Remote Sensing of Environment 114 (2010) 1263–1276

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics

Jari Vauhkonen [a],*, Ilkka Korpela [b], Matti Maltamo [a], Timo Tokola [a]

[a] University of Eastern Finland, Faculty of Science and Forestry, School of Forest Sciences, P.O. Box 111, FI-80101 Joensuu, Finland
[b] University of Helsinki, Faculty of Agriculture and Forestry, Department of Forest Sciences, P.O. Box 27, FI-00014 University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 30 October 2009
Received in revised form 22 January 2010
Accepted 23 January 2010

Keywords:
LiDAR
Non-parametric estimation
Nearest neighbor
k-Most Similar Neighbor
Random Forest

ABSTRACT

Forest inventories based on single-tree interpretation of airborne laser scanning (ALS) data often rely on an allometric estimation chain in which inaccuracies in the estimates of the diameter at breast height (DBH) propagate to other characteristics of interest such as the stem volume. Our purpose was to test nearest neighbor imputation by the *k*-Most Similar Neighbor (*k*-MSN) and the Random Forest (RF) methods for the simultaneous estimation of species, DBH, height and stem volume using ALS data. The predictors included computational alpha shape metrics and variables based on the height and intensity distributions in the ALS data. Separate data sets covering 1898 and 1249 dominant to intermediate trees in a typical Scandinavian stand structure were used for training and validation, respectively. RF proved to be a flexible method with an ability to handle 1846 predictors with no need for their reduction. Classification of Scots pine, Norway spruce and deciduous trees showed an accuracy of 78%, and the estimates of DBH, height and volume had root mean square errors of 13%, 3%, and 31%, respectively, when evaluated against the validation data. The two selection strategies implemented here reduced the number of candidate variables effectively without any substantial effect on the accuracy relative to the use of all predictors. Differences in *k*-MSN and RF imputations were marginal when the reduced sets of variables were used. Estimation accuracies could be maintained practically unchanged with only 12.5% of the initial reference data (237 trees), provided the distribution of the observations was similar in the reference and target data. Since we used information collected in the field for extracting the ALS point clouds for individual trees, our results represent an optimal case and should nevertheless be validated against automated tree delineation.

1. Introduction

A measurement and estimation chain that links photogrammetric single-tree measurements with allometric estimation of diameter at breast height (DBH) has motivated several studies in Scandinavia (Ilvessalo, 1950; Jakobsons, 1970; Talts, 1977, Kalliovirta & Tokola, 2005; Korpela & Tokola, 2006; Maltamo et al., 2007). Kalliovirta and Tokola (2005), for example, formulated national and regional species-specific models that used tree height and crown width for predicting DBH in Finland. Provided an accurate digital terrain model exists for the area, tree height can be accurately retrieved from airborne data (e.g. St-Onge et al., 2004; Maltamo et al., 2004). It is known, however, that various factors such as stand density and silvicultural history can affect the relationship between tree height and DBH (Korpela, 2004; Maltamo et al., 2007; Kaitaniemi & Lintunen, 2008). The accuracy at tree level is restricted by the imprecision of the allometric relationships between measurable tree dimensions and the variables of interest, being 10% in terms of root mean squared error (RMSE) in the case of DBH in Finland (Korpela & Tokola, 2006).

Stem volumes are commonly predicted by using DBH and height estimates based on airborne data in species-specific volume equations such as those produced by Laasasenaho (1982) for use in Finland, but the errors in the DBH estimates are compounded when applied to stem volume models, which themselves also include inaccuracy. Maltamo et al. (2007), for example, simulated the accuracy of a single-tree inventory of 472 sample plots on the assumption that all the trees had been found, the tree height estimates were correct with no underestimation and tree species recognition had yielded 100% accuracy. Only DBH was predicted from tree height. Despite the simplifying assumptions that could hardly be justified in a real-world application (cf. Korpela & Tokola, 2006), the simulated RMSE for the stem volume was about 23% at plot level, indicating a need for either additional predictors or an entirely novel estimation approach.

Airborne laser scanning (ALS), which has become a very common data source during the 2000s, allows numerous variables to be extracted in addition to tree height and crown width. Takahashi et al. (2005) and Villikka et al. (2007), for example, used percentile variables based on the tree-level distribution of ALS height values

* Corresponding author.
E-mail address: jari.vauhkonen@uef.fi (J. Vauhkonen).

0034-4257/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2010.01.016

for predicting the stem volume of sugi (*Cryptomeria japonica* D. Don.) and Norway spruce (*Picea abies* L. Karst.), respectively. Chen et al. (2007) introduced "canopy geometric volume", defined as the area of a tree segment multiplied by its height, to estimate tree-level basal area and biomass. The computational volume and complexity of ALS point data were explored further by Vauhkonen et al. (2008) to predict the DBH in Scandinavian tree species. All of these authors concluded that the ability to use additional variables will improve the estimates for the variables of interest relative to models based on tree height and crown diameter or area. Furthermore, various statistical measures (Holmgren & Persson, 2004), variables quantifying the shape and structure of the tree crown (Holmgren & Persson, 2004; Vauhkonen et al., 2009) and intensity metrics (Holmgren & Persson, 2004; Ørka et al., 2009; Vauhkonen et al., 2009; Korpela et al., 2009a) have been developed for estimating species from ALS data.

The increased number of possible predictors requires caution in the estimation phase, however, as collinearity between the variables may cause a parametric model to be unstable. Also, normality and homoscedasticity assumptions need to be met in the case of linear models. On the other hand, in the case of ALS data various non-parametric methods have recently been employed for predicting species-wise stand characteristics and diameter distributions (Maltamo et al., 2006; Packalén & Maltamo, 2007, 2008; Peuhkurinen et al., 2008; Hudak et al., 2008; Niska et al., in press), and also for use at the tree level (Maltamo et al., 2009c). These studies have particularly been focused on nearest neighbor (NN) imputation, in which the estimates for the attributes of interest are produced as weighted averages of the attributes of those reference observations that are similar in terms of a distance metric calculated in the predictor space formed by the independent variables. As such approaches require no prior knowledge of the distribution of the data, their use may be highly relevant when non-linear and possibly diverse relationships exist between the independent and dependent variables.

The most popular alternative in those studies that have combined non-parametric estimation with ALS data has been the Most Similar Neighbor (MSN) method (Moeur & Stage, 1995), in which the nearest reference observations are determined by distances computed in a projected canonical space, and *k*-MSNs (e.g. Maltamo et al. 2006) are the *k* minima of those distances. Niska et al. (in press) compared the *k*-MSN approach with methods based on artificial neural networks for the prediction of plot-level forest volume per species, and reported only small differences in estimation accuracy between the approaches. On the other hand, Hudak et al. (2008) found Random Forest (RF) (Breiman, 2001) applied to the NN search (Crookston & Finley, 2008) to be the most robust and flexible approach for predicting species-specific basal area and tree density, being superior to several other imputation methods, including *k*-MSN. RFs are combinations of numerous (e.g. thousands to tens of thousands) classification trees fitted from a random sample of reference data, and therefore represent a fundamentally different approach from the other NN imputation methods. RFs have been found to be highly flexible, especially for the classification of complex forest phenomena such as succession stages (Falkowski et al., 2009) or mire site types (Korpela et al., 2009a) from ALS data. Nevertheless, the approach is also applicable to regression (e.g. Hudak et al., 2008).

The use of imputation methods places very high requirements on the reference data, as these should cover the entire phenomenon of interest. This means that variable imputation may seem problematic, especially at the level of single trees. Maltamo et al. (2009c) nevertheless used the *k*-MSN approach for predicting tree-level characteristics from a reference data set comprising only 133 trees. They found the *k*-MSN estimates to be generally more accurate than parametric sets of models constructed simultaneously by the Seemingly Unrelated Regression (SUR) approach, with tree-level RMSEs of 5, 2 and 11% for DBH, height, and volume, respectively. The result is promising, although based on a local data set and ignoring species identification as the data applied to Scots pine trees only. According to the simulations of Korpela and Tokola (2006), a species recognition accuracy of 95% (three species groups in Finland) is required in an allometric estimation chain based on species-specific national–regional equations. Recent studies in Finland suggest that accuracies from 50% (Vauhkonen et al., 2008) up to 95% (Vauhkonen et al., 2009) can be expected, the likely level being about 80% (Korpela et al., 2009b). In non-parametric estimation, however, the species estimate can be produced simultaneously with other characteristics, i.e. the choice of allometric equation is implicit. Thus, the effect of misinterpreting a species might not be obvious as far as the accuracy of predicting stem dimensions is concerned.

We tested the *k*-MSN and RF imputation methods for the simultaneous estimation of species, DBH, height, and merchantable stem volume, which are the principal input parameters for forest management systems in Finland. We used separate reference and validation data sets containing 1898 and 1249 trees, respectively, to evaluate the estimates based on different combinations of a total of 1846 initial ALS-based variables. We also examined the effect of variation in the number of reference observations and the size of the neighborhood in order to assess the feasibility of the approaches under different conditions.

2. Methods

2.1. Forest area and data

The experiments were carried out near the Hyytiälä forest station in southern Finland (61°50′N, 24°20′E), where rotation periods are typically 80–130 years and trees on mineral soils attain a height of 21–33 m, depending on site fertility. Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* L. Karst.) dominate and form both mixed and pure stands. The birch (*Betula* spp.) stands are younger than 40 years, because of the past silvicultural preference for conifers. Isolated birches are also to be found in the coniferous stands. Mineral soils with gentle slopes prevail and the elevation is 135–198 m above sea level. The basins contain lakes, open mires, spruce mires and pine bogs. The mires have largely been drained, and the state-owned forests are managed on a commercial basis. The forest area studied here is of size 2 × 6 km and encompasses a large number of permanent forest plots in both managed and pristine forests, on mineral and peat soils and on pristine and drained mires. Aerial images are available for 1946–2009 and four ALS data acquisition campaigns have taken place from 2004 onwards. We used ALS data from 2006 and 2007 (Table 1), when the campaigns had an average pulse density of 6–8 per m^2 (55% overlap). Point spacing and point patterns vary, mainly because of overlapping swaths and the zigzag-scanning pattern. The data were stored in 207-byte binary records that contained the full pulse geometry including the position and attitude of the ALS system for accurate range calculations. The ALTM3100 post-processing software supported this feature (comprehensive format), but the ALS50 data required additional post-processing to include the trajectory data. Both 2006 and 2007 data were used in the collection of the tree point data, but only 2007 data were included in the later analysis.

Table 1
Characteristics of the ALS data sets.

Instrument	ALTM3100	ALS50-II
Date	July 25, 2006	July 4, 2007
Pulse frequency	100 kHz	115.8 kHz
Scan frequency	70 Hz	52 Hz
Footprint	25–28 cm	17–18 cm
Flying height, nominal	1000 m a.s.l.	930 m a.s.l.
Scan angle	±14°	±15°
Air humidity, 2 m	48–52%	60–75%
AGC	–	8 bits

We used two subsets of forest plots in the area, a set of 59 circular, 0.04-ha plots (as also used by Korpela et al., 2007a), and a set of 18 rectangular plots (0.08–0.24 ha, in total 2.2 ha). These were measured in May 2007 and July 2007, respectively, employing a photogrammetric–geodetic mapping procedure (Korpela et al., 2007b). In it, the treetops were first positioned using multi-scale template matching of aerial images to serve as field control points for the positioning of the other targets by trilateration and/or triangulation. All trees were recorded in the field with respect to species, DBH and crown status, and 30–60% of the trees on each plot were measured for height and crown base height (CBH). The best available height observation was computed for each tree, being either the field measurement, the height obtained in the treetop positioning, or an estimate derived from the plot-level regression curve for the unseen trees. Each tree was assigned a relative height by reference to the local dominant height. Stem volumes were calculated using DBH and height in the species-specific equations presented by Laasasenaho (1982). The dominant species, Scots pine and Norway spruce, were treated as separate classes in the analysis, whereas the minor number of aspens (*Populus tremula*), alders (*Alnus* spp.), and rowans (*Sorbus* spp.) were lumped together with the birches to form a class of deciduous species. Only trees that were discernible in the images and/or visualized in the ALS data were included in the tree sets. Most suppressed and intermediate trees with relative heights of less than 60% were rejected.

The trees measured on the circular plots ($N = 1898$) were used consistently as a reference data set throughout the study, while the rectangular plot data ($N = 1249$) were used for validation. The main characteristics of the data sets are presented in Table 2. The distributions of the data match each other relatively well, especially in the case of smaller trees (Fig. 1), but there were two plots in the validation data that included observations of considerably larger trees than were present in the reference data set, as can also be seen in Fig. 1. As it was assumed that imputing the attributes of the two plots would cause higher error levels due to mismatch in the reference data, the 1173 trees matching the reference data set and the 76 deviating trees were treated as separate subsets of the validation data. The three validation data sets, consisting of 1173, 76, and combined 1249 trees, will be referred to as validation data sets A, B and C, respectively.

2.2. Extracting ALS point data

The collecting of point data on a particular tree is an ill-posed task especially in the lower parts of crowns where neighboring trees overlap. Here the extraction of ALS data and derivation of features was incorporated into a crown modeling procedure (Korpela, 2007) in which a curve of revolution is fitted to the ALS point clouds near the treetop using a weighted least squares (WLS) adjustment. This WLS technique reduces the underestimation of the crown envelope that is inherent in this kind of modeling.

In practice, 871 field measurements (from 2002) were first used to construct species-specific regression models that predicted the crown width (d_{cr}) from DBH and tree height. The regression estimate for d_{cr} was multiplied by a factor of 1.2 and in this way transformed into an initial (somewhat overestimated) approximation of the crown model. A cylinder of this width was then used to collect the crown points from the combined sensors for the crown modeling.

Crown radius *y* at the height *x* down from the treetop was solved by fitting a three parameter curve $y = a + b \times h \times x^c$ to the point cloud using the WLS adjustment. Initial values for the parameters *a*, *b* and *c* were set to fulfill the criteria $a_0 + b_0 \times h \times 1^{c_0} = 1.2 \times d_{\mathrm{cr}}$. In the adjustment, the ALS points were treated as observations of crown radius and the observations that fell outside the crown envelope were weighted by a factor of 5. The length of the crown model was fixed, and the CBH was always 40% down from the top. ALS points inside the envelope or within one RMSE of it were saved for feature computations. Echoes below the 40% height were stored inside a cylinder having a diameter equal to the maximum crown width + the RMSE of the fit. The modeled crown widths and numbers of echoes extracted from the data are shown in Table 2.

2.3. Outline of the analysis

The *k*-MSN and RF methods were compared using the same ALS-based variables that have recently been used for predicting species and stem dimensions under Finnish conditions (Vauhkonen et al., 2008, 2009; Maltamo et al., 2009c; Korpela et al., 2009b). The variables were grouped into those considered informative with respect to the estimated crown dimensions, particularly volume (predictor group denoted by *CrVol*), area (*CrArea*), length (*CrLength*) and complexity (*CrCompl*), and those based on the height (*Height*) and intensity (*Int*) distributions of the tree point clouds. Also, a number of statistical transformations of the variables were calculated and used as predictors. The total number of candidate variables was 1846. The ability of the RF algorithm to manage all the variables (Breiman, 2001) was tested, but variable reduction was performed for the *k*-MSN method in order to reduce data redundancy and improve the overall interpretability of the model, as also with the RF method. Two variable reduction procedures based on internal importance measures applied to the RF algorithm were implemented for that purpose. The reduced number of variables was used to study the effects of the size of the neighborhood, i.e. the value of *k*. Finally, we used different data selection strategies to simulate reference data consisting of a lower number of field observations in order to assess the sensitivity of the approaches to the amount of reference data.

Table 2
Means and standard deviations (in parentheses) of the attributes of Scots pine, Norway spruce and deciduous trees in the reference and validation data sets. "Echoes" refers to the number of "only" or "first of many" returns within the tree crown.

Data set	Variable	Pine	Spruce	Deciduous trees
Reference data	Number of observations	459	529	261
	DBH, cm	23.3 (8.1)	19.2 (7.3)	15.2 (6.5)
	Height, m	19.3 (4.8)	17.2 (5.2)	16.2 (4.1)
	Volume, dm^3	493.5 (465.2)	329.6 (376.0)	188.0 (236.4)
	Age, a	79.7 (33.5)	60.9 (31.5)	45.8 (33.5)
	Crown diameter, m	3.2 (0.8)	3.3 (0.7)	3.1 (1.2)
	Echoes	102.4 (78.7)	124.5 (117.6)	93.6 (110.8)
Validation data	Number of observations	855	722	321
	DBH, cm	17.7 (5.6)	17.8 (6.1)	14.2 (5.7)
	Height, m	15.6 (3.7)	16.1 (4.2)	15.3 (4.1)
	Volume, dm^3	227.5 (196.8)	249.0 (216.1)	152.3 (147.6)
	Age, a	44.6 (16.3)	55.0 (22.4)	44.1 (14.4)
	Crown diameter, m	2.9 (0.7)	3.2 (0.7)	2.9 (1.1)
	Echoes	90.4 (80.7)	107.6 (92.6)	87.4 (97.4)

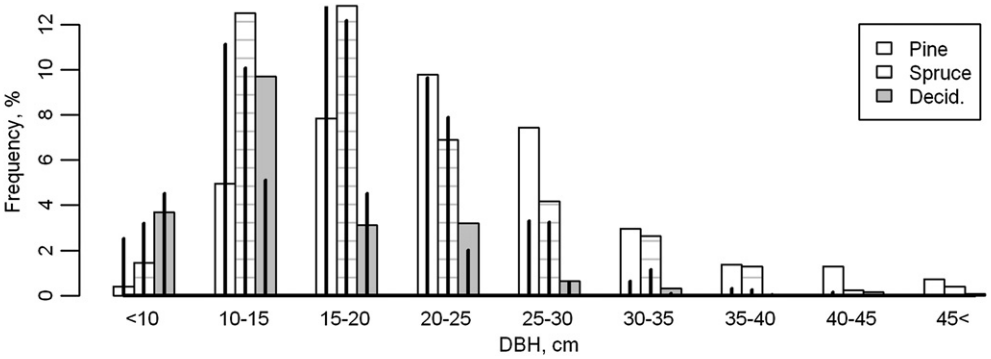

Fig. 1. Distribution of DBHs in the data sets (vertical lines — reference data, bars — validation data).

2.4. Imputation methods

The theoretical background to the methods will not be covered in this context, but the reader is referred to the original works on the *k*-MSN and RF methods by Moeur and Stage (1995) and Breiman (2001), respectively, or to Eskelson et al. (2009) or Sironen (2009), for example, for a brief description of non-parametric methods in general. The focus in the following will be upon the present implementation of the methods. The imputations were carried out in the *R* statistical computing environment (http://www.r-project.org), using a package called *yaImpute* (Crookston & Finley, 2008), version 1.0-10. Another package, *randomForest* (Liaw & Wiener, 2002), version 4.5-30, was used within *yaImpute* for the RF method, and also separately for selecting the variables, as described below.

The field attributes of interest (species and stem dimensions: DBH, height, and volume) were produced simultaneously for each target tree using *yaImpute*. As categorical variables were not allowed with the *k*-MSN method, the species attribute was coded into dummy variables *pine*, *spruce*, and *decid*, which were then imputed using *k*-MSN. The number of canonical vectors to be used with the *k*-MSN method was set by the function (Crookston & Finley, 2008, p. 3).

The original RF algorithm was modified such that *yaImpute* was required to build separate forests for each dependent variable in the case of multivariate imputation (Crookston & Finley, 2008). Altogether 2000 trees (500 per imputed attribute) were fitted in the each RF run, and $\sqrt{N}$ variables, where N is the total number of possible predictors, were randomly permuted at the nodes of a classification tree. The values for these parameters were as recommended by Diaz-Uriarte and Alvarez de Andrés (2006), and were not tested any further in this context. The RFs were fitted in the classification mode in all cases, i.e. continuous variables were converted internally into classes forcing RF to build classification trees for them. As reported by Hudak et al. (2008), only minor effects on the imputed characteristics were observed when regression trees were used instead of classification.

In both the *k*-MSN and RF imputations, the user needs to decide the size of the neighborhood, i.e. the value of the parameter *k*. Increasing *k* improves the precision of the imputation but shifts the prediction towards the sample mean, thereby increasing the bias of the extreme values of the imputed variables (see Sironen, 2009; Eskelson et al., 2009). Values of *k* from 1 to 10 were tested here. When a parameter *k* other than $k=1$ was considered, the imputed value was produced as a weighted average over the *k* neighbors, or as a mode, where the count was the sum of the weights rather than each having a weight of 1, as in the case of a categorical variable. The weight w_{ij} of the reference tree j for target tree i was determined according to distance d_{ij}:

$$w_{ij} = \frac{1}{1+d_{ij}}. \qquad (1)$$

The random element in both the RF imputations and the simulated data was taken into account by running 50 iterations in each case. The number of iterations was restricted by the computational burden involved. A single RF run with the 1846 initial variables took around 15 min with a 1.9 GHz notebook processor when the R environment was operated under Linux Ubuntu 8.04. The number of iterations was considered adequate for RF imputation, however, as the number of fitted trees was also assumed to stabilize the result (cf. Diaz-Uriarte & Alvarez de Andrés, 2006). After the iterations, the resulting characteristics were calculated as the mode in the case of RF, whereas in the case of *k*-MSN the species was determined correspondingly but the continuous variables were calculated as means of the characteristics for the species concerned. If the mode could not be defined (two equally sized classes in the iteration results), either of the two classes was randomly selected to serve as the result.

2.5. Candidate variables for the imputation

2.5.1. Tree crown dimensions

To include information on tree allometry, the alpha shape metrics (Vauhkonen et al., 2008, 2009), i.e. various measures of crown volume (predictor group denoted by *CrVol*), area (*CrArea*), length (*CrLength*) and complexity (*CrCompl*), were determined from the tree-level point data, relying on the concept of 3-D alpha shapes (Edelsbrunner & Mücke, 1994). An alpha shape is derived from the Delaunay triangulation of a 3-D point cloud such that each simplex of the triangulation is compared with the specified alpha value in the computation phase. Those simplices which have an empty circumsphere with a squared radius larger than the defined alpha value are removed. Thus, an alpha shape can be regarded as an alpha-weighted Delaunay triangulation.

Variables extracted from alpha shapes have been successfully applied to the prediction of DBH and species by Vauhkonen et al. (2008, 2009), and their relation to actual tree characteristics is further explained in those publications. In the present work we calculated interior (*int_*) and exterior (*ext_*) volumes and numbers of solid components (*nsc_*) in alpha shapes generated with point data from above 50, 60, …, 90, and 95% (*h50_*, *h60_*, etc.) of the maximum height, using alpha values of 0.5, 1.5, and 4 m (*a0.5_*, *a1.5_*, and *a4_*). Secondly, the corresponding variables were calculated using the point data on 10% height zones surrounding heights of 55, 65, …, and 95% (*z55_*, *z65_*, etc.) of the maximum height using the above alpha values. An approximate 3-D convex hull volume (*cvol_*) was calculated within the above height zones using an extremely large alpha value (99 999 m). In addition to variables of the basic form, variables normalized with reference to the maximum height value (*norm_*) were included. Following given notation, *int_h50_a0.5* would denote “interior alpha shape volume calculated using point data from above 50% of the maximum height using an alpha of 0.5 m”. The computations regarding the previous variables were carried out using the functionality of the Open Source library CGAL (Da & Yvinec, 2007).

Tree crown area was estimated at different height levels as the area of a 2-D convex hull, generated from point data below 10, 20, …, 90, 95, and 100% (*a10*, *a20*, etc.) of the maximum height, and areas relative to the maximum area were calculated as *a10*/*a100*, *a20*/*a100*, …, *a95*/*a100* (*ar10*, *ar20*, etc.). Also, each area value was normalized by reference to the maximum height (prefix *norm_*). An estimate for crown height was calculated using an approach presented by Vauhkonen (in press). In the approach, return frequencies are extracted using 10% height bins with bin values of 5, 10, …, 95% of the maximum height. The value within a height bin is normalized using the largest bin value, resulting in values between 0 and 1. Values less than 0.1 are considered to be zeros. The crown height is defined as the lowest point above the first of 2 sequential zero bins below the maximum. Both the estimates of crown height (*crheight*) and crown length (*crlength*), calculated by subtracting the crown height value from the maximum height, were included.

2.5.2. Variables derived from the height distribution

Following Maltamo et al. (2009c), who found ALS height metrics calculated at both the tree and plot level useful for tree-level predictions, we included the predictor group *Height* among the candidate variables. This group included the maximum (*hmax*), mean (*hmean*), and standard deviation (*hstd*) of the ALS height values, the vegetation ratio (*vege*), defined as the proportion of all height values above a ground threshold, and several variables describing the distribution of the height values, namely height percentiles (denoted by *h5*, *h10*, *h20*,…, *h90*, *h95*) and corresponding proportional densities (*p5*, *p10*, *p20*,…, *p90*, *p95*) for 5, 10, 20, …, 90, and 95% of the maximum height. Percentile h_q is the height from the ordered list of ALS heights corresponding to A_q calculated as:

$$A_q = \frac{q}{100}\sum_{i=1}^{n} h_i, \qquad (2)$$

where q is the desired percentile value, A_q is the proportion of return heights cumulated below q, h_i is the height of return i, and n is the number of returns in the area considered in the calculation (Korhonen et al., 2008). Percentile variables normalized with reference to the maximum height (prefix *norm_*) were included in addition to variables of the basic form.

In order to reproduce possible species-specific differences in these variables (cf. Brandtberg, 2007), we calculated them with respect to different echo categories: “all echoes” (suffix *_all*), “first and only echoes” (*_0*), “first of many echoes” (*_1*), “last of many echoes” (*_2*), “first echoes” (*_first*), and “last echoes” (*_last*), where the last two categories included “first of many” and “last of many” echoes, respectively, with “first and only echoes” duplicated in both.

All these variables were calculated at both the tree and plot levels with a small letter in the notation (e.g. *h5*) denoting a tree-level variable and a capital letter (e.g. *H5*) a plot-level variable. In all cases a ground threshold value of 0.5 m was used. The plot-level variables were calculated from point data extracted from a 250 m^2 circle (r = 8.92 m) centered on the coordinates of the tree top. This circle size was selected subjectively with a view to the purpose of the plot-level characteristics, which was to describe the status of the trees within their immediate neighborhood.

2.5.3. Variables derived from the intensity distribution

Although ALS intensity observations are not solely related to the reflectance properties of the vegetation (e.g. Moffiet et al., 2005), different intensity metrics have been shown to be useful in species recognition (Korpela et al., 2009b; Ørka et al., 2009). The intensity features employed here (predictor group *Int*) were adaptations of those used by Korpela et al. (2009b). We calculated the minimum (*imin*), maximum (*imax*), mean (*imean*), and standard deviation (*istd*) of all the intensity values of the first returns with a height value more than the ground threshold, the mean intensities of the returns at 0–10%, 10–20%, 20–30%, and 30–40% of the distance down from the tree top (*imean#*, # = 1–4), and transformations of these *imean#1*/*imean#2* (denoted *imean#1#2*). In addition, percentiles (*ip5*, *ip10*, *ip20*, …, *ip90*, *ip95*) of the intensity distribution were calculated in a manner that to the height percentiles above. The percentiles were further normalized with respect to *imax* (prefix *norm_*). All intensity features were derived from “first returns” (see the previous section), assuming that they were least affected by intra-crown transmission losses (e.g. Gaveau & Hill, 2003).

2.5.4. Statistical transformations of the predictors

Finally, various transformations of the independent variables (Maltamo et al., 2006, 2009c) were included. The natural logarithm (prefix *ln_*) and square root (*sqrt_*) of each variable were included, and the cubic roots (*cubic_*) of the crown volume variables (see Vauhkonen et al., 2009) were also calculated.

2.6. Variable reduction

Two procedures, both based on RF and its ability to bootstrap the data, were tested for reducing the number of variables, the purpose in both of them being to search for the best predictors by fitting RF separately to predict species and species-specific stem dimensions. A secondary purpose was to examine the number of predictors needed and to reduce them aggressively due to the high number of initial candidate variables.

As the first variable reduction procedure, we utilized the *R* package *varSelRF* (Diaz-Uriarte, 2009), the idea of which is to fit several forests by successively eliminating 20% of the least important variables determined by internal measures of importance (Diaz-Uriarte, 2009; Breiman, 2001). After the iterations, the out-of-bag (OOB) error rates for all the fitted forests are examined, and the variables that are considered to represent the best solution according to the given criteria are selected. In this case “the best solution” was determined as the one with an error rate within one standard error of the minimum of all forests that included the smallest number of variables (cf. Diaz-Uriarte, 2009). The result of *varSelRF* was then iterated 10 times, preserving the most frequent variables, where “frequency” was determined as the mode of occurrence of those variables present on more than one occasion.

The second variable selection procedure was adapted from Hudak et al. (2008), who iterated RF discarding the least important variable in each run until a single predictor variable remained. The final models were, however, built on combinations of variables consistently important between the iterations. As with the *varSelRF* procedure above, Hudak et al. (2008) used the internal importance measures in RF to determine the importance of the variable. In the present case we relied on previous research and assumed that 9–10 variables similar to those used here would be required for a reliable estimate of species (Vauhkonen et al., 2009; Korpela et al., 2009b), whereas 2–3 variables would be enough for predicting the DBH (Vauhkonen et al., 2008). We therefore iterated RF in a manner similar to Hudak et al. (2008) but retained the 15 most important predictors of species and 5 most important predictors of stem dimensions in each run, assuming that the other stem dimensions would require up to the same number of predictors as DBH. As a result, up to 60 individual variables could theoretically be selected in a single RF run, although in practice there were less, because the same variables were the strongest predictors of several attributes to be estimated. As with the first method, we performed 10 iterations, but eventually retained only those variables that were present in all of them.

Finally, we performed a sensitivity analysis to find out effects the number of predictors had on the obtained results. In it, we randomly selected predictors from the combined subset produced by the reduction strategies detailed above, and examined imputation accuracy produced by the RF and *k*-MSN methods using these predictors. Furthermore, we examined the effect of correlation structure within the selected variables by identifying groups with high and low inter-correlations and performing a similar analysis for these groups. The groups were formed by clustering the Pearson correlation matrix using a *k*-means algorithm (Hartigan and Wong, 1979). The variable selection was repeated 50 times, with the *k* value fixed to 1, and the averages of these iterations are reported.

2.7. Simulation of reference data

Data sets consisting of 50%, 25%, and 12.5% of the observations in the initial data set, i.e. 949, 474, and 237 reference trees, were generated by applying three selection strategies. First data selection strategy (denoted *a-random-plots*) corresponded to the manner of collecting reference data from randomly sampled field plots, in that entire plots were selected at random until the required number of trees was obtained. In the second approach (*b-random-trees*), trees were selected randomly from the pooled tree set. In the third case (*c-systematic*), it was assumed that the ALS data was acquired prior to the field work, serving the role of an auxiliary information source for the selection of the reference data (cf. Hawbaker et al., 2009; Maltamo et al., 2009a). The trees were selected systematically from the initial reference data sorted by tree species and height, and within each species, the number of observations to be selected was determined by reference to the proportion of that species in the validation data.

2.8. Evaluation criteria and performance measures

The performance of the species classification was evaluated with the overall classification accuracy (%), user's accuracy (proportion of observations assigned to correct classes), and producer's accuracy (proportion of field observations classified correctly), whereas RMSE and bias were used for the accuracy of the stem dimensions:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}, \text{ and} \tag{3}$$

$$bias = \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)}{n}, \tag{4}$$

where n is the number of observations, and x_i and $\hat{x}_i$ are the reference and imputed stem dimensions (DBH, height, or volume), respectively, for tree i. The relative RMSEs were calculated by dividing the absolute RMSE values by the mean of the reference dimension.

3. Results

The analysis carried out led to a wide-ranging set of results of which we present a selection. First, we present the detailed results of the RF imputation carried out using all available predictors, noting that the main results were rather similar when other imputation alternatives (*k*-MSN and RF with the reduced number of predictors) were considered. Second, we present the results of reducing the number of variables. Third, we give the results for the effects of the size of the neighborhood and the amount of reference data on both *k*-MSN and RF imputation methods when using the reduced sets of variables. Finally, we present some conclusive results with respect to the comparison of the methods.

3.1. RF imputation of species, DBH, height and stem volume using all predictors

The results obtained using RF imputation with all predictors are presented in Table 3. When evaluated by cross-validation in the reference data set itself, accurate and practically unbiased estimates for all the characteristics were obtained. The estimates generally included errors of less than 10%, the RMSE for stem volume being about 19% (Table 3), but the accuracies were considerably lower when separate validation data sets were applied.

The accuracies were notably higher in the validation data A consisting of those trees likely to have a similar observation within the reference data. Here the species recognition accuracy was about 78%, and the RMSE was 13%, 3%, and 31% for the DBH, height and volume, respectively (Table 3). In particular, the bias in the stem volume estimate increased relative to the reference data set.

The accuracy of estimating tree species did not differ appreciably between the validation data sets (Table 3), whereas the accuracies of DBH, tree height and stem volume were lower in B and C. The estimates saturated at the level of the largest reference observations (Fig. 2). Otherwise, the imputed values corresponded to those measured in the field relatively well, especially in the case of tree height (Fig. 2b). The accuracy of DBH and volume decreased with the tree size, however (Fig. 2a, c).

The results of the tree species classification are presented in Tables 4 and 5. Trees could usually be successfully determined to species in the case of the conifers (Table 4). Altogether 83% and 90% of the field-observed pines and spruces, respectively, were distinguished from other species in validation data C, but only 44% of the observed deciduous trees were correctly classified, being confused with both conifer species. The expected coniferous trees, on the other hand, included both other conifer species and deciduous trees (user's accuracy of 75–79%), whereas 85% of the trees assigned to the class of deciduous trees were correctly classified (Table 4).

The classification of the smallest trees was less accurate than that of the larger ones (Table 5), but recollecting that mainly deciduous trees filled the two smallest diameter classes (Fig. 1), the inaccuracy is likely due to species rather than to size. On the other hand, the largest

Table 3
Accuracies obtained by imputing attributes of different data sets using RF with all predictors.

Characteristic		Data set			
		Reference	A	B	C
Species	Classification accuracy	93.3	78.4	73.7	78.1
DBH	RMSE, cm	1.4 (8.2%)	2.5 (12.9%)	6.3 (17.5%)	2.8 (14.2%)
	Bias, cm	0.1 (0.7%)	0.5 (2.7%)	1.6 (4.4%)	0.6 (2.9%)
Height	RMSE, m	0.5 (3.0%)	0.6 (3.2%)	2.6 (8.9%)	0.8 (4.6%)
	Bias, m	0.0 (0.1%)	0.00 (0.0%)	1.8 (6.2%)	0.1 (0.6%)
Volume	RMSE, dm^3	41.6 (18.6%)	91.7 (30.1%)	553 (38.5%)	162.8 (44.3%)
	Bias, dm^3	2.8 (1.2%)	16.6 (5.6%)	236 (16.4%)	29.9 (8.1%)

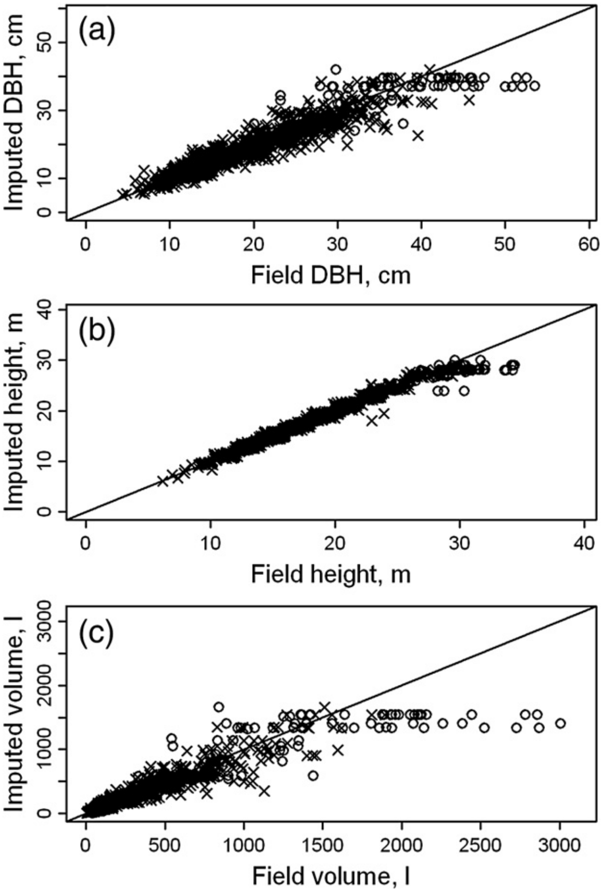

Fig. 2. Measured vs. imputed DBH (a), height (b), and volume (c) using RF with all predictors (crosses — observations in validation data A, black dots — B).

of all (DBH>45 cm) were not accurately classified either, though there were no corresponding observations in the reference data.

The accuracy of relating the ALS characteristics to the field attributes was affected by increasing stem density, causing inaccuracies in the derived characteristics due to interlaced tree crowns. As seen in Fig. 3, the accuracy of the volume estimates may be related to plot-level basal area. Absolute inaccuracy in the imputed volume was higher for those trees that were located on plots with a basal area above 22 m^2/ha.

3.2. Reduction in variables

The variables selected as a result of reduction procedures #1 and #2 are presented in Tables 6 and 7, respectively. Altogether 130 variables (7% of the candidate variables) were preserved using the first approach, whereas 24 (1%) variables were included in the second.

In most cases, several transformations of a predictor variable were included. Altogether 58 (Table 6) and 8 (Table 7) separate variables, together with 72 and 16 transformations of these, were included by reductions #1 and #2, respectively. Among those selected by #1, crown volume variables were most often involved (31 separate variables), followed by height distribution variables (9), intensity distribution variables (9), crown area variables (7) and one crown complexity and one crown length variable (Table 6). Procedure #2 gave 4 crown volume variables, 3 height distribution variables and an intensity variable (Table 7).

Table 4
Species classification results using RF with all predictors and the full validation data (C). Overall accuracy 78.1%.

Estimated species	Observed species			Total	User's accuracy, %
	Pine	Spruce	Deciduous		
Pine	383	42	85	510	75.1
Spruce	64	478	61	603	79.3
Deciduous	12	9	115	136	84.6
Total	459	529	261	1249	
Producer's accuracy	83.4%	90.4%	44.1%		

Table 5
Effect of tree size on species classification accuracy with the full validation data (C).

DBH, cm	*N*	Accuracy, %
<10	69	60.9
10–15	339	66.4
15–20	297	88.5
20–25	248	87.5
25–30	153	81.7
30–35	74	75.7
35–40	33	69.7
40–45	21	71.4
45<	15	66.7

Usually a single ALS predictor was involved in several field attributes and stem dimensions of more than one species, but there were exceptions. Variables selected for species were normally not included for other field attributes. Interestingly, reduction #1 selected a number of variables for predicting the height of deciduous trees, but none of these was included in the second. On the other hand, all the variables selected by reduction #2 were among those included by #1.

According to the sensitivity analysis performed with respect to the number of predictors (Fig. 4), a combination selected by either procedure was sub-optimal, however, as equal or even better accuracies could be obtained using a number of predictors other than 24 or 130 selected by the variable reduction procedures. With RF, the tree species could be estimated with slightly better accuracy using only 10–20 predictors selected randomly from the initial 130 candidates. A combination of some 40–60 variables was also enough to obtain similar volume imputation accuracies than using the initial predictor set, with both RF and *k*-MSN. However, the model noise caused by higher number of predictors did not have any major effect on the accuracy of any imputation alternative, as confirmed in both reference and validation A data. In addition, RF with all available predictors was always the best or at least nearly the best alternative. The *k*-MSN method in particular resulted in better accuracies with a number of predictors less than 130, indicating that further results of the approach could be improved by 2–4 percentage points using variable selection procedure optimized for it.

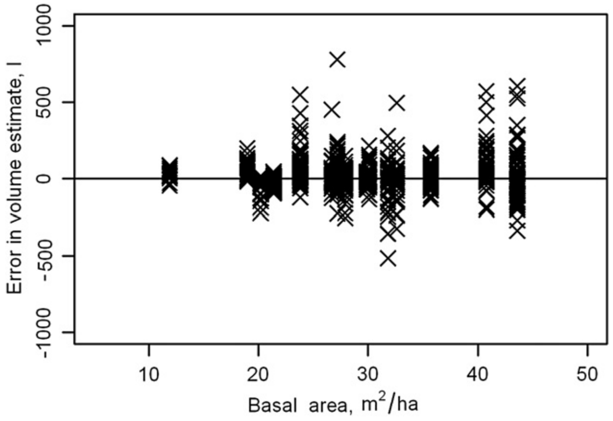

Fig. 3. Plot-level basal area vs. error in the imputed volume when the imputation was carried out using RF with all predictors.

Table 6
Predictors of field attributes retained when using procedure #1 for reducing the number of variables. The order of the variables does not indicate their relative importance.

Variable group	Variable	Transformations[a]	Field attribute[b]	*R* group[c]
Int	imean3	i, ln, sqrt	**Sp**, d2, v1	2
Int	imean4	i, sqrt	**Sp**	2
Int	ip40	i, ln, sqrt	**Sp**, d2	2
Int	ip50	i, ln, sqrt	**Sp**	2
Int	ip60	i, ln, sqrt	**Sp**	2
Int	ip70	i, ln, sqrt	**Sp**	2
Height	norm_h20_0	i, ln, sqrt	**Sp**	2
Height	P60_1	i, ln, sqrt	**Sp**	2
Height	P70_1	i, ln, sqrt	**Sp**	2
CrVol	ext_h50_a0.5	i, ln, sqrt, cubic	**d1**, **d2**, **d3**	1
CrVol	ext_h60_a0.5	i, ln, sqrt, cubic	**d1**, **d2**, **d3**	1
CrVol	ext_h70_a0.5	i, ln, sqrt, cubic	**d1**, **d3**, d2	1
CrVol	ext_h80_a0.5	i, ln, sqrt, cubic	**d1**, d3	1
CrVol	ext_h50_a1.5	i, ln, sqrt, cubic	**d1**, **d3**, d2, v2	1
CrVol	ext_h60_a1.5	i, ln, sqrt, cubic	**d1**, d2, d3	1
CrVol	ext_h70_a1.5	i, ln, sqrt, cubic	**d1**, **d3**, d2	1
CrVol	ext_80_a1.5	ln, sqrt, cubic	**d1**, d3	1
CrVol	ext_h70_a4	ln	**d1**, d2, d3	1
CrVol	ext_h80_a4	i, sqrt	**d1**, d2, d3	1
CrVol	int_cvol_h50	i, ln, sqrt, cubic	**d1**, **d2**, **d3**, v2	1
CrVol	int_cvol_h60	i, ln, sqrt, cubic	**d1**, **d2**, **d3**	1
CrVol	int_cvol_h70	i, ln, sqrt, cubic	**d1**, **d3**, d2	1
CrVol	int_z85_a4	i, ln, cubic	**d1**, d3	1
CrVol	int_h50_a4	i, ln, sqrt, cubic	**d2**, d3	1
CrVol	int_h60_a4	i, ln, sqrt, cubic	**d3**, d1, d2	1
CrVol	int_h70_a4	ln, sqrt, cubic	**d3**, d1	1
CrArea	a100	i, ln, sqrt	**d2**, d3, v2	1
CrArea	a95	i, ln, sqrt	**d2**, d3, v2	1
CrArea	a90	i, ln, sqrt	**d2**, d3	1
CrArea	a80	i, ln, sqrt	**d2**, d3	1
Height	hmax	i, ln, sqrt	**h1**, **h2**, h3, d1, d2, d3, v2	1
Height	hmax_first	i, ln, sqrt	**h1**, **h2**, h3, d1, d2, d3, v2	1
Height	Vege_last	ln	**h3**, d2	2
Height	norm_h10_3	ln	**h3**	2
Height	h10_3	ln	**h3**	2
CrVol	ext_h95_a4	ln	**h3**	2
CrVol	ext_h90_a4	ln	**h3**	2
CrVol	ext_z55_a4	ln	**h3**, v2, v3	1
CrVol	int_z55_a4	ln	**h3**	2
CrVol	int_z75_a1.5	ln	**h3**	2
CrCompl	nsc_z65_a1.5	i	**h3**	2
CrArea	norm_a30	ln	**h3**	2
CrArea	norm_a10	ln	**h3**	2
CrArea	a100	ln	**h3**, v1	
CrVol	ext_z85_a4	ln	**v1**, d2, h2	2
CrVol	int_z65_a4	ln	**v1**, d2, h3	1
CrVol	int_h95_a0.5	ln	**v1**, h2, h3	2
CrVol	int_z65_a0.5	ln	**v2**	2
CrVol	int_z55_a1.5	ln	**v2**	2
CrVol	ext_cvol_h95	i	**v2**	2
CrVol	ext_cvol_h85	ln	**v2**, **v3**, v1	2
CrVol	ext_cvol_h55	ln	**v3**	2
CrVol	ext_z75_a4	i	**v3**, h2, h3	1
CrLength	norm_crheight	ln	**v3**, v2, d3	2
Height	Vege_3	ln	**v3**, d3	2
Int	imean14	ln	**v3**, d2	2
Int	imean23	ln	**v3**, d2, d3, h3	2
Int	norm_ip80	ln	**v1**, d2, d3, v2	2

[a] *i* denotes the basic form of the variable.

[b] *Sp* denotes species, *d* DBH, *h* height, *v* volume, and numbers 1–3 pine, spruce, and deciduous trees, respectively. The field attributes for which the variable was included are marked in bold; narrow letters indicate that the variable was considered in the selection procedure but not accepted on the final criteria.

[c] Correlation structure group according to the *k*-means partition of the Pearson correlation matrix. 1 = high and 2 = low inter-correlation.

Table 7
Predictors of field attributes retained when using procedure #2 for reducing the number of variables. The order of the variables does not indicate their relative importance.

Variable group	Variable	Transformations[a]	Field attribute[b]
Height	P70_1	i, ln, sqrt	**Sp**
Height	norm_h20_0	i, ln, sqrt	**Sp**
Int	ip50	sqrt	**Sp**
CrVol	ext_h50_a1.5	i, ln, sqrt, cubic	**d1**, v1, v3
CrVol	ext_h50_a0.5	i, ln, sqrt, cubic	**d2**, **v2**, v3
CrVol	ext_h60_a0.5	sqrt, cubic	**d3**, d1, d2, v2, v3
CrVol	ext_h70_a0.5	i, ln, sqrt, cubic	**v3**, d1, d3
Height	hmax	i, ln, sqrt	**h1**, **h2**, **h3**, v1, v2

[a] *i* denotes the basic form of the variable.

[b] *Sp* denotes species, *d* DBH, *h* height, *v* volume, and numbers 1–3 pine, spruce, and deciduous trees, respectively. The field attributes for which the variable was included are marked in bold; narrow letters indicate that the variable was considered in the selection procedure but not accepted on the final criteria.

The inter-correlation in the variables, on the other hand, did not have any particular effect on the accuracies. Groups of 79 and 51 variables (Table 6) were identified to have different correlation structure, such that variables selected to predict stem dimensions were more often inter-correlated than those selected for species. Thus, the ability of these groups to predict species and stem volume, for example, was fundamentally different, but as a function of the number of predictors they behaved in practically similar way. For clarity, the curves for these groups were left out from Fig. 4.

3.3. Effects of the k parameter in NN estimation

The effects of the *k* parameter were studied using data with a reduced number of predictors (Tables 6 and 7). With respect to the accuracy of estimating stem dimensions, all the imputation methods behaved in almost the same way as a function of *k* (Fig. 5). Increasing *k* first sharply reduced the inaccuracy, which then stabilized and finally started to increase in some cases. The poorest accuracies in DBH and stem volume estimates were always obtained using a low value of *k* with the *k*-MSN method, the difference between the minimum and maximum accuracies being around 3–5 percentage points (Table 8), while the accuracies of the RF imputations were more stable, or rather diminished along with the value of *k*, especially in the case of tree species (Fig. 5, Table 8). Using the RF method with all available predictors, clearly best species recognition accuracy was obtained with *k* = 1 in both the reference and validation data sets (Fig. 5). The imputation alternatives which employed only 24 predictor variables seemed to be more sensitive to the parameter *k*, in the sense that the range between the extreme values was generally wider (Table 8). The trend in the accuracy did not differ markedly between the reference and validation data sets (Fig. 5).

Unlike the accuracy, the precision of the stem volume estimate in particular decreased with increasing *k*, this effect being especially clear in validation data A in terms of bias (Fig. 6). The smallest bias was always obtained with *k* = 1 and the largest with *k* = 10, except in the case of using RF with all available predictors, in which *k* = 5 produced slightly smaller bias than other values of *k* (Table 8; Fig. 6). The bias increased most when the RF imputations with variables selected by reduction procedure #2 were considered, and more so for tree height and stem volume than for DBH.

3.4. Effects of the amount of reference data

The accuracy of the imputation methods was tested in the simulated reference data using the reduced sets of variables (Tables 6 and 7) and *k* fixed at a value of 1. The results are illustrated in Fig. 7, which shows the errors in species and volume estimates in validation data set A. Neither the species nor the volume error was clearly reduced as a function of the amount of reference data. Errors in species classification in particular remained at the level achieved with the full reference data. The inaccuracies in the data selected by the *c-systematic* strategy, however, clearly increased upon reduction, being generally higher than those for the other two data selection strategies.

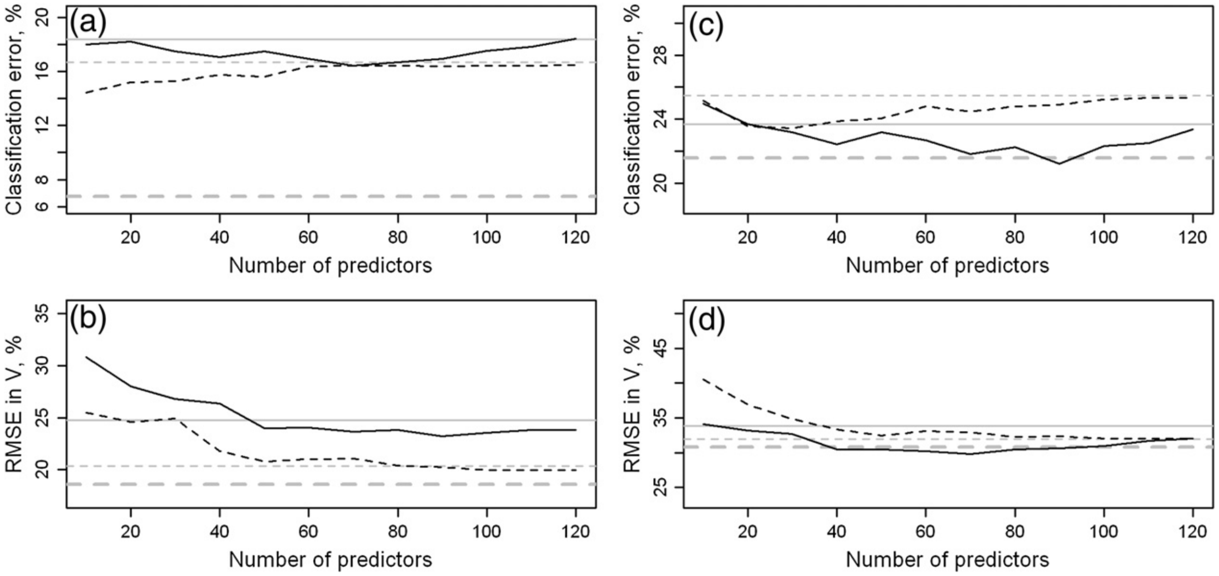

Fig. 4. Error in imputed species (a — reference data, c — validation data A) and volume (b, d) as a function of the number of predictors used (solid lines — *k*-MSN, dashed lines — RF). Grey lines indicate the error levels with the initial 130 variables, and thick dashed lines those using RF with all available predictors.

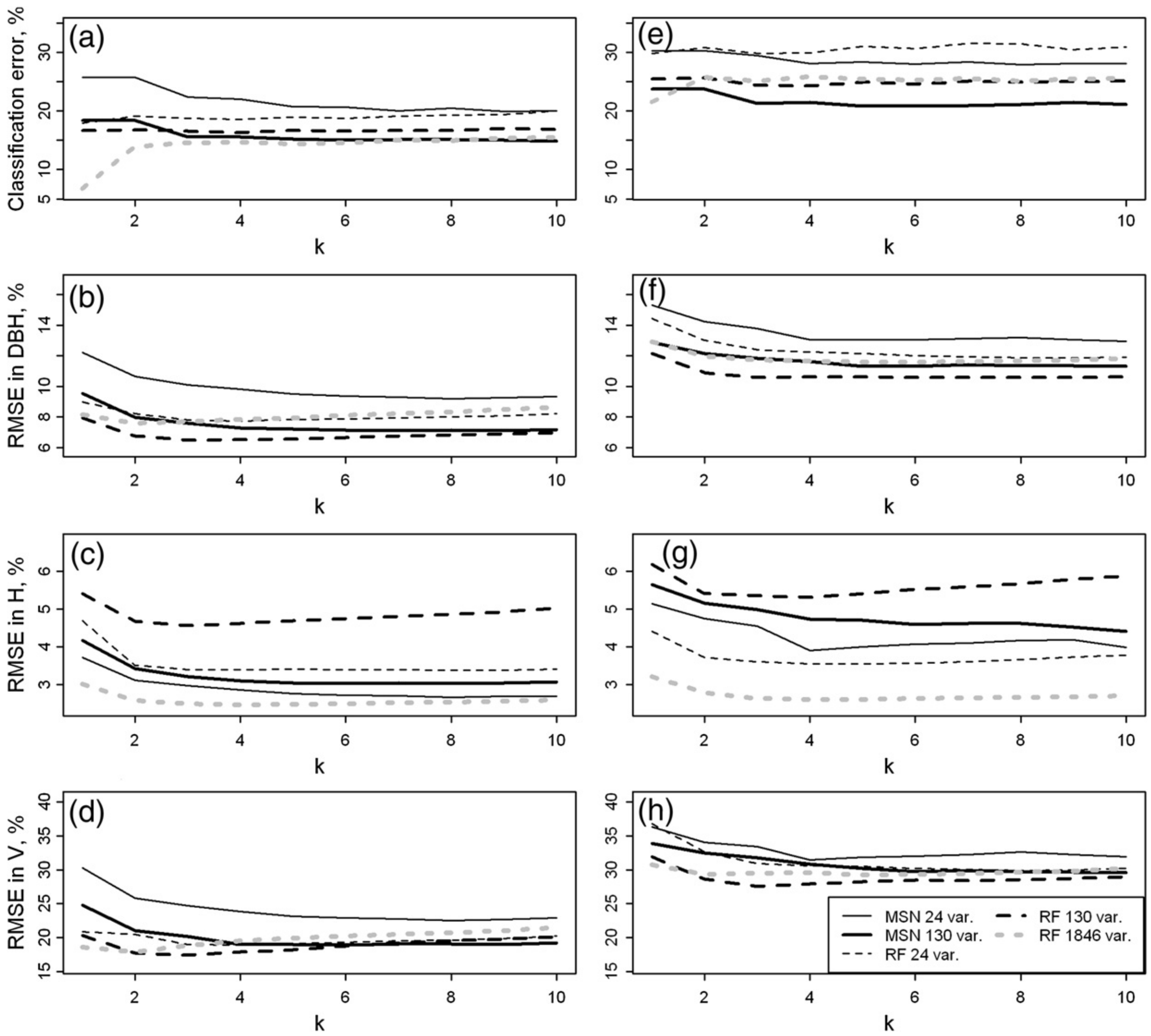

Fig. 5. Accuracy of the imputed characteristics as a function of *k*. Species classification error when considering the reference (a) and validation data A (e), and correspondingly, relative RMSEs of DBH (b, f), height (c, g), and volume (d, h).

Table 8
Ranges of the reliability characteristics of tree species (*Sp*) and stem volume (*v*) estimates when using different imputation alternatives in validation data A. The value of *k* is given in parentheses.

Method[a]	*Sp* accuracy, %	*v* RMSE, %	*v* bias, %
RF_{1846}	74.2 (4)–78.4 (1)	29.2 (5)–30.8 (1)	5.3 (5)–5.8 (10)
RF_{130}	74.3 (2)–75.7 (4)	27.5 (3)–31.9 (1)	4.2 (1)–5.9 (10)
RF_{24}	68.5 (7)–70.2 (1,3)	29.9 (8)–36.8 (1)	1.8 (1)–4.0 (10)
MSN_{130}	76.3 (1,2)–78.9 (10)	29.6 (10)–33.9 (1)	2.0 (1)–4.5 (10)
MSN_{24}	69.7 (1,2)–72.8 (8)	31.4 (4)–36.4 (1)	2.5 (1)–4.0 (10)

[a] The subscript indicates the number of predictor variables used.

3.5. Comparison of the imputation methods

According to the results in Table 8, species classification accuracies of 69.7–78.9% and RMSEs of 29.6–36.4% were obtained for stem volume using *k*-MSN, whereas the corresponding figures for the RF method were 68.5–78.4% and 27.5–36.8%, respectively, when considered in validation data A. Thus, *k*-MSN resulted in a slightly better accuracy with respect to predicting tree species, the model with 130 variables being the most accurate (Fig. 5a, e; Table 8). The poorest *k*-MSN imputation was also slightly better than the poorest result obtained using RF imputation. On the other hand, RF produced both the best and the worst result in the estimation of stem volume (Table 8).

Using $k=1$, RF with all predictors was the best method in all cases except for DBH imputation (Fig. 5), the reduced sets of variables generally resulting in higher accuracies when the value of *k* was increased (Fig. 5; Table 8). The biases in the stem volume estimates were in practice at the same level using both *k*-MSN, regardless of the number of predictors, and RF with 20 variables, but considerably higher in the case of RF with 130 or 1846 variables.

By contrast with the method used for imputation, the number of predictor variables affected the results in the sense that better accuracies were mainly obtained by using a higher number of predictors, the difference being up to 10 percentage points (Fig. 5). The *k*-MSN and RF models with 130 predictors were in most cases at least as accurate as RF with 1846 predictors (Fig. 5; Table 8). The accuracies (Fig. 4b–d, f–h) and precisions (Fig. 6) of the stem dimensions were generally poorer in the validation data sets, but the relative differences between the imputation methods remained approximately unaltered. The differences between the methods generally diminished with increasing values of *k*.

The accuracy of imputing tree height differs considerably from that of the other characteristics in most cases. RF with all available variables was always most accurate, whereas RF with 130 predictor variables stands out from the other cases in achieving poorer results in terms of both RMSE (Figs. 4c, g) and bias (Fig. 5b, e) of tree height. On the other hand, RF with both 24 and 1846 predictors was in practice as accurate in the validation data as in the modeling data when estimating the tree height, whereas the *k*-MSN methods produced poorer results (Fig. 4c, g) in the validation data.

The accuracy of RF with all the predictors decreased along increased *k* and with the reduced number of reference data, effects which were not as evident with other methods (Fig. 7). With respect to tree species, the methods maintain their relative position independent of the amount of reference data (Fig. 6a–c). In the case of stem volume, *k*-MSN was more instable, especially with 130 predictor variables (Fig. 6d–e).

4. Discussion

Forest planning systems in Finland function at the level of single trees, because tree-level models for volume, growth, mortality and regeneration have advantages over more aggregated ones (e.g. Lämås & Eriksson, 2003). The detection of individual trees from airborne data would produce optimal input data for such systems, unless it had substantial limitations caused by the undetected trees and the inaccuracies in species recognition and allometric estimation of the attributes of interest (Korpela & Tokola, 2006; Korpela et al., 2007a). Detectable trees constitute 90–100% of our commercial timber (e.g.

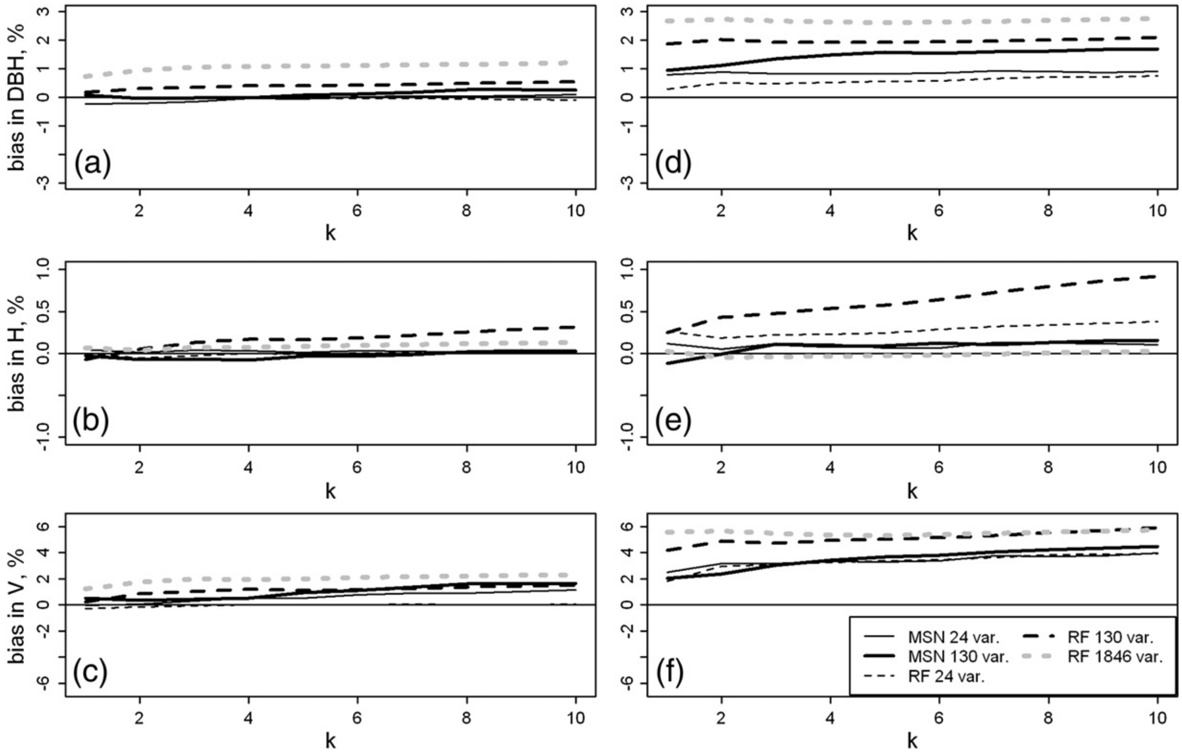

Fig. 6. Bias of DBH (a and d referring to the reference and validation data A, respectively), height (b, e), and volume (c, f) in the imputations as a function of *k*.

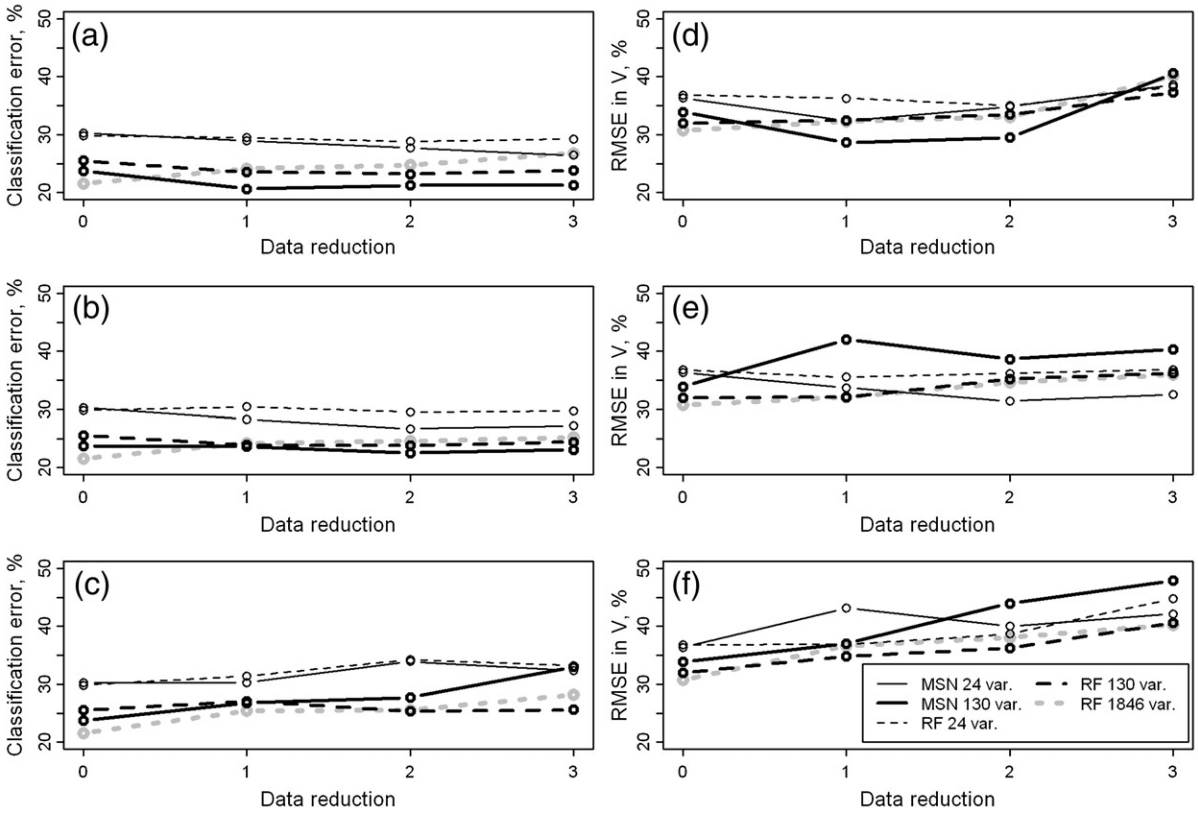

Fig. 7. Error in imputed species (a, random-plots; b, random-trees; c, systematic data selection) and volume (d, e, f) as a function of the amount of reference data, when using validation data A. Data reduction level 0 on the horizontal axis denotes the initial data set (1898 reference observations).

Korpela, 2004), however, which motivates applications for timber cruising. Accuracies of 80–95% in distinguishing between Scandinavian commercially important species have been reported recently (Holmgren et al., 2008; Korpela et al., 2009b; Vauhkonen et al., 2009). ALS data are usually regarded as having a greater potential for characterizing the canopy structure than passive optical imagery, but images have been favored for species recognition (Holmgren et al., 2008). Recent studies (Korpela et al., 2009b; Vauhkonen et al., 2009) have focused on deducing species solely from ALS data, however, because the use of several data sources would complicate the inventory system further, increase costs and entail difficulties from the operational point of view. Error propagation in the allometric estimation (Korpela et al., 2007a) has nevertheless remained a problem.

We employed *k*-MSN and RF imputation methods here for the simultaneous non-parametric estimation of the species and stem dimensions using ALS variables. Depending on the method, the value of the parameter *k* and the set of predictor variables, species classification accuracies ranging from 68% to 79% and estimates for DBH, height and stem volume with RMSEs of 11–15, 3–6, and 28–37%, respectively, were obtained with validation data set A, which matched the characteristics of the reference data. Earlier, Korpela et al. (2007a) reported tree-level RMSEs of about 20, 5, and 46%, respectively, for the stem dimensions considered here, following an estimation chain based on allometric species-specific models (Kalliovirta & Tokola 2005) that employ tree height and maximum crown width as predictors. Species determination was done visually with an accuracy of 95%. On the other hand, Maltamo et al. (2009c) reported accuracies of 5, 2, and 11% for stem dimensions in cross-validated reference data consisting of 133 pines. The accuracies were slightly lower in the reference data of this study, but the data here consisted of considerably more observations in three species classes.

Selection of the estimation method is restricted by the requirements for the attributes to be estimated. The requirement set out here was to estimate categorical variables (species) and continuous variables (stem dimensions) simultaneously, and the *k*-MSN and RF imputation methods were selected based on experiences gained from previous studies on the estimation of forest attributes from ALS data (Maltamo et al., 2009c; Hudak et al., 2008). Maltamo et al. (2009c) found *k*-MSN to be more accurate than the parametric SUR approach for estimating tree-level characteristics from ALS data, whereas Hudak et al. (2008) concluded that RF was a more robust and flexible approach than several other NN-imputation methods for plot-level attributes. We found no clear differences between the *k*-MSN and RF methods, when estimation accuracy was considered. The variable subsets selected using RF favor that method, however, the inoptimality in the *k*-MSN estimates assessed to be about 2–4% only. The accuracy of the imputation decreased considerably upon extrapolation beyond the reference data, but this defect was common to all the estimation alternatives. Besides the NN-imputation techniques considered here, supervised learning approaches such as support vector machines (Vapnik & Kotz, 2006; Drucker et al., 1997) or neural networks (see Hyyppä et al., 2000) could be used for classification and regression in a corresponding manner, although no large improvements can be expected relative to simple k-NN classifiers (Venables & Ripley, 2002, pp. 346–349; Niska et al., in press).

Both the methods used here require a set of field reference observations, within which the user is required to set the size of the neighborhood for the nearest observations, i.e. the value of the parameter *k*. This affected the results in that higher values improved the accuracy of the estimates at the expense of the increased bias, as commonly observed in previous studies (e.g. Eskelson et al. 2009). While *k* of 3–5 have been found to be optimal with remote sensing data (Muinonen et al., 2001; Packalén & Maltamo, 2007; see also Eskelson et al., 2009), Sironen (2009) used $k > 10$ for predicting tree growth from tens of thousands of potential reference observations. It may therefore be concluded that the selection of *k* depends on the problem and the available data (cf. Eskelson et al. 2009), and also on the user's preferences, in the sense of whether the user is more interested in predictive accuracy or preserving the variance structure of the reference observations in the imputed values. The present results nevertheless suggest that a *k* value of 2–4, depending somewhat on the method, could be proposed for a situation similar to that considered here. Values in this range did not result in biased estimates, and larger *k* values did not improve the results to any appreciable extent. It was clear, however, that increasing the value of *k* affected the results of all the methods tested in approximately the same way.

The need to acquire reference data is a crucial element to be considered in non-parametric estimation. The estimates of Korpela et al. (2007a), for example, were based on national–regional models formulated using measurements made on permanent sample plots used in the National Forest Inventory (NFI) in Finland (Kalliovirta & Tokola, 2005), so that the approach can basically be applied without need for field work. Also, a single stand-specific observation may be enough to calibrate the estimates produced by those models to local conditions. In the present work, only a slight increase in inaccuracy was discovered when a rather small reference data set of 237 trees was used, and since *k* was fixed at 1 in our simulations, it can be argued that even better results could have been gained by using a larger neighborhood. We had local reference data (same climate zone, forest owner, management regime) available, however, and therefore the results cannot be generalized to cases with wide geographical extent. Also, the characteristics of the reference and validation sets matched each other fairly well, so that the random selection of reference data worked well. Selection using auxiliary information about the species and height composition was also tested (cf. Hawbaker et al., 2009; Maltamo et al., 2009a), but this tended to impair the results, partly due to the fact that tree height did not serve very well to describe the variation in stem volume. More importantly, the random selections were iterated 50 times, and the average outcome of these iterations was taken as the result, which means that the result does not describe the worst possible case and does not apply when the validation data can be expected to differ markedly from the reference data.

Since data representing local conditions and tree species proportions are usually required for the non-parametric estimation approach (e.g. Packalén & Maltamo 2007), the possibility of using existing field data such as NFI plots (Maltamo et al., 2009b) would reduce costs of field measurements. The present approach places further demands on the reference data, however, as trees that are positioned in the field are required. To the best of our knowledge, no such exist at least in Finland. An inventory based on the current approach would therefore include the collection of reference data, which definitely places constraints on its applicability. A tolerable amount of reference data was required here, but it should be noted that a larger or a more complex area would presumably require more extensive data. On the other hand, an alternative means of producing tree-level characteristics could be to theoretically predict theoretical diameter distributions using an area-based approach (Packalén & Maltamo, 2008), in which case the data acquisition costs are likely to be lower (positioning of plots vs. trees). The usefulness of producing accurate tree-level data by means of remote sensing should therefore be assessed carefully with respect to alternative methods and the costs involved (cf. Kangas, in press). In all, a number of open questions remain and constitute topical problems as ALS data with high pulse densities that allow the detection of individual trees are expected to become more commonly available (Hyyppä et al., 2008).

Two strategies for selecting the most important variables were implemented here, both based on internal importance measures incorporated into the RF algorithm (Breiman, 2001; see also Hudak et al., 2008), but rather than using these directly, we iterated the solutions and selected those variables most frequently present in the iterations. Only 10 iterations were performed, but it was observed that the same variables would have been preserved also after 4–5 iterations only. The reduction procedure may be considered to have been successful on the following grounds:

1. Reduction strategies #1 and #2 both reduced the number of candidate variables effectively, retaining only 7% and 1% of the candidate variables, respectively.
2. The most essential variables with respect to the attributes of interest were retained, and the imputations with the reduced sets of variables resulted in similar accuracies as compared with the RF model with all predictors.
3. The variables selected were logical in terms of tree allometry and related to the phenomena to be predicted. Also, the variables selected by reduction #2 were a subset of those selected by #1.

Both variable reduction strategies resulted in sub-optimal sets of predictors, however, as the results of the sensitivity analysis suggest an adequate number of variables to be 10–90, depending on the method and attribute to be estimated. Inclusion of additional predictors did not improve the model performance, although the increased model noise had practically no effect on that either, as verified in the separate validation data. The most important predictors were apparently included in the 130 variables retained by the reduction #1 approach, and the result could possibly be refined by including additional criteria for further reduction.

The reduction procedure #2 bears resemblance to that of Hudak et al. (2008), although the latter iterated the RF solution until a single predictor remained and then combined the final models by selecting from those variables with consistently high importance. Our initial solutions included a considerably higher number of independent variables, because separate sets of variables were selected for predicting species and species-specific stem dimensions. The most frequent variables were eventually retained, however, but no attempt was made to ensure the optimality of the number of predictors included for each field attribute. The partly contradictory results obtained using the set of 130 variables may have been due to this fact. Packalén and Maltamo (2007), for example, selected their variables by optimizing the cost in terms of a weighted average of the RMSEs of the dependent variables, the weights being assigned subjectively according to importance weighting. A similar weighting function could have improved our results, although they were in any case fairly good, at least when compared with those obtained using all predictor variables.

Most of the variables employed here have been introduced very recently, and therefore the results obtained add to the current state of the art. At a minimum, crown volume measurements taken at different height levels that describe the tapering of the crown and its maximum height were required here for estimating the stem dimensions. Furthermore, height and intensity distribution variables were employed for estimating species. Previously, Vauhkonen et al. (2008) used a corresponding combination of variables for predicting DBH in linear regression, and a similar correspondence between the estimated tree heights and those measured in the field to that observed here was reported by Maltamo et al. (2004) and St-Onge et al. (2004), for example. Instead of alpha shape metrics (Vauhkonen et al., 2009), it was intensity and structure features (Korpela et al., 2009b) that were the most important with respect to species. The species classification results obtained here are in line with the observations of Korpela et al. (2009b), who classified more than 13 000 trees in the same area. Their later study (Korpela et al., 2010), however, indicates that even higher classification accuracies can be obtained using intensity features normalized with reference to the scanning range. Interestingly, variables based on the plot-level distribution of ALS height values were used here for tree species, the same variables having previously been found to improve the accuracy of predicting tree attributes and quality characteristics (Maltamo et al., 2009c). A number of statistical transformations of the independent variables were always included.

Previous studies have incorporated variable reduction after extracting some 40–70 independent variables (Packalén & Maltamo, 2007; Hudak et al., 2008; Korpela et al., 2009a), so that selecting the best ones would be a laborious although not impossible manual task. Here, the total of 1846 initial variables absolutely precluded any manual selection approach. Basically the number of candidate variables could have been reduced by excluding some variables or groups of variables on the basis of an expert opinion (cf. Korpela et al., 2009a), but the inclusion of the alpha shape metrics, for example, resulted in numerous very similar candidate variables (cf. Vauhkonen et al., 2009), the rating of which for predictor importance would have been difficult. Since selecting the most important out of a set of tens or hundreds of candidate variables is always an ambiguous task and depends on the problem to be solved (Packalén & Maltamo, 2007; Korpela et al., 2009a), we acknowledge the ability of RF to handle all independent variables. The high number of predictor variables apparently adds redundant information, but from a view of any practical application, the ability to avoid delicate variable selection process would be an advantage (see also Peuhkurinen et al. 2008). No instability due to the high number of predictors was discovered here when the results were validated against separate data sets. The variables were computed automatically, and although their number increases the processing time, practically no man-hours are involved no matter how high the number of possible predictors may be.

Finally, we stress that the method used to extract the ALS data for the trees was not an automatic individual tree detection method, but based on trees that were mapped in the field. Only dominant, co-dominant and intermediate trees were considered. Our results may have been affected by the tree delineation, however, and therefore they tend to give an over-optimistic impression of the methods. However, automatic tree detection methods also find trees mainly from the dominant layer (e.g. Persson et al., 2002), and since interlaced tree crowns affected to our results, they should be applicable to an automatic procedure. Yet, this needs to be validated. Assessment of plot-level accuracies of the imputation methods is a theme that remains to be taken up in future work.

Acknowledgements

The airborne data used in this study was funded by a consortium formed by the Finnish Forest Research Institute, Metsähallitus, Metsämannut, Metsäteho, Stora Enso, UPM-Kymmene and the universities of Helsinki and Joensuu. We thank Dr. Markus Holopainen and Prof. Pauline Stenberg for allowing the collection of field reference data during their courses at the University of Helsinki, and the forestry students who carried out the actual work. Funding for this research was received from Academy of Finland (Project 12193: High Resolution Remote Sensing Potential to Measure Single Trees and Site Quality). In addition to the previous, we are grateful to the three anonymous reviewers for their attempts to improve the quality of this paper.

References

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.

Chen, Q., Gong, P., Baldocchi, D., & Tian, Y. Q. (2007). Estimating basal area and stem volume for individual trees from lidar data. *Photogrammetric Engineering & Remote Sensing*, *73*, 1355—1365.

Crookston, N. L., & Finley, A. O. (2008). yaImpute: An R package for kNN imputation.*Journal of Statistical Software*, *23*(10) 16 p. Available at http://www.jstatsoft.org/v23/i10

Da, T. K. F., & Yvinec, M. (2007). 3D alpha shapes. In CGAL editorial board (Ed.), *CGAL user and reference manual*, 3.3 edition Available at http://www.cgal.org/Manual/3.3/doc_html/cgal_manual/packages.html#Pkg:AlphaShapes3 [accessed 23 June 2009].

Diaz-Uriarte, R. (2009). *varSelRF: Variable selection using random forests. R package version 0.7-1.* http://ligarto.org/rdiaz/Software/Software.html Accessed 25 September 2009.

Diaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest.*BMC Bioinformatics*, *7*(3) 13 p. Available at http://www.biomedcentral.com/1471-2105/7/3

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155—161.

Edelsbrunner, H., & Mücke, E. P. (1994). Three-dimensional alpha shapes. *ACM Transactions on Graphics*, *13*, 43—72.

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

Falkowski, M. J., Evans, J. S., Martinuzzi, S., Gessler, P. E., & Hudak, A. T. (2009). Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. *Remote Sensing of Environment*, *113*, 946—956.

Gaveau, D. L. A., & Hill, R. A. (2003). Quantifying canopy height underestimation by laser pulse penetration in small-footprint airborne laser scanning data. *Canadian Journal of Remote Sensing*, *29*, 650—657.

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, *28*, 100—108.

Hawbaker, T. J., Keuler, N. S., Lesak, A. A., Gobakken, T., Contrucci, K., & Radeloff, V. C. (2009). Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design.*Journal of Geophysical Research*, *114*(G00E04) 11 pp.

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

Holmgren, J., Persson, Å., & Söderman, U. (2008). Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images. *International Journal of Remote Sensing*, *29*, 1537—1552.

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

Hyyppä, J., Hyyppä, H., Inkinen, M., Engdahl, M., Linko, S., & Zhu, Y. -H. (2000). Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, *128*, 109—120.

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*, 1339—1366.

Ilvessalo, Y. (1950). On the correlation between crown diameter and the stem of trees. Publications of Forest Research Institute in Finland.*Communicationes Instituti Forestalis Fenniae*, *38*(2) 32 pp.

Jakobsons, A. (1970). Sambandet mellan trädkronans diameter och andra trädfaktorer, främst brösthöjdsdiametern: analyser grundade på riksskogstaxeringens provträdsmaterial. Summary: The correlation between the diameter of the tree crown and other tree factors — mainly the breast-height diameter: Analyses based on sample-trees from the National Forest Survey. Department of Forest Survey, Royal College of Forestry, Stockholm, Sweden. *Rapporter och uppsatser*, 14. 75 p. [in Swedish].

Kaitaniemi, P., & Lintunen, A. (2008). Precision of allometric scaling equations for trees can be improved by including the effect of ecological interactions. *Trees*, *22*, 579—584.

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.

Kangas, A. (in press). Value of forest information. *European Journal of Forest Research*. doi:10.1007/s10342-009-0281-7.

Korhonen, L., Peuhkurinen, J., Malinen, J., Maltamo, M., Suvanto, A., Packalén, P., & Kangas, J. (2008). The use of airborne laser scanning to estimate sawlog volumes. *Forestry*, *81*, 499—510.

Korpela, I. (2004). Individual tree measurements by means of digital aerial photogrammetry.*Silva Fennica Monographs*, *3* 93 pp.

Korpela, I. (2007). Incorporation of allometry in single-tree remote sensing with LiDAR and multiple images. In C. Heipke, K. Jacobsen, & M. Gerke (Eds.), *Proceedings of ISPRS Hannover workshop 2007 on High Resolution Earth Imaging for Geospatial Information. Hannover, Germany, May 29–June 1, 2007* IAPRS XXXVI Part I/W51, ISSN No. 1682-1777. Available at http://www.ipi.uni-hannover.de/fileadmin/institut/pdf/Korpela.pdf

Korpela, I., & Tokola, T. (2006). Potential of aerial image-based monoscopic and multiview single-tree forest inventory — A simulation approach. *Forest Science*, *52*, 136—147.

Korpela, I., Dahlin, B., Schäfer, H., Bruun, E., Haapaniemi, F., Honkasalo, J., Ilvesniemi, S., Kuutti, V., Linkosalmi, M., Mustonen, J., Salo, M., Suomi, O., & Virtanen, H. (2007). Single-tree forest inventory using lidar and aerial images for 3D treetop positioning, species recognition, height and crown width estimation. In P. Rönnholm, H. Hyyppä, & J. Hyyppä (Eds.), *Proceedings of ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007. September 12–14, 2007, Espoo, Finland* (pp. 227—233). IAPRS XXXVI Part 3 / W52. Available at http://www.commission3.isprs.org/laser07/final_papers/Korpela_2007.pdf

Korpela, I., Tuomola, T., & Välimäki, E. (2007). Mapping forest plots: An efficient method combining photogrammetry and field triangulation. *Silva Fennica*, *41*(3), 457—469.

Korpela, I., Koskinen, M., Vasander, H., Holopainen, M., & Minkkinen, K. (2009). Airborne small-footprint discrete-return LiDAR data in the assessment of boreal mire surface patterns, vegetation, and habitats. *Forest Ecology and Management*, *258*, 1549—1566.

Korpela, I., Tokola, T., Ørka, H. O., & Koskinen, M. (2009). Small-footprint discrete-return LIDAR in tree species recognition. In C. Heipke, K. Jacobsen, S. Müller, & U. Sörgel (Eds.), *Proceedings of ISPRS Hannover workshop 2009 on High Resolution Earth*

Imaging for Geospatial Information. Hannover, Germany, June 2–5, 2009 IAPRS XXXVIII-1-4-7/W5. Available at http://www.isprs.org/proceedings/XXXVIII-1-4-7_W5/paper/Korpela-137.pdf

Korpela, I., Ørka, H.O., Maltamo, M., Tokola, T. and Hyyppä, J. (2010). Tree species classification using airborne LIDAR — Effects by stand and tree parameters, downsizing the training set, intensity normalization and sensor type. Manuscript.

Laasasenaho, J. (1982). Taper curve and volume function for pine, spruce and birch. Publications of Forest Research Institute in Finland.*Communicationes Instituti Forestalis Fenniae*, *108* 74 pp.

Lämås, T., & Eriksson, L. O. (2003). Analysis and planning systems for multiresource, sustainable forestry: The Heureka research programme at SLU. *Canadian Journal of Forest Research*, *33*, 500−508.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest.*R News*, *2* (3), 18−22 Available at http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

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

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., & Kangas, J. (2006). Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*, *36*, 426−436.

Maltamo, M., Packalén, P., Peuhkurinen, J., Suvanto, A., Pesonen, A., & Hyyppä, J. (2007). Experiences and possibilities of ALS based forest inventory in Finland. In P. Rönnholm, H. Hyyppä, & J. Hyyppä (Eds.), *Proceedings of ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007. September 12–14, 2007, Espoo, Finland* (pp. 270−279). IAPRS XXXVI Part 3 / W52. Available at http://www.commission3.isprs.org/laser07/final_papers/Maltamo_2007_keynote.pdf

Maltamo, M., Bollandsås, O. M., Næsset, E., Gobakken, T., & Packalén, P. (2009). Different sampling strategies for field training plots in ALS inventory. *Proceedings of SilviLaser 2009 held in October 14–16, 2009*. Texas, U.S.A: College Station 10 pp.

Maltamo, M., Packalén, P., Suvanto, A., Korhonen, K. T., Mehtätalo, L., & Hyvönen, P. (2009). Combining ALS and NFI training data for forest inventory — A case study in Kuortane, Western Finland. *European Journal of Forest Research*, *128*, 305−317.

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.

Moeur, M., & Stage, A. R. (1995). Most Similar Neighbor: An improved sampling inference procedure for natural resources planning. *Forest Science*, *41*, 337−359.

Moffiet, T., Mengersen, K., Witte, C., King, R., & Denham, R. (2005). Airborne laser scanning: Exploratory data analysis indicates potential variables for classification of individual trees or forest stands according to species. *ISPRS Journal of Photogrammetry & Remote Sensing*, *59*, 289−309.

Muinonen, E., Maltamo, M., Hyppänen, H., & Vainikainen, V. (2001). Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. *Remote Sensing of Environment*, *78*, 223−228.

Niska, H., Packalén, P., Skön, J.-P., Tokola, T., Maltamo, M. and Kolehmainen, M. (in press). Predicting species-specific tree volumes using artificial neural networks and multi-objective feature selection. *IEEE Transactions in Geoscience and Remote Sensing* doi:10.1109/TGRS.2009.2029864.

Ørka, H. O., Næsset, E., & Bollandsås, O. M. (2009). Classifying species of individual trees by intensity and structure features derived from airborne laser scanning data. *Remote Sensing of Environment*, *113*, 1163−1174.

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*, 328−341.

Packalén, P., & Maltamo, M. (2008). Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Canadian Journal of Forest Research*, *38*, 1750−1760.

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

Peuhkurinen, J., Maltamo, M., & Malinen, J. (2008). Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: A distribution-based approach. *Silva Fennica*, *42*(4), 625−641.

Sironen, S. (2009). Estimating individual tree growth using non-parametric methods. *Dissertationes Forestales*, 94. 54 p. Available at http://www.metla.fi/dissertationes/df94.htm

St-Onge, B., Jumelet, J., Cobello, M., & Véga, C. (2004). Measuring individual tree height using a combination of stereophotogrammetry and lidar. *Canadian Journal of Forest Research*, *34*, 2122−2130.

Takahashi, T., Yamamoto, K., Senda, Y., & Tsuzuku, M. (2005). Predicting individual stem volumes of sugi (*Cryptomeria japonica* D. Don) plantations in mountainous areas using small-footprint airborne LiDAR. *Journal of Forest Research*, *10*, 305−312.

Talts, J. (1977). Mätning i storskaliga flygbilder för beståndsdatainsamling. Summary: Photogrammetric measurements for stand cruising. Department of Forest Mensuration and Management, Royal College of Forestry, Stockholm, Sweden. *Research notes NR 6 - 1977*. 102 p. [In Swedish].

Vapnik, V., & Kotz, S. (2006). *Estimation of dependences based on empirical data.*: Springer ISBN 0387308652. 510 pp.

Vauhkonen, J. (in press). Estimating crown base height for Scots pine by means of the 3-D geometry of airborne laser scanning data. *International Journal of Remote Sensing* doi:10.1080/01431160903380615.

Vauhkonen, J., Tokola, T., Maltamo, M., & Packalén, P. (2008). Effects of pulse density on predicting characteristics of individual trees of Scandinavian commercial species using alpha shape metrics based on airborne laser scanning data. *Canadian Journal of Remote Sensing*, *34*(Suppl. 2), S441−S459.

Vauhkonen, J., Tokola, T., Packalén, P., & Maltamo, M. (2009). Identification of Scandinavian commercial species of individual trees from airborne laser scanning data using alpha shape metrics. *Forest Science*, *55*, 37−47.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*, Fourth edition : Springer ISBN 0-387-95457-0. 495 pp.

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.