thus, three parameters in the RF need to be set: one is how many trees to construct (N) and another is how many predictor variables to try at each node for splitting (M). The third parameter to be specified is the node size (NS), which determines how deep the regression tree will grow.

When the regression tree is constructed each time, a new training set (bootstrap samples) is drawn with replacement from the original training set. About one third of the samples are left out of the new training set. The samples that are not included in the new training set are called out-of-bag samples. They can act as a testing set in the approach, i.e., each time a regression tree is constructed with drawn training samples, the out-of-bag samples are passed down the regression tree to obtain new estimates. By averaging the estimates over the regression trees for a sample whenever it is an out-of-bag sample and then comparing the average obtained with the field measurement, we could obtain a prediction error estimate for the RF constructed.

Compared with other regression approaches, several advantages have made the RF an attractive tool for regression: it does not

Table 2
Features extracted from ALS data for trees.

Feature	Description
X[a]	Tree location (easting)
Y[a]	Tree location (northing)
Hmean	Arithmetic mean of laser heights
HSTD	Standard deviation of heights
Hrange	Heights range
CA	Crown area as area of convex hull
CV	Crown volume as convex hull in 3D
H0	Heights 0th percentile
H10	Heights 10th percentile
HP20	Heights 20th percentile
H30	Heights 30th percentile
H40	Heights 40th percentile
H50	Heights 50th percentile
H60	Heights 60th percentile
H70	Heights 70th percentile
H80	Heights 80th percentile
H90	Heights 90th percentile
H100	Heights maximum
DS1	Percentage of returns below 10% of total height
DS2	Percentage of returns below 20% of total height
DS3	Percentage of returns below 30% of total height
DS4	Percentage of returns below 40% of total height
DS5	Percentage of returns below 50% of total height
DS6	Percentage of returns below 60% of total height
DS7	Percentage of returns below 70% of total height
DS8	Percentage of returns below 80% of total height
DS9	Percentage of returns below 90% of total height
MaxD	Maximum crown diameter when crown was considered an ellipse

[a] X, Y are listed in the table, but not used in regression analyses.

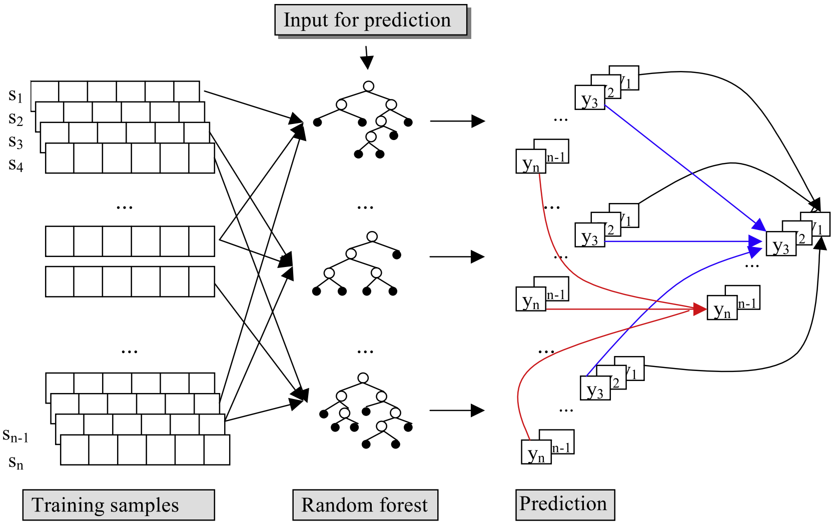


Fig. 3. Workflow for random forests.

overfit when the number of regression trees increases (Breiman, 2001) and it needs no variable selection, which could be a difficult task if the number of the predictor variables is huge. Furthermore, RF contains a built-in cross-validation method and needs no separate testing dataset for evaluating the performance because the out-of-bag samples act in the same way as the testing dataset and give realistic prediction error estimates. In addition, RF makes no distributional assumptions about the predictor or response variables, and can handle situations in which the number of predictor variables greatly exceeds the number of observations (Cutler et al., 2007).

To assess the importance of a specific predictor variable (feature), the values of the variable in the out-of-bag samples are randomly permuted and then the modified out-of-bag samples are passed down the tree to get new predictions. The increase of estimation error for the modified and original out-of-bag data provides a useful measure for determining the feature importance, although feature selection is not needed in RF (Breiman and Cutler, 2008).

3.3. Evaluation of random forests for individual tree attribute determination

To evaluate the performance of the method for predicting individual tree attributes, namely tree height, DBH and stem volume, two thirds of the matched trees were randomly selected and used for automatic construction of the RF. The remaining trees were left for testing, although the method needs no separate set of testing data to give an error estimate. The idea is to determine how good the error estimation is from the out-of-bag samples. The candidate predictors included the 26 tree features derived from the ALS data (Table 2). The response variables were the tree height, DBH and stem volume. The RF was run 50 times, each time with two thirds of the randomly selected data as a training set to grow and combine a given number (60) of regression trees. The remaining one third of the data (as testing data) was then passed down the RF to obtain a test error estimate for each run. Different forests were grown for each tree attribute. Five predictors were randomly selected at each node for the best splitting. The maximum size of the leaf nodes was set at 5. RMSEs between the predicted and observed values were used as a measure for error estimates and the correlation coefficient (R) as a measure for the goodness of fit of the model. The mean and standard deviation of these two measures were calculated for 50 runs to assess the consistency and stability. The same computation was followed for the out-of-bag samples.

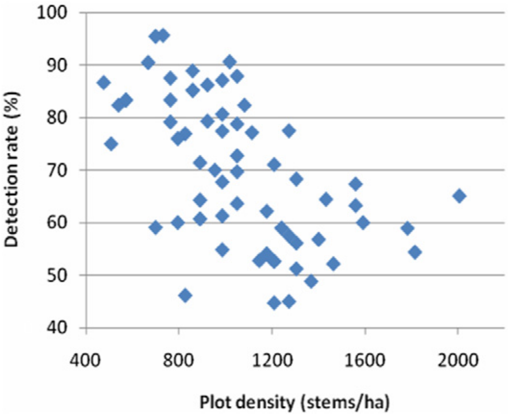


Fig. 4. Tree detection accuracy versus plot density for 69 plots.

To assess the performances of the RF, linear regressions were also carried out using the same datasets and experimental settings as in the RF, i.e., the linear regression was run 50 times, each time with two thirds of the randomly selected data for developing the models and the remaining one third for testing. The mean R and RMSE between the predicted and observed values were computed for 50 runs. In the modelling, the tree heights were regressed as a function of maximum laser height (H100), and the DBH as a function of H100 and crown diameter. The stem volume was calculated for trees using a volume equation (Laasasenaho, 1982) with the laser-estimated DBH and H100 as inputs, and regressed against the reference volumes.

4. Results

4.1. Individual tree detection and matching with field-measured trees

The automatically detected trees were compared with the field-measured trees. Based on the matching results between the field-measured trees and laser-detected trees, 43–96% (mean 69%) of the trees were detected for 69 plots. Fig. 4 shows the detection accuracy as a function of stem density. It can be seen that detection accuracy decreased as the stem density increased. In all, 1476 trees

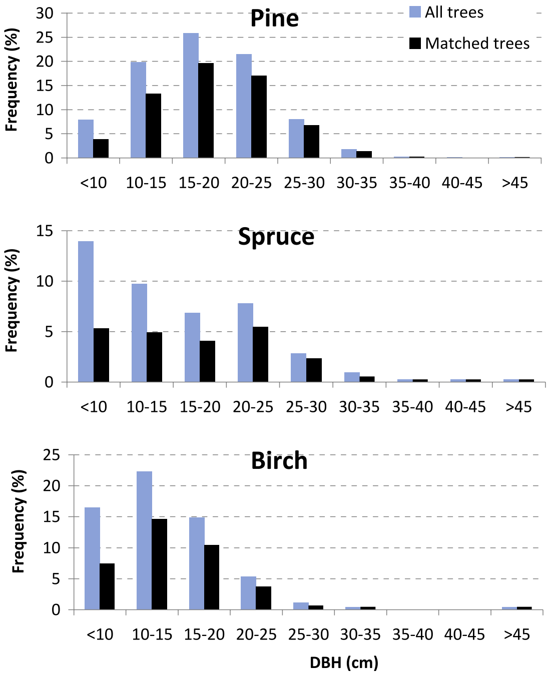


Fig. 5. Diameter distributions of all trees versus matched trees for pine, spruce and birch.

were matched and used in further analyses. Among them, 1073 were pine, 171 spruce, 160 birch and 72 other deciduous trees. The mean differences for tree location were 0.03 and 0.05 m in the *X* and *Y* directions. The corresponding standard deviations were 0.82 and 0.87 m for these matched trees. Fig. 5 shows the diameter distributions of all trees and matched trees for pine, spruce and birch. As the results show, the detection rate for larger trees was significantly better than that for smaller trees.

4.2. Prediction of individual tree variables

Estimates of individual tree attributes with the RF approach were compared with field-measured or derived values. Table 3 summarizes the results of over 50 trials. Fig. 6 shows the scatterplots of predicted versus reference values for tree height, DBH and stem volume and for training and testing datasets. The tree height was estimated with an RMSE of 10.03% and an *R*-value of 0.93. For DBH estimation, the corresponding values were 21.35% and 0.79, respectively. The worst case was the estimates for tree volume, with an RMSE of 45.77% and an *R*-value of 0.87. The biases and standard deviations are also given in Table 3.

When the field-measured values were regressed against the laser-derived variables, the linear models gave RMSEs of 9.7% and *R*-values of 0.93 for tree height, 21.49% and 0.79 for DBH and 45.95% and 0.88 for stem volume estimates, respectively (Table 4). There were no big differences between the RF and linear regression in terms of RMSE and *R*. However, the RF gave more stable and consistent results than linear regression, as indicated by the smaller standard deviations of the RMSEs over 50 runs, especially for volume estimation.

4.3. Error estimation from out-of-bag samples

In Table 3, the prediction error estimations from the out-of-bag samples (training set) are listed against the error estimations from the testing set. The close agreement of bias, *R* and RMSE between the out-of-bag samples and the testing set suggested that the out-of-bag samples could be used as a testing set and the error estimate from them as a measure to indicate how well the RF performs.

4.4. Feature importance

In Fig. 7, 26 laser-derived features are evaluated for their importance in tree height, DBH and stem volume estimation. The higher the value, the more important the feature. As can be seen, the most significant feature for all three tree attribute estimations is H100, followed by higher percentiles of canopy height distribution. The features describing crown shape have more predictive power in the volume and DBH estimates than in the tree height estimates.

5. Discussion

The results show the potential use of the method for predicting the tree attributes of individual trees in real forests. Significant laser-based features used for prediction are tree height-related features, with the H100 as the most important one, as expected. The density-related features did not play an important role in tree attribute estimation. This could be explained by the lower pulse density of the ALS data. The full potential of the approach should be explored using high-density data and more laser-derived features.

It is difficult to compare our results with those from previous studies, because the studies were conducted under varying forest conditions with different scanning systems and data characteristics. Kaartinen and Hyyppä (2008) concluded that forest conditions play a major role in determining the results of estimation. Hyyppä et al. (2005) obtained random errors between 25% and 30% for pine tree volumes, using volume equations with the laser-measured tree height and DBH (derived from the laser tree height and crown diameter) as inputs. Using a linear model fitted to the data, acquired over boreal forests in Sweden, Persson et al. (2002) obtained RMSEs of 0.63 m for tree height estimates, 3.8 cm (10% of the mean value) for the DBH, using the laser-measured tree height and crown diameter as independent variables in a regression model, and 0.21 m^3 (22% of the mean value) for the stem volume at the individual tree level. Villikka et al. (2007) investigated several alternatives for predicting the tree-level stem volume of Norway spruce using ALS data over an area dominated by Norway spruce, Scots pine and birch in eastern Finland. RMSEs ranging from 27% to 52% were achieved. The best one was obtained using a volume equation with the laser tree height and DBH (modelled from all laser-derived features) as inputs.

A nonparametric estimation method, *k*-MSN, was used by Maltamo et al. (2009), achieving RMSEs of better than 10%, but the test was limited to 133 isolated Scots pines, which are easier to assess than those overshadowed by others.

The accuracies from these studies were better than that obtained here. The difference in accuracy could be attributed to several factors, while some of the errors could be explained by the inadequacy of measurements in both the reference and laser data. For example, the densities of laser data obtained in this study were 1.8–3.4 points per m^2, with about 5–9 points per m^2 in Persson et al. (2002), 5–10 points per m^2 in Hyyppä et al. (2005) and about 4 points per m^2 in Villikka et al. (2007). As a result, the tree crowns were less accurately described by the laser-derived features, and laser hits have a higher probability of missing the treetops in this study. The field measurements were also more accurate in Persson et al. (2002) than those obtained here. In addition, only single-tree species were considered in Hyyppä et al. (2005) and Villikka et al. (2007). Some of the errors could have contributed to the fact that there is a 1–2-year difference between the ALS

Table 3
Bias, standard deviation (SD), correlation coefficient (R) and RMSE between estimated and observed individual tree attributes for training and testing sets with random forests (values are means over 50 runs and values in brackets are standard deviation of the mean. For training, values were calculated from out-of-bag samples).

Attribute	Training				Testing			
	Bias	SD	R	RMSE (%)	Bias	SD	R	RMSE (%)
Tree height (m)	−0.01	1.67	0.93 (0.00)	10.20 (0.23)	−0.02	1.65	0.93 (0.00)	10.03 (0.30)
DBH (cm)	0.00	3.89	0.78 (0.01)	21.58 (0.30)	−0.03	3.88	0.79 (0.01)	21.35 (0.55)
Volume (m^3)	0.00	0.12	0.88 (0.01)	46.54 (1.33)	0.00	0.12	0.87 (0.02)	45.77 (2.30)

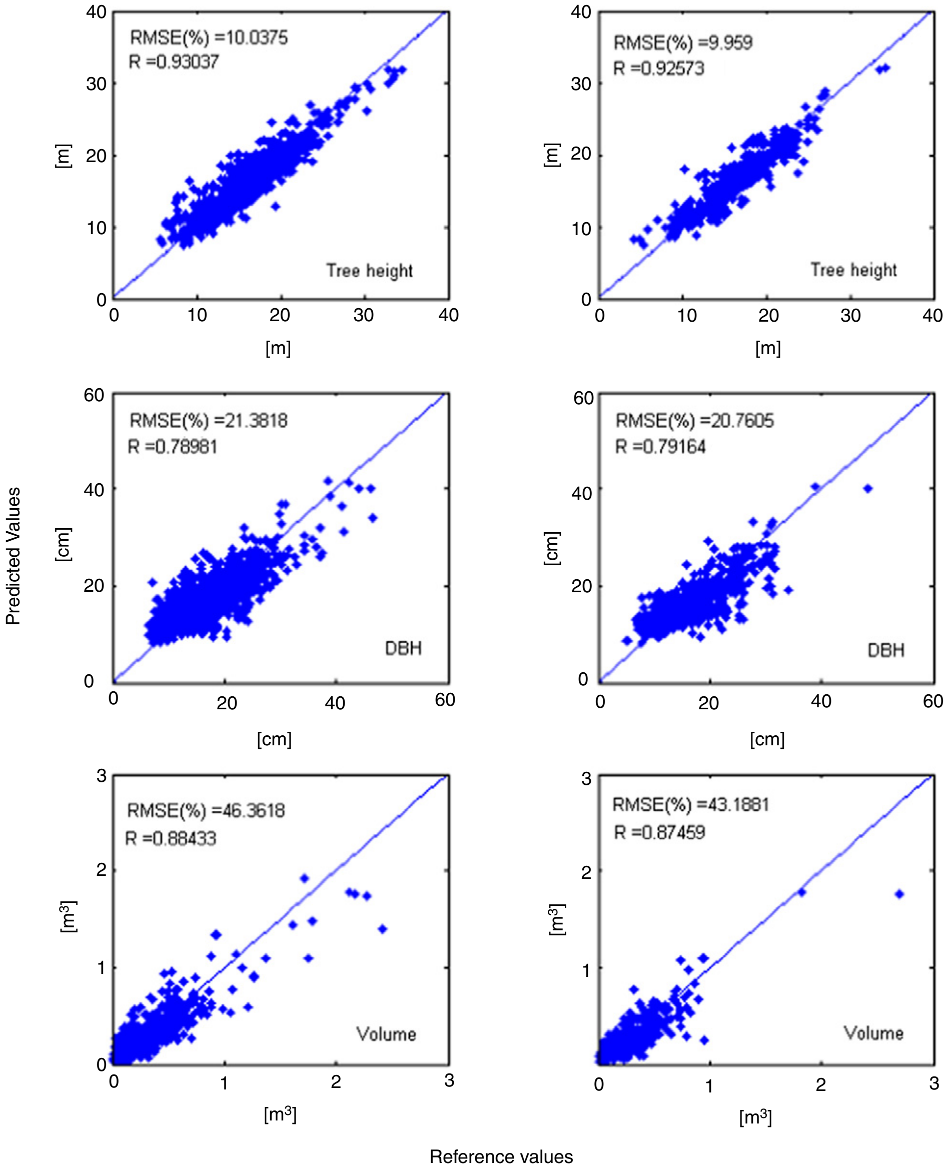


Fig. 6. Scatterplots of predicted versus reference values for tree height, DBH and stem volume, respectively, and for training (left column) and testing (right column). The line indicates a 1:1 relationship.

Table 4
Biases, standard deviations (SD), correlation coefficients (R) and RMSEs between estimated and observed individual tree attributes for testing sets with linear regressions (values are means over 50 runs and values in brackets are standard deviation of the mean, H100 and CW are the maximum laser height and crown diameter respectively, V is the volume calculated by volume equations (Laasasenaho, 1982) with predicted DBH and H100 as inputs).

Attribute	Model	Bias	SD	R	RMSE (%)
Tree height (m)	$=1.23 + 0.96 * \text{H100}$	−0.01	1.59	0.93 (0.01)	9.70 (0.46)
DBH (cm)	$= -0.9 + 1.07 * \text{H100} + 0.48 * \text{CW}$	0.04	3.88	0.79 (0.01)	21.49 (0.72)
Volume (m^3)	$= -0.02 + 1.16 * \text{V}$	−0.00	0.12	0.88 (0.02)	45.95 (2.75)

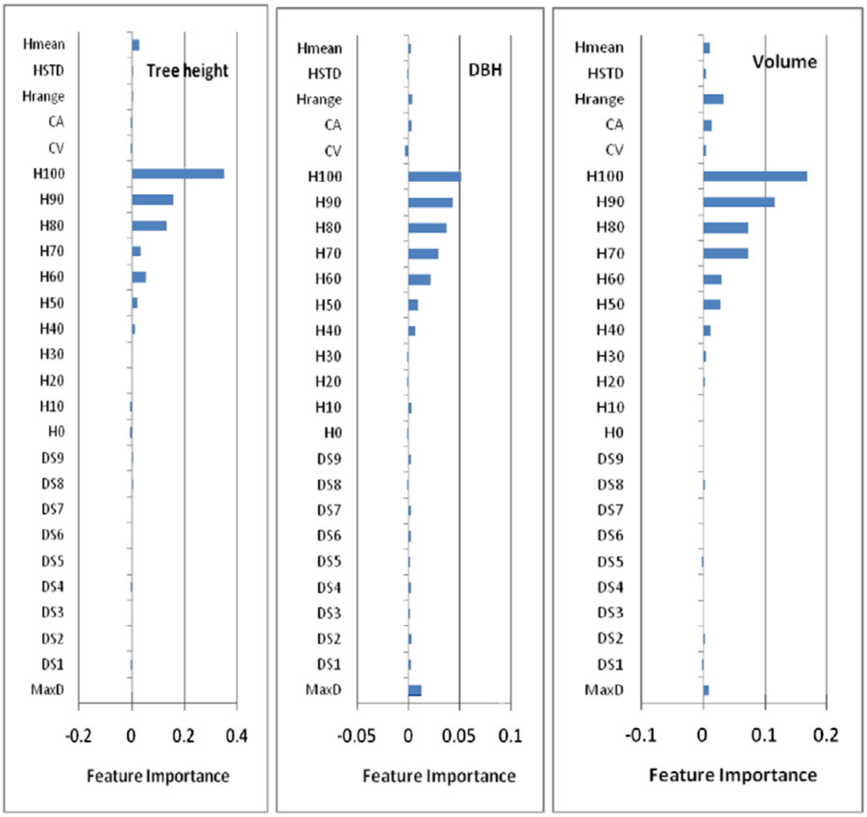


Fig. 7. Importance of laser-derived features for estimation of tree height, DBH and tree volume. Higher values indicate features that are more important to the estimation. Feature descriptions are given in Table 2.

acquisitions (2006) and field measurements (2007 and 2008). The time difference was not corrected for, since growth models were not applicable in this case. A significant portion of the errors in the volume estimates came from old-growth trees, which comprised 1.2% of the total number of stems. These small numbers of samples reduced the possibility that there were enough training samples for this group of trees. Consequently, their estimates were more erroneous. When we kept trees in this group always in the training set, an RMSE of 38% for the volume estimates was obtained, which is comparable to that achieved by Villikka et al. (2007) using laser estimates of the tree height and crown diameter in a linear regression.

The estimation accuracy was similar to that obtained with linear regression conducted with the same datasets and the same tree detection algorithm. The similar accuracy could be explained by the comparable features used in the linear models and regarded as important (indicated by feature importance values, Fig. 7) in the RF. For example, in the RF, the most important feature was H100 for all three tree attribute estimates, while, in linear regression, H100 was also a predictor (directly or indirectly) in all three models. For DBH estimation with the RF, crown diameter or crown area was one of the most important features after higher percentiles of canopy height distribution (Fig. 7), and it was one of the predictors in the linear regression as well. It is worth noting that the volume equations and species information were used in the linear regression for volume estimates. It may improve the estimation accuracy if the same information were used in the RF.

The RF can be considered as a black box method to users, because we could not examine individual regression trees separately. However, it provides several measures to aid in interpretation. One of these is the feature importance, which gives some insights into the method. The ability to provide a measure for feature importance is a very crucial and useful function of RF. Utilization of out-of-bag samples is one of the advantages with RF, as a separate testing set is not needed to evaluate the performance. As a consequence, the number of field measurements and therefore the cost of measurements could be reduced.

There are factors that could affect the estimation, such as the number of regression trees constructed, number of features randomly selected for splitting, and number of training samples. In the present study, the selections of these parameters were based on trial and error. In our experience, the algorithm is not very sensitive to the parameter settings as long as there are enough regression trees constructed. From a practical point of view, the amount of training data is a primarily important issue. In the present study, two thirds of the data (984 trees) were used for training to ensure the quality of the results. In practice, a comprehensive distribution of representative sample sets is more important, i.e., at least several samples from each tree species, and diameter and height class in different forest site types.

Some studies have shown that use of the crown diameter can improve tree volume estimates (Popescu et al., 2003). Crown dimension-related features did not play as important a role as height-related features in the present study. It is likely that the crown diameter estimated from an overhead sensor would be significantly different from that measured from the ground, which is typically used in forest models (e.g. Kalliovirta and Tokola, 2005). Due to the penetration of the pulses and low point density, even the upper crown diameter can also be underestimated. However, as the sampling density of laser scanning increases, tree crowns could be better described, and, as a consequence, the crown dimension could play a more significant role in prediction.

Laser-derived features have been produced based on the original segmentation of individual trees. Thus, the performance of segmentation could impact the extraction of the laser features, especially those features related to the crown shape. In this study, the tree segmentation algorithm tended to merge smaller trees and split taller trees with larger crowns. This may partially explain the tendency that taller trees were more likely to be underestimated and smaller trees overestimated.

The importance of developing species-specific models for forest inventories has been examined (e.g. Holopainen et al., 2008; Packalén and Maltamo, 2006, 2007). Unfortunately, species classification at the individual tree level has not been very successful using relatively low pulse density ALS data and traditional classification methods. The RF could be a potential technique for developing species-specific methods for forest inventories, because it can function as a classifier (Pal, 2005; Ørka et al., 2009) and incorporate categorical information, such as species, with the regression (Breiman and Cutler, 2008; Prasad et al., 2006).

In practice, nonparametric estimation methods are widely used in area-based ALS inventories (e.g. Holopainen et al., 2008; Maltamo et al., 2006; Packalén and Maltamo, 2007), and forest organizations are implementing these methods into operational management planning. In the individual tree-based method, the feasibilities of nonparametric methods have recently been investigated (e.g. Maltamo et al., 2009).

6. Conclusions

This study has shown that the RF technique can provide a stable and reliable solution for predicting the tree height, DBH and volume of individual trees based on both physical and statistical features derived from airborne laser scanning data with an average density of 2.6 points per m^2. The method needs no physical models linking laser-derived features and physical attributes of the trees.

Further studies may examine the effects of the size of the training set and parameter settings on the estimates. Improvement is possible using other operators instead of averaging the regression trees for the final estimation and using a combination of features in the prediction. Tree species classification prior to volume estimation is also expected to improve the accuracy significantly.

One application of RF in forest inventories would be in using data collected with logging machines or terrestrial laser scanners as a reference. With this type of procedure, stem forms and distributions of timber assortments could be estimated in addition to traditional tree characteristics, a possibility needing further investigation.

Acknowledgements

Financial support from the Academy of Finland (projects "Improving forest supply chain by means of advanced laser measurements" and "Forest Tomography") is gratefully acknowledged.

References

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing 33 (Part B4), 110–117.

Bortolot, Z., Wynne, R.H., 2005. Estimating forest biomass using small footprint LiDAR data: an individual tree-based approach that incorporates training data. ISPRS Journal of Photogrammetry and Remote Sensing 59 (6), 342–360.

Brandtberg, T., 2007. Classifying individual tree species under leaf-off and leaf-on conditions using airborne LiDAR. ISPRS Journal of Photogrammetry and Remote Sensing 61 (5), 325–340.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Cutler, A., 2008. Random forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (accessed 20.11.08).

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Falkowski, M.J., Smith, A.M.S., Hudak, A.T., Gessler, P.E., Vierling, L.A., Crookston, N.L., 2006. Automated estimation of individual conifer tree height and crown diameter via two-dimensional spatial wavelet analysis of LiDAR data. Canadian Journal of Remote Sensing 32 (2), 153–161.

Hall, S.A., Burke, I.C., Box, D.O., Kaufmann, M.R., Stoker, J.M., 2005. Estimating stand structure using discrete-return LiDAR: an example from low density, fire prone ponderosa pine forests. Forest Ecology and Management 208 (1–3), 189–209.

Holmgren, J., 2004. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. Scandinavian Journal of Forest Research 19 (6), 543–553.

Holmgren, J., Persson, Å., 2004. Identifying species of individual trees using airborne laser scanner. Remote Sensing of Environment 90 (4), 415–423.

Holopainen, M., Haapanen, R., Tuominen, S., Viitala, R., 2008. Performance of airborne laser scanning- and aerial photograph-based statistical and textural features in forest variable estimation. In: Proceedings of SilviLaser 2008: 8th International Conference on LiDAR Applications in Forest Assessment and Inventory. Heriot-Watt University, Edinburgh, UK. 17–19 September. pp. 105–112.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., Falkowski, M.J., 2008. Nearest neighbour imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment 112 (5), 2232–2245.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., Maltamo, M., 2008. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. International Journal of Remote Sensing 29 (5), 1339–1366.

Hyyppä, J., Inkinen, M., 1999. Detecting and estimating attributes for single trees using laser scanner. The Photogrammetric Journal of Finland 16 (2), 27–42.

Hyyppä, J., Kelle, O., Lehikoinen, M., Inkinen, M., 2001. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. IEEE Transactions on Geoscience and Remote Sensing 39 (5), 969–975.

Hyyppä, J., Mielonen, T., Hyyppä, H., Maltamo, M., Yu, X., Honkavaara, E., Kaartinen, H., 2005. Using individual tree crown approach for forest volume extraction with aerial images and laser point clouds. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (Part 3/W19), 144–149.

Kaartinen, H., Hyyppä, J., 2008. Tree extraction. Final Report of EuroSDR/ISPRS Project, Commission II "Tree Extraction". EuroSDR Official Publication No. 53.

Kalliovirta, J., Tokola, T., 2005. Functions for estimating stem diameter and tree age using tree height, crown width and existing stand database information. Silva Fennica 39 (2), 227–248.

Laasasenaho, J., 1982. Taper curve and volume functions for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae. 108. 74p.

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 Sensing of Environment 67 (1), 83–98.

Lim, K., Treitz, P., Baldwin, K., Morrison, I., Green, J., 2002. LiDAR remote sensing of biophysical properties of northern tolerant hardwood forests. Canadian Journal of Remote Sensing 29 (5), 658–678.

Lim, K., Treitz, P., Wulder, M., St-Onge, B., Flood, M., 2003. LiDAR remote sensing of forest structure. Progress in Physical Geography 27 (1), 88–106.

Magnussen, S., Eggermont, P., LaRiccia, V.N., 1999. Recovering tree heights from airborne laser scanner data. Forest Science 45 (3), 407–422.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., Kangas, J., 2006. Non-parametric estimation of stem volume using laser scanning, aerial photography and stand register data. Canadian Journal of Forest Research 36 (2), 426–436.

Maltamo, M., Mustonen, K., Hyyppä, J., Pitkänen, J., Yu, X., 2004. The accuracy of estimating individual tree variables with airborne laser scanning in boreal nature reserve. Canadian Journal of Forest Research 34 (9), 1791–1801.

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.

Means, J.E., Acker, S.A., Fitt, B.J., Renslow, M., Emerson, L., Hendrix, C.J., 2000. Predicting forest stand characteristics with airborne scanning LiDAR. Photogrammetric Engineering & Remote Sensing 66 (11), 1367–1371.

Næsset, E., 1997. Estimating timber volume of forest stands using airborne laser scanner data. Remote Sensing of Environment 61 (2), 246–253.

Næsset, E., 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment 80 (1), 88–99.

Ørka, H.O., Næsset, E., Bollandsås, O.M., 2009. Comparing classification strategies for tree species recognition using airborne laser scanner data. In: Proceedings of SilviLaser 2009: 9th International Conference on LiDAR Applications for Assessing Forest Ecosystems. Texas A&M Univerisity, College Station, USA. 14–16 October. pp. 46–53.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26 (1), 217–222.

Packalén, P., Maltamo, M., 2006. Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. Forest Science 52 (6), 611–622.

Packalén, P., Maltamo, M., 2007. The *k*-MSN method in the prediction of species specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment 109 (3), 328–341.

Persson, Å., Holmgren, J., Söderman, U., 2002. Detecting and measuring individual trees using an airborne laser scanner. Photogrammetric Engineering & Remote Sensing 68 (9), 925–932.

Popescu, S., Wynne, R., Nelson, R., 2003. Measuring individual tree crown diameter with LiDAR and assessing its influence on estimating forest volume and biomass. Canadian Journal of Remote Sensing 29 (5), 564–577.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9 (2), 181–199.

Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319.

Tomppo, E., 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 92 (1), 1–20.

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

van Aardt, J.A.N., Wynne, R.H., Scrivani, J.A., 2008. LiDAR-based mapping of forest volume and biomass by taxonomic group using structurally homogenous segments. Photogrammetric Engineering & Remote Sensing 74 (8), 1033–1044.

Villikka, M., Maltamo, M., Packalén, P., Vehmas, M., Hyyppä, J., 2007. Alternatives for predicting tree-level stem volume of Norway spruce using airborne laser scanner data. The Photogrammetric Journal of Finland 20 (2), 33–42.

Wallerman, J., Holmgren, J., 2007. Estimating field-plot data of forest stands using airborne laser scanning and SPOT HRG data. Remote Sensing of Environment 110 (4), 501–508.

Xu, P., Jelinek, F., 2009. Random forests in language modeling. http://www.clsp.jhu.edu/people/xp/publications/emnlp04.pdf (accessed 10.10.09).

Yu, X., Hyyppä, J., Kukko, A., Maltamo, M., Kaartinen, H., 2006. Change detection techniques for canopy height growth measurements using airborne laser scanning data. Photogrammetric Engineering & Remote Sensing 72 (12), 1339–1348.

Zhao, K., Popescu, S., Nelson, R., 2009. LiDAR remote sensing of forest biomass: a scale-invariant estimation approach using airborne lasers. Remote Sensing of Environment 113 (1), 182–196.