



# Comparing nearest neighbor configurations in the prediction of species-specific diameter distributions

Janne Rätty<sup>1</sup> · Petteri Packalen<sup>1</sup> · Matti Maltamo<sup>1</sup>

Received: 8 December 2017 / Accepted: 12 February 2018 / Published online: 6 March 2018  
© INRA and Springer-Verlag France SAS, part of Springer Nature 2018

## Abstract

• **Key message** We examine how the configurations in nearest neighbor imputation affect the performance of predicted species-specific diameter distributions. The simultaneous nearest neighbor imputation for all tree species and separate imputation by tree species are evaluated with total volume calibration as a prediction method for diameter distributions.  
• **Context** This study considers the predictions of species-specific diameter distributions in Finnish boreal forests by means of airborne laser scanning (ALS) data and aerial images.  
• **Aims** The aim was to investigate different configurations in non-parametric nearest neighbor (NN) imputation and to determine how changes in configurations affect prediction error rates for timber assortment volumes and the error indices of the diameter distributions.  
• **Methods** Non-parametric NN imputation was used as a modeling method and was applied in two different ways: (1) diameter distributions were predicted at the same time for all tree species by simultaneous NN imputation, and (2) diameter distributions were predicted for one tree species at a time by separate NN imputation. Calibration to a regression-based total volume prediction was applied in both cases.  
• **Results** The results indicated that significant changes in the volume prediction error rates for timber assortment and for error indices can be achieved by the selection of responses, calibration to total volume, and separate NN imputation by tree species.  
• **Conclusion** Overall, the selection of response variables in NN imputation and calibration to total volume improved the predicted diameter distribution error rates. The most successful prediction performance of diameter distribution was achieved by separate NN imputation by tree species.

**Keywords** NN imputation · Area-based approach · Airborne laser scanning · Diameter distribution

## 1 Introduction

### 1.1 Background

The application of airborne laser scanning (ALS) in forest management has been widely investigated in recent decades.

---

**Handling Editor:** Aaron R. Weiskittel

**Contribution of the co-authors** The corresponding author carried out the survey, analyzed the results, and prepared the manuscript. The other authors participated in the design of the study and also commented and refined the manuscript.

---

✉ Janne Rätty  
janne.ratty@uef.fi

<sup>1</sup> Faculty of Forestry, University of Eastern Finland, Yliopistokatu 7, P.O. Box 111, FIN-80101 Joensuu, Finland

As a result, ALS (in combination with other remote sensing techniques) has been utilized in forest inventories in many countries (see White et al. 2013; Maltamo and Packalen 2014; Næsset 2014). In Finland, species-specific stand attributes have traditionally formed the basis of traditional forest management and procurement planning. Therefore, stand-level assessments of forest resources have been widely used to measure forest attributes over the latter part of the twentieth century (Koivuniemi and Korhonen 2006). The use of diameter distributions to describe the size distribution of trees has been an essential part of forest inventories in stand-level assessments, as well as in ALS-based inventories. In modern forestry practices, diameter distributions are used, for example, to simulate the future development of forests, and hence are necessary instruments in planning and scheduling forest management decisions. Diameter distribution can also be used to describe tree layers, which is useful for conservation

purposes. Traditionally, diameter distributions in Finland have been predicted separately in a species-specific manner because of species-specific timber procurements (e.g., Kangas and Maltamo 2000).

## 1.2 Diameter distributions based on ALS

Currently, the area-based approach (ABA) is usually employed with low-density ALS data, with the ALS-derived statistics calculated at a plot or a grid level (Næsset 1997). During the era of ALS, studies related to diameter distributions, especially where tree species are considered, are still somewhat rare. Most global forest types are occupied by numerous tree species, and in forest management, diameter distributions have been predicted at the total or genus level. In Europe, ALS-based diameter distribution studies without tree species separation have been published, for example, from Norway (Gobakken and Næsset 2004), Sweden (Saad et al. 2015), Finland (Maltamo et al. 2006), and Germany (Breidenbach et al. 2008). ALS data have also been used in diameter distribution prediction in North America (Thomas et al. 2008; Lamb et al. 2017; Shang et al. 2017). Studies have shown that diameter distributions can be predicted for individual tree species by means of ALS and aerial images under Finnish conditions (Packalén and Maltamo 2008; Peuhkurinen et al. 2008), although, the error rates associated with the prediction of minor tree species and timber assortment volumes need to be improved if these variables are deemed to be of interest.

In regard to area-based ALS studies, diameter distributions have been predicted with both parametric and non-parametric methods. Parametric methods are usually based on probability distributions, such as the Weibull density function (Gobakken and Næsset 2004). The use of probability distributions to describe diameter distributions can be seen as contradictory—they are suitable for describing even-aged monocultures, although multimodal distributions may not be described at a satisfactory level (Maltamo et al. 2017). In addition, current ALS-based forest inventories require field data in which the diameter at breast height (DBH) has been measured for all trees. Therefore, the utilization of data from trees measured in the field, rather than the prediction of parameters based on theoretical distributions, is justified. Moreover, the local variability of diameter distributions in the inventory area is taken into account better compared to general parameter models.

The non-parametric nearest neighbor imputation (hereafter NN imputation) produces diameter distributions as a byproduct from the NN imputation of forest attributes. The diameter distribution is constructed from trees that occur in sample plots. NN imputation for species-specific predictions has been used, for example, by Packalén and Maltamo (2008) and Peuhkurinen et al. (2008). The results of the former study indicated that NN imputation for the prediction of diameter

distributions under boreal forest conditions was superior to the parametric method based on a Weibull distribution. Bollandsås et al. (2013) also assessed the performance of seemingly unrelated parametric regression and non-parametric NN imputation for the prediction of total-level diameter distributions and concluded that NN imputation performed better than parametric approaches.

One of the most common approaches employed in NN imputation has been the most similar neighbor (MSN) method (Moeur and Stage 1995), which is usually applied for  $k$  neighbors ( $k$ -MSN), with the  $k$  value commonly set to 5. (e.g., Packalén and Maltamo 2008; Maltamo et al. 2009). The potential of MSN for diameter distribution modeling was observed before the era of ALS (Maltamo et al. 2003). The benefit of multivariate NN imputation is that a user can simultaneously predict all of the attributes of interest in a single step. Packalén and Maltamo (2008) selected response variables for diameter distribution modeling that consisted of different species-specific sum and mean attributes in field plots. For the prediction of sum and mean attributes, the selection of the responses is a straightforward task as the attributes of interest can be actual responses. However, for diameter distributions, the attribute is not usually used as such, although some indicators related to the shape and structure of diameter distribution might be included as responses. Related to this, Maltamo et al. (2009) studied response configurations in the modeling of diameter distribution by  $k$ -MSN at the total tree stock level, and their results suggested that the selection of responses has an effect on the performance of diameter distributions.

## 1.3 Goodness of diameter distribution

Performance assessments of diameter distributions have been implemented by error indices (Gobakken and Næsset 2004; Packalén and Maltamo 2008; Peuhkurinen et al. 2008), error rates in predicted timber assortments (Packalén and Maltamo 2008; Peuhkurinen et al. 2008), and statistical tests (Strunk et al. 2017; Gorgoso et al. 2007; Poudel and Cao 2013). The root mean squared error (RMSE) and bias (BIAS) of timber assortment and tree species volumes are easy to interpret, and their use is justified by the fact that logwood volume is the main forest attribute evaluated economically. The use of timber assortment as a validation criterion requires that tree heights are predicted, and assortments are calculated using taper curve models (e.g., Siipilehto 1999). However, assessments related to the frequencies of diameter distribution are reasonable because the shape of the diameter distribution might not always be observed correctly from the RMSE values of predicted timber assortment volumes. Some error indices have been developed for diameter distributions; for example, Reynolds et al. (1988) proposed an index that

compares the frequencies between observed and predicted distributions.

#### 1.4 Study objectives

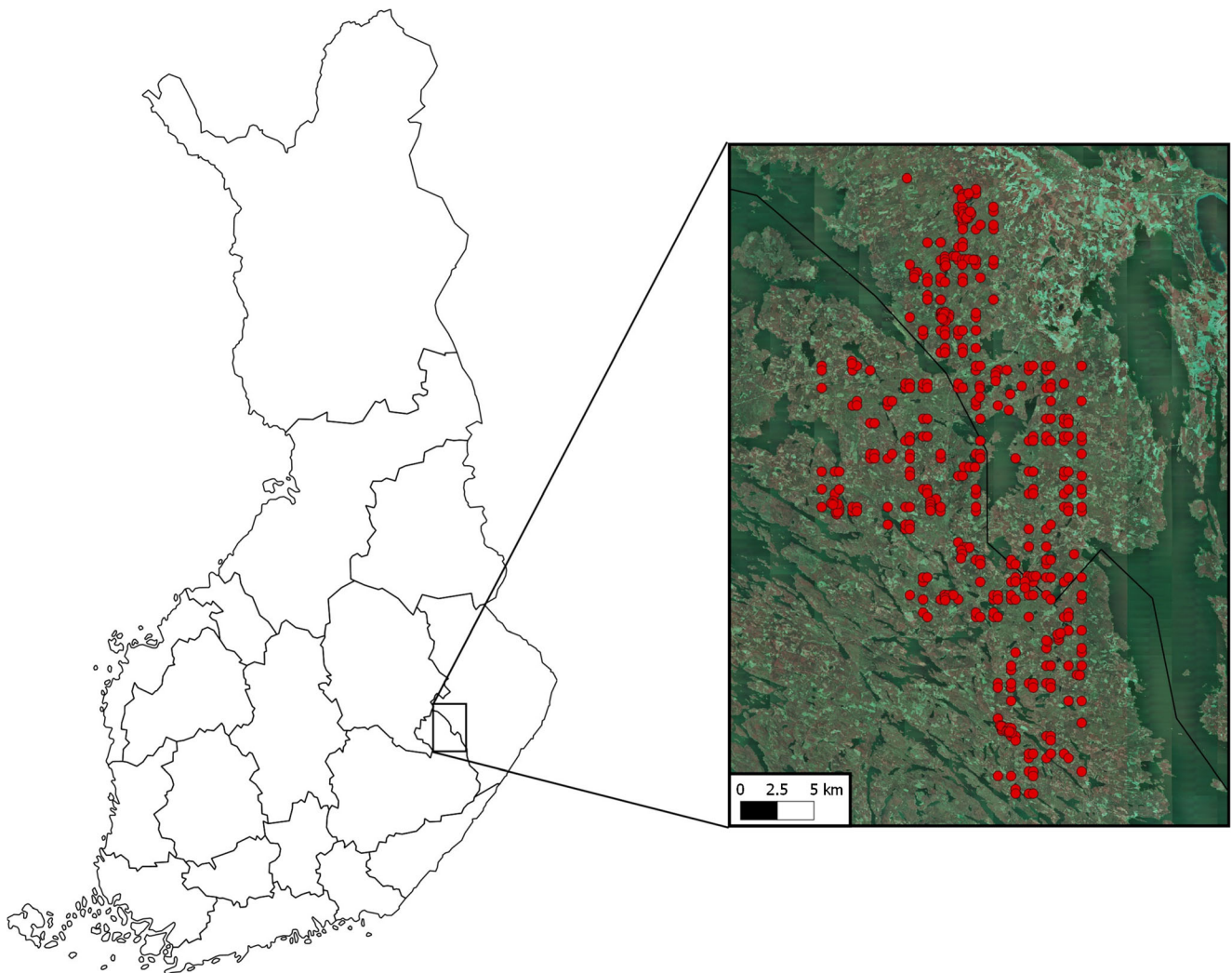
The primary aim of this study is to investigate species-specific diameter distributions predicted by NN imputation using various response configurations. Low-density leaf-off ALS and aerial image metrics are used as predictor variables. In remote sensing applications, the species-specific diameter distribution for the target observation has traditionally been fetched simultaneously for all tree species from the same  $k$  nearest neighbors. In this study, a secondary objective is to examine whether the species-specific diameter distributions could be better predicted when the attributes of pine, spruce, and deciduous tree species were used independently as responses in separate NN imputations. Moreover, a straightforward regression model-based calibration to total volume is employed for

diameter distribution frequencies to ensure that NN-imputed species-specific volumes sum up to the total volume prediction.

## 2 Material and methods

### 2.1 Study area

The study site is located around the municipality of Liperi (62°31' N 29°23' E) in the province of Eastern Finland (see Fig. 1). The study area extends partly over the province of Southern Savonia. The total area of the study site covers about 43,000 ha. The forest structure can be considered as a typical Finnish managed forest in which coniferous species, such as Scots pine (*Pinus sylvestris* [L.]) and Norway spruce (*Picea abies* [L.] Karst.) are usually the dominant tree species. In the field data, the main tree species is Norway spruce and



**Fig. 1** The location of the inventory area and the field plots

**Table 1** Means and standard deviations of plot attributes by tree species at the plot level

	Population	Mean	Sd
Volume (m <sup>3</sup> ha <sup>-1</sup> )	Scots pine	76.0	83.4
	Norway spruce	87.7	109.1
	Deciduous	22.8	36.1
	Total	186.4	100.4
Basal area (m <sup>2</sup> ha <sup>-1</sup> )	Scots pine	8.8	9.1
	Norway spruce	9.9	10.5
	Deciduous	2.9	4.2
	Total	21.6	8.2
Stem number (ha <sup>-1</sup> )	Scots pine	371.0	458.5
	Norway spruce	513.8	523.5
	Deciduous	239.3	392.3
	Total	1142.0	641.4
DGM (cm)	Scots pine	21.0	6.5
	Norway spruce	17.1	8.4
	Deciduous	14.8	7.5
	Total	–	–
HGM (m)	Scots pine	17.6	4.8
	Norway spruce	14.8	6.5
	Deciduous	15.3	5.5
	Total	–	–

deciduous species (e.g., silver birch (*Betula pendula* Roth.), downy birch (*Betula pubescens* Ehrh.), and aspen (*Populus tremula*)) are in the minority. According to the field data, the most common site fertility classes are grass-herb (19%; *Oxalis-type*), moist (58%; *Myrtillus-type*), and dry sites (19%; *Vaccinium-type*). The proportions of stand development classes are 27, 52, and 21% for young, middle-aged, and mature forests, respectively.

## 2.2 Field data

The fieldwork was carried out between June and September 2016. The majority of sample plots (73%) was distributed in the study area using systematic cluster sampling. The clusters include four plots, and the distance between the clusters is 1200 m. The remainder of the sample plots was distributed and measured by the Finnish Forest Centre. The sample plot network used in this study consists of 424 circular plots with

either a radius of 9 m (71%) or 12.62 m (29%) depending on the stem number inside the plot. All the sample plots were located entirely within the forest stand. Seedling or sapling plots and dead trees were excluded from the data. The main sample plot attributes are presented in Table 1.

Post-corrected global navigation satellite system (GNSS) measurements were used to accurately locate the center of each sample plot. For every plot, DBH, height, species, and tree class of trees with DBH ≥ 5 cm were measured. Using DBH and height measurements, the basal area and volume were calculated for all trees and were subsequently aggregated at the plot level by tree species and multiplied to the hectare level together with the number of stems. Diameter and height of the basal area median tree (DGM and HGM) were also calculated. Using plot level DBH measurements, diameter percentiles were calculated from 10 to 90% (D10, D20, ..., D90) by tree species. Tree level stem volumes were computed using the species-specific models of Laasasenaho (1982) with DBH and height as predictor variables. Tree species were classified into three classes: pine, spruce, and deciduous trees. Deciduous trees were aggregated together due to (a) the minority of deciduous species other than birches and (b) the knowledge that deciduous species are practically impossible to separate with the remote sensing systems currently utilized under Finnish conditions. Species-specific timber assortment volumes were calculated by means of taper curves (Laasasenaho 1982), and the taper curve for birch was used for all deciduous species. The bucking parameters related to logwood and pulpwood calculations are presented in Table 2.

## 2.3 ALS data and aerial photographs

Leaf-off ALS was carried out between April 30, 2016, and May 3, 2016, using a fixed-wing Piper PA-31-350 Chieftain airplane with a Leica ALS60 laser scanning device. The data were collected from an altitude of 2400 m above ground level using a half-angle of 20°. The ALS data were collected mainly for terrain modeling purposes by the National Land Survey of Finland. As a result of the data acquisition setup, a nominal sampling density of 0.88 measurements per square meter was achieved. The side overlap between the flight lines was 20%.

The Leica ALS60 device can capture four echoes per emitted pulse, and the echoes were classified into *only*,

**Table 2** Bucking parameters used when computing pulpwood and logwood volumes

	Logwood minimum diameter (cm)	Pulpwood minimum diameter (cm)	Logwood minimum length (m)	Logwood maximum length (m)
Scots pine	15.0	7.0	3.7	6.1
Norway spruce	16.0	7.0	3.7	6.1
Deciduous species	18.0	7.0	3.7	6.1

*intermediate*, *first-of-many*, and *last-of-many* categories. Here, we used three of the echo categories: *first*, *last*, and *intermediate*. The *only* category was joined to the *first-of-many* and the *last-of-many* categories to constitute the *first* and *last* categories. The ALS points were classified into ground hits and other hits following the method of Axelsson (2000). The digital terrain model (DTM) was interpolated with Delaunay triangulation by means of ground classified ALS echoes. Height normalization of the echoes was implemented by subtracting the DTM from the height values of the ALS echoes. The intensity values of the ALS echoes were calibrated for the range (Korpela et al. 2010).

Aerial photographs were captured at a flight altitude of 4100 m above ground level on the 23–24 May 2016 by the National Land Survey of Finland. The aerial photographs were taken with a Z/I Imaging Intergraph (01-0128) camera with a focal length of 30 mm. The camera has  $3456 \times 1920$  pixels, and the ground sampling distance (GSD) of the aerial images was about 160 cm. The abovementioned parameters are for multispectral bands; we did not use panchromatic band or pan-sharpening. This camera model records four spectral bands (red, green, blue, and near infrared).

## 2.4 Predictor variables

The ALS echo categories; *first*, *last*, and *intermediate* were included in the analysis, and most predictors were calculated separately for all categories. The height cutoff was set at 1.3 m to separate ground hits from the vegetation metrics. Mean, standard deviation, max, min, kurtosis, skewness, percentiles, and fixed density metrics were computed at the plot level. The canopy height percentiles (h) were computed for 5, 10, ..., 95%. The density values (d) that describe the proportion of echoes under the given height value were calculated for heights of 0.5, 2, 5, 10, 15, and 20 m. In the case of density metrics, a height cutoff threshold was not used and those metrics were always computed according to all echoes of used echo categories. The proportions of echoes by echo categories were also computed. Moreover, percentiles (i5, i10, i15, ..., i95) together with mean, standard deviation, max, min, kurtosis, and skewness were computed for the intensity values of the ALS echoes.

Aerial image metrics were computed as explained in Packalén et al. (2009). First ALS echoes belonging to the *first-of-many* and *only* echo categories were projected to unrectified aerial images using external and internal orientations. Every echo hits in multiple images because of the overlap in aerial imaging. Therefore, an average pixel value was retrieved for every echo by bands. Finally, the mean, standard deviation, minimum, and maximum by bands were computed from retrieved pixel values by plots.

## 2.5 Workflow

The overall workflow applied in the study is illustrated in Fig. 2, which sums up the stages from data elaboration to performance assessments. The routines for NN imputation and the selection of predictor variables were implemented following the algorithms applied in Packalén et al. (2012). The routines for construction and assessment of diameter distributions were carried out in R-environment (R Core Team 2017).

## 2.6 NN imputation

The NN imputation was used to search the similar neighbors for target observations. The tree list (i.e., diameter distribution) can then be compiled from the tree-level data of the nearest plots. The NN imputation was implemented by a k-MSN. The MSN method applies canonical correlation analysis to define the distance metrics. The distance metric determines which k reference observations (k nearest neighbors) are most similar to the object of prediction. In this study, the k value was fixed at 5 for all NN imputations. For simplicity, the MSN distance metric was solved only once in every NN run and the target observation was always excluded from the group of k neighbors (TRAINCV approach in Packalén et al. 2012).

The squared distance for nearest neighbors was calculated as follows:

$$D_{uj}^2 = (X_u - X_j) \Gamma \Lambda^2 \Gamma' (X_u - X_j)' \quad (1)$$

where  $D_{uj}^2$  is the squared distance between target  $u$  and neighbor  $j$ ,  $X_u$  is a vector of predictor variables from the target plot,  $X_j$  is a vector of predictor variables from the reference plot,  $\Gamma$  is the matrix of canonical coefficients of the predictors, and  $\Lambda^2$  is the diagonal matrix of squared canonical correlations.

The nearest neighbors were weighted by taking an inverse of the squared distance as follows:

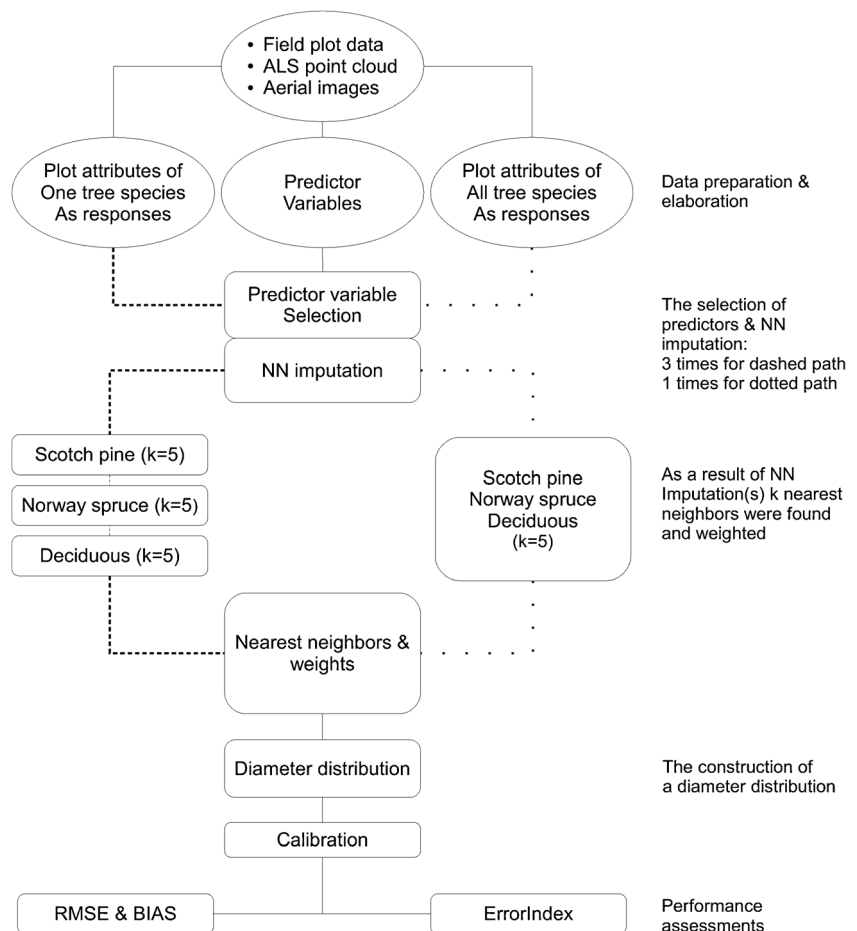
$$W_{uj} = \frac{1}{D_{uj}^2} \quad (2)$$

where  $W_{uj}$  is the weight value between the target plot  $u$  and reference plot  $j$ ,  $k$  is the number of nearest neighbors, and  $D_{uj}^2$  is the squared MSN distance.

## 2.7 The selection of response variables

The starting point in the comparison is the reference set of responses (hereafter SET<sub>1</sub>) consisting of 15 responses: species-specific volume (V), basal area (G), stem number

**Fig. 2** The schematic illustration of the workflow used in the study. The dashed path demonstrates the separate non-parametric nearest neighbor (NN) imputation by tree species and the dotted path refers to simultaneous NN imputation



(N), basal area median height (HGM), and diameter (DGM) (as presented in Packalén and Maltamo 2008). Subsequently, numerous response configurations were compiled. Due to the high number of response configurations, the results of all configurations cannot be presented in a succinct manner. Consequently, five response configurations are presented here and consisted of two response configurations for simultaneous NN imputation and three response configurations for separate NN imputation by tree species. The selection was based on the RMSEs of the predicted timber assortment volumes, and the error indices of diameter distributions. The selection of predictor variables was run individually for every configuration of response variable. The presumption was that diameter distribution related responses (e.g., DGM, G, and N) are important for the description of diameter distribution. Therefore, those attributes were included in most of the configurations of responses compared. In addition to sum and mean attributes, other attributes, such as diameter percentiles (D10, D20,..., D90), diameter distribution frequencies with a 2-cm class width (F6, F8, ..., F32), and Weibull-parameters ( $c$  = shape,  $b$  = scale) were compared. All response configurations taken into account are presented in Section 3.1 (see also Fig. 3).

## 2.8 The selection of predictor variables

When diameter distributions are predicted in a species-specific manner by NN imputation, the crucial part is tree species discrimination. Thus, careful selection of predictor variables is critical. Since the selection of predictor variables may be laborious owing to the large amount of predictor candidates, heuristic optimization methods have been proposed to lighten the computational workload of the process (Tuominen and Haapanen 2013; Packalén et al. 2012). For tree species discrimination, aerial images (e.g., Packalén and Maltamo 2007) have often been used together with ALS data. To some extent, leaf-off ALS data can also discriminate between coniferous and deciduous tree species in boreal conditions (Villikka et al. 2012).

The variable selection for NN imputation was implemented according to the algorithm proposed by Packalén et al. (2012). The method is based on heuristic optimization by Simulated Annealing (Kirkpatrick et al. 1983), which aims to minimize the mean RMSE over all responses by solving the NN model repeatedly until a solution is found. The minimization is based on the cost function and the temperature parameter. The latter has a key role in avoiding local optima by determining the

**Fig. 3** The results of the response selection process. (1) In the simultaneous non-parametric nearest neighbor (NN) imputation, all responses are included for every tree species at a time (e.g., V for pine, spruce, and deciduous species). (2) In the separate NN imputation by tree species, the responses of only one species are included as responses at a time. For abbreviations used, please refer to section 2.7

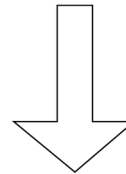
### Simultaneous NN imputation<sup>1</sup>

SET<sub>1</sub>: V + G + N + DGM + HGM  
 SET<sub>2</sub>: G + N + DGM  
 SET<sub>3</sub>: G + N + DGM + HGM  
 SET<sub>4</sub>: V + N + DGM + HGM  
 SET<sub>5</sub>: G + N + DGM + D30 + D80  
 SET<sub>6</sub>: G + N + DGM + c + b  
 SET<sub>7</sub>: V + Vlog + Vpulp  
 SET<sub>8</sub>: Vlog + N/G

### Separate NN imputation by tree species<sup>2</sup>

SET<sub>sep1</sub>: V + G + N + DGM + HGM  
 SET<sub>sep2</sub>: G + N + DGM  
 SET<sub>sep3</sub>: D10 + D20 + ... + D90 + N  
 SET<sub>sep4</sub>: D10 + D20 + ... + D90  
 SET<sub>sep5</sub>: V + Vlog + Vpulp  
 SET<sub>sep6</sub>: F6 + F8 + ... + F34

Selection according to the error rates of the predicted volumes of timber assortments and ErrorIndex values



SET<sub>1</sub> = V + G + N + DGM + HGM  
 SET<sub>2</sub> = G + N + DGM

SET<sub>sep1</sub> = V + G + N + DGM + HGM  
 SET<sub>sep2</sub> = G + N + DGM  
 SET<sub>sep3</sub> = D10 + D20 + ... + D90 + N

probability of also accepting moves to poorer solutions. Simulated annealing is a stochastic method, and therefore, the optimization routine was iterated 1000 times with a randomized initial solution for every variable selection case. The initial temperature value was set to 1. The selected set of predictor variables produced the smallest cost according to the mean RMSE of predictions for all responses during the iterative process. In this study, the procedure to select predictor variables was carried out separately for every response configuration. The number of predictor variables to be selected was fixed at 11.

## 2.9 Constructing diameter distributions

As nearest neighbors were found by means of NN imputation, the diameter distributions of neighbors can be used in the construction of a tree list for a target. Since the DBH and the height of all trees are measured; volumes (total, pulpwood, and logwood) and basal area can be calculated. Ultimately, every tree gets a weighting value that corresponds to the given weight value in NN imputation for the plot in question (Eq. 2). In reality, for each imputed tree, the weight describes the frequency that the tree represents in the target plot. The construction of the tree list for the target plot  $u$  from the nearest neighbors ( $k=5$ ) can be described as follows:

$$F_u = \{W_{u1}S_{u1}, W_{u2}S_{u2}, \dots, W_{u5}S_{u5}\} \quad (3)$$

where  $F_u$  is the tree list on a target plot  $u$ ,  $W_{uj}$  is the plot level weight as presented in Eq. 2, and  $S_{uj}$  is a list of trees in the neighbor plot  $j$  (here  $j$  refers to neighbor number 1, 2, 3, 4, or 5).

Finally, the trees in a tree list were scaled to the hectare level by multiplying frequencies by a scaling factor that corresponds to the area of the reference plot. In cases where only one tree species is seen as a response, the same principle can be followed as presented in Eq. 3, although the procedure must be applied separately for each tree species (Fig. 2). Then the final, separately NN imputed by tree species, diameter distribution for a target plot consists of trees of 15 nearest neighbors.

## 2.10 Calibration to total volume

The NN imputation of several forest attributes at a time may not optimally predict the total volume of a plot because the response is different while solving the distance metric. However, the total volume can be predicted with a lower error rate individually, for example, by means of linear regression analysis. Diameter distribution frequencies can be fixed according to multiple predicted forest attributes by calibration estimation (see Kangas and Maltamo 2000; Maltamo et al. 2007). Here, the calibration to total volume applies the same principle but only one attribute is used in the calibration. After the calibration, the total volume of the diameter distribution sums up to the regression-based prediction of total volume. Simultaneously, the predicted species-specific distributions might also correspond better with the observed distributions.

These phenomena motivated us to fit a linear model of three predictor variables for total volume. The predictor selection for the model was implemented by searching for the best performing combination of three predictor variables with

respect to RMSE. A square root transformation for the response variable was considered optimal to produce a constant variance for residuals. The bias caused by the transformation was corrected according to Lappi (1993).

The calibration factor calculated for a sample plot is presented in Eq. 4. The calibration factor was applied for a predicted diameter distribution by multiplying the tree frequencies by the calibration factor.

$$CFactor_u = \frac{\hat{V}_{tot.reg_u}}{(\hat{V}_{pine.nn_u} + \hat{V}_{spruce.nn_u} + \hat{V}_{decid.nn_u})} \quad (4)$$

where  $CFactor_u$  is a calibration factor for a plot  $u$ ,  $\hat{V}_{tot.reg}$  refers to the volume prediction by a regression model for plot  $u$ , and  $\hat{V}_{species.nn}$  in the denominator refers to the NN imputation-based volume predictions for plot  $u$ .

### 2.11 Performance assessments

Root mean squared error (RMSE) and mean difference (BIAS) figures were applied for the predictions of timber assortments and for tree species volumes. The RMSEs and BIASs were expressed in a relative way in which the absolute value was divided by the mean value of the corresponding observed attribute. The absolute RMSE and BIAS were calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (5)$$

$$BIAS = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{n} \quad (6)$$

An error index (hereafter ErrorIndex) was employed as a performance measure (Eq. 7). The same index was applied, for example, in the study by Gobakken and Næsset (2004). Originally, the ErrorIndex has been presented by Reynolds et al. (1988). The ErrorIndex is presented as mean values of all plot-level figures and is calculated for the total diameter distribution as well as for a logwood part. If the observed stem number was zero, the plot was excluded from the ErrorIndex calculations.

$$ErrorIndex = \sum_{i=1}^l 100 \left| \frac{f_i - \hat{f}_i}{N} \right| \quad (7)$$

where  $f_i$  and  $\hat{f}_i$  refer to the classes of diameter distribution,  $l$  is the number of classes in the diameter distribution, and  $N$  represents the number of observations in the observed distribution.

**Data availability** The ALS data can be downloaded from the website of the National Land Survey of Finland. Other data sources are not deposited in publicly available repositories.

## 3 Results

### 3.1 The selection of response configurations

For simultaneous NN imputation, a total of eight response configurations were originally compared (see the leftmost column in Fig. 3) and the results from two response configurations are presented here; SET<sub>1</sub> (as it is regarded as a common practice) and SET<sub>2</sub> as it performed most successfully. SET<sub>2</sub>, SET<sub>3</sub>, and SET<sub>4</sub> were modified configurations from SET<sub>1</sub>. A set of responses where species-specific logwood and the relationships between stem number and basal area (SET<sub>8</sub>) were simultaneously predicted was also tested and appeared to predict logwood volumes quite well, although the remainder of the diameter distribution was not predicted successfully. No significant benefit was found when different species-specific diameter percentiles (SET<sub>5</sub>) or Weibull parameters (SET<sub>6</sub>) were added to SET<sub>2</sub>. The response set comprising volume attributes (SET<sub>7</sub>) was not adequate for the modeling of diameter distributions in simultaneous NN imputation. Therefore, the aforementioned response sets were rejected and their results are not presented here. Diameter frequencies were also used as response variables, although the use of distribution frequencies as responses for simultaneous NN imputation of species-specific diameter distributions posed problems since the number of responses rose too high when a bin width of 1 or 2 cm was used.

For the separate NN imputation by tree species, a total of six response configurations were compared (see the rightmost column in Fig. 3) and the results from three response configurations are presented here; SET<sub>sep1</sub> and SET<sub>sep2</sub> were selected because of their correspondence with the simultaneous NN imputation sets (SET<sub>1</sub> and SET<sub>2</sub>), and in SET<sub>sep3</sub>, species-specific stem number was added to the response set of SET<sub>sep4</sub>. The inclusion of stem number improved most of the prediction error rates for logwood and pulpwood volumes compared to the response configuration that consisted solely of percentiles. The separate NN imputations by tree species where frequencies were a response set (i.e., SET<sub>sep6</sub>) did not perform as well as the other response sets tested. In the separate NN imputation, the set that consisted of volume attributes (SET<sub>sep5</sub>) performed quite well, and the RMSEs for coniferous logwood were at the same level as SET<sub>2</sub>. The response sets SET<sub>sep1</sub>, SET<sub>sep2</sub>, and SET<sub>sep3</sub> performed most successfully and provided a valuable comparison with the simultaneous NN imputation.



**Table 3** The relative root mean squared error (RMSE) and bias (BIAS) (in parenthesis) of logwood volume predictions at the plot level. For SET abbreviations, please refer to Fig. 3. For calibration see Eq. 4

	Uncalibrated				Calibrated			
	Vlog <sub>total</sub>	Vlog <sub>pine</sub>	Vlog <sub>spruce</sub>	Vlog <sub>decid</sub>	Vlog <sub>total</sub>	Vlog <sub>pine</sub>	Vlog <sub>spruce</sub>	Vlog <sub>decid</sub>
SET <sub>1</sub>	36.4 (−1.3)	67.1 (−1.5)	69.7 (0.8)	211.0 (−21.1)	28.8 (0.0)	61.6 (−0.5)	64.3 (2.6)	204.1 (−22.9)
SET <sub>2</sub>	35.0 (−0.5)	62.9 (−2.6)	62.4 (2.8)	177.6 (−15.8)	29.1 (0.6)	56.3 (−0.4)	59.3 (3.4)	171.6 (−18.2)
SET <sub>sep1</sub>	35.5 (1.0)	66.8 (1.4)	57.2 (1.4)	155.9 (−5.3)	26.2 (0.8)	55.9 (0.9)	52.8 (1.6)	155.8 (−6.9)
SET <sub>sep2</sub>	40.4 (1.7)	79.6 (1.1)	67.0 (2.8)	159.1 (−2.9)	27.6 (0.6)	65.5 (0.3)	58.9 (1.4)	160.9 (−6.0)
SET <sub>sep3</sub>	40.7 (−2.4)	73.2 (−3.3)	68.9 (−0.7)	221.7 (−12.3)	27.0 (−0.2)	61.4 (−0.3)	58.6 (0.5)	208.7 (−10.3)

### 3.2 The error rates of timber assortments

The relative RMSEs and BIASs are presented in Tables 3, 4, and 5 for logwood, pulpwood, and tree species volumes, respectively. When the tree species volumes and timber assortment volumes were calculated from uncalibrated diameter distributions, the results showed that the error rates predicted by SET<sub>1</sub> could be improved. For example, percentage improvements of 6.3, 10.5, and 15.8% for pine, spruce, and deciduous, respectively, were achieved for logwood error rates when the response variables of SET<sub>1</sub> were replaced by those of SET<sub>2</sub>. Moreover, the results indicated that separate NN imputation by tree species performed better compared to simultaneous NN imputation. Since the lowest error rates were mostly achieved by means of separate NN imputation by tree species, the findings imply that separate NN imputation by tree species might be a better alternative for the prediction of diameter distribution.

The most challenging part of the prediction was for deciduous species, which are minor components in the study area, and it was very difficult to find similar neighbors for a target plot. In addition to high-error rates, the predictions for deciduous species were more biased than predictions for the other species. The best performing configuration for deciduous volume, pulpwood, and logwood was acquired with separate NN imputation by tree species rather than simultaneous NN imputation.

**Table 4** The relative root mean squared error (RMSE) and bias (BIAS) (in parenthesis) of pulpwood volume predictions at the plot level. For SET abbreviations, please refer to Fig. 3. For calibration see Eq. 4

	Uncalibrated				Calibrated			
	Vpulp <sub>total</sub>	Vpulp <sub>pine</sub>	Vpulp <sub>spruce</sub>	Vpulp <sub>decid</sub>	Vpulp <sub>total</sub>	Vpulp <sub>pine</sub>	Vpulp <sub>spruce</sub>	Vpulp <sub>decid</sub>
SET <sub>1</sub>	38.2 (2.2)	60.9 (2.1)	64.9 (4.2)	96.7 (−1.2)	37.9 (0.5)	61.4 (1.2)	64.7 (4.2)	92.3 (−8.2)
SET <sub>2</sub>	36.0 (1.3)	62.3 (0.5)	62.4 (3.1)	96.1 (−0.6)	35.8 (−0.3)	63.0 (0.4)	61.2 (2.2)	93.4 (−6.3)
SET <sub>sep1</sub>	33.9 (−0.2)	56.3 (−1.0)	54.4 (−0.2)	78.3 (1.1)	34.3 (−1.2)	52.7 (−2.6)	58.4 (1.6)	87.9 (−3.3)
SET <sub>sep2</sub>	36.7 (−0.8)	63.0 (−0.8)	50.7 (−0.7)	88.1 (−1.1)	33.3 (−0.9)	56.4 (−0.8)	50.9 (0.7)	84.3 (−4.1)
SET <sub>sep3</sub>	41.5 (−0.3)	63.9 (2.0)	65.6 (−1.4)	109.4 (−2.5)	34.7 (0.1)	54.4 (0.5)	61.8 (0.8)	97.7 (−1.9)

### 3.3 Examination of ErrorIndex values

The ErrorIndex values computed with SET<sub>2</sub> were slightly better than those computed with SET<sub>1</sub> (Fig. 4). In addition, the ErrorIndex values indicated that SET<sub>sep3</sub> may be a better alternative than SET<sub>2</sub> if the ErrorIndex values for deciduous species were ignored. Overall, the results did not indicate a strong superiority of any of the response sets tested.

The ErrorIndex values for the diameter distributions predicted for the logwood part of the distribution by the response configurations of SET<sub>2</sub> and SET<sub>sep1</sub> are presented in Table 6. The diameter threshold of 18 cm was used to reduce the diameter distribution for the logwood part of the calculations. The bin width was fixed to 2 cm. For the logwood part, the majority of the ErrorIndex values supported the indication of the better prediction performance by means of SET<sub>sep1</sub> compared to SET<sub>2</sub> (Table 6).

### 3.4 Calibration to total volume

The regression model used for calibration to the total volume had an RMSE of 17.2% (the model formula:  $\sqrt{V_{tot}} = \sqrt{f_{h60}} + \sqrt{f_{d5}} + \text{inv}(f_{d15})$ ; for predictor abbreviations refer to section 2.4.). In general, calibration improved the error rates for the predictions of timber assortments and species-specific volumes (Tables 3, 4, and 5).

**Table 5** The relative root mean squared error (RMSE) and bias (BIAS) (in parenthesis) of volume predictions for tree species at the plot level. For SET abbreviations, please refer to Fig. 3. For calibration see Eq. 4

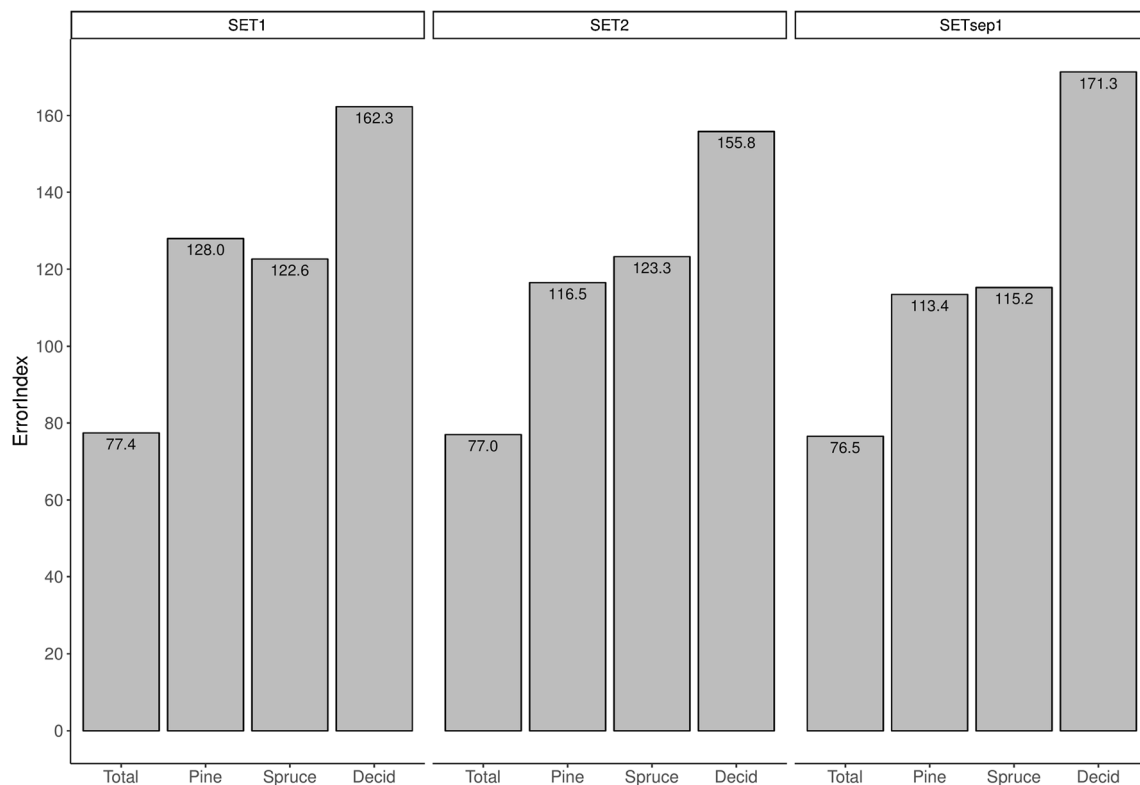
	Uncalibrated				Calibrated			
	V <sub>total</sub>	V <sub>pine</sub>	V <sub>spruce</sub>	V <sub>decid</sub>	V <sub>total</sub>	V <sub>pine</sub>	V <sub>spruce</sub>	V <sub>decid</sub>
SET <sub>1</sub>	23.1 (0.1)	50.3 (0.0)	50.1 (1.8)	76.3 (−5.9)	17.2 (0.0)	46.8 (0.2)	45.0 (2.9)	72.6 (−11.6)
SET <sub>2</sub>	21.4 (0.0)	46.9 (−4.7)	44.1 (2.7)	70.7 (−5.0)	17.2 (0.0)	43.3 (−0.2)	41.2 (2.7)	68.0 (−9.8)
SET <sub>sep1</sub>	24.3 (0.5)	50.2 (0.5)	40.6 (0.9)	61.3 (−0.7)	17.2 (0.0)	40.8 (−0.5)	38.6 (1.6)	67.8 (−4.3)
SET <sub>sep2</sub>	28.3 (0.7)	58.1 (0.4)	46.8 (1.4)	67.6 (−1.1)	17.2 (0.0)	45.2 (−0.1)	40.0 (1.1)	66.5 (−4.1)
SET <sub>sep3</sub>	30.5 (−1.5)	56.6 (−1.3)	52.0 (−0.9)	97.2 (−4.4)	17.2 (0.0)	44.1 (0.2)	43.6 (0.8)	82.4 (−3.6)

The effect of calibration was not unambiguous, as the calibration did not always improve the predictions of diameter distributions according to the performance assessments. However, the degradations were always minor. For example, uncalibrated diameter distributions imputed with the SET<sub>sep1</sub> responses produced slightly better (maximum 2 units) ErrorIndex values for total, spruce and deciduous species than those for calibrated (Fig. 4). Instead, the calibration works effectively for the logwood part with the response set of SET<sub>sep1</sub> (Table 6), and thus the inconsistency in the calibration process in regard to the whole distribution may be affected by changes in tree frequencies with small diameters. The effect can be noticed, for example, in the diameter distribution (Fig. 6) where small diameter classes are prone to higher

changes, since the distribution is positively skewed and high frequencies usually occur in small diameter classes. In Fig. 6, the calibration has worked somewhat effectively as the bar graphs present the diameter distribution of dominant tree species in conditions where the species-specific volumes are quite easy to predict.

## 4 Discussion

The use of diameter distributions in Finnish forest management systems is well established since most growth simulators use tree-level data, and timber assortments are calculated based on diameter distributions. In this study, we examined

**Fig. 4** ErrorIndex values for total volume calibrated diameter distributions by tree species. Two centimeter diameter classes were used. For SET abbreviations, please refer to Fig. 3

**Table 6** ErrorIndex values for uncalibrated and calibrated diameter distributions by tree species. 2 cm classes and diameter classes with midpoints  $\geq 18$  cm were used. For SET abbreviations, please refer to Fig. 3

		Uncalibrated ErrorIndex	Calibrated ErrorIndex
SET <sub>2</sub>	Total	90.18	89.46
	Pine	107.14	107.34
	Spruce	114.19	114.62
	Decid	131.59	129.21
SET <sub>sep1</sub>	Total	88.29	87.42
	Pine	105.39	102.87
	Spruce	113.06	111.61
	Decid	138.26	137.85

the response configurations in NN imputation, the application of the NN imputation method for the prediction of species-specific diameter distributions, and the application of straight-forward calibration to predicted total volume for predicted diameter distributions. The results highlighted the relationship between the response configurations and the performance of NN imputation. Our results showed that the prediction performance of NN imputation can be improved by response selection and by total volume calibration. Moreover, separate NN imputation by tree species produced the lowest error rates in the prediction of species-specific timber assortment.

Comparing our findings to previous studies is not straightforward since the use of remote sensing has changed in the last 10 years. Although numerous ALS-related studies have been published, diameter distributions as predicted by NN imputation and studies that specifically consider species-specific distributions are rare. The findings of this study are in line with the findings of previous studies (see Packalén and Maltamo 2007; Packalén and Maltamo 2008; Peuhkurinen et al. 2008). However, our results are at the plot level, which has a considerable effect on error rates when compared to stand level studies in general (e.g., Næsset 2004). The comparison of different response configurations for the NN imputation has been studied previously without considering tree species. Maltamo et al. (2009) tested numerous response configurations and found that distribution-related attributes (e.g., volume, basal area, stem number, diameter percentiles, and basal area weighted mean diameter) were appropriate for the prediction of diameter distribution. The results of our study confirm that the same also holds true at the species-specific level.

Increasing the performance of the diameter distribution predictions by modifying the initial response variable set of 15 species-specific attributes in NN imputation is to be expected. Although height variables tend to correlate well with ALS predictors, the correlation between tree diameter and height may not be as clear. Tree diameter is affected by soil fertility, tree age, geographical location, silvicultural operations etc.,

and these variables pose a challenge for remote sensing-based diameter distribution modeling. Taking this into account, the exclusion of the height-related responses can be justified in diameter distribution modeling. We evaluated various response configurations although the configurations comprising of sum and mean (G, DGM, N, etc.) attributes were observed to be successful in general. The response sets that included only volumes (species-specific attributes from performance assessments: V, Vlog, and Vpulp) could be used as responses only if those attributes are of interest. For diameter distribution modeling purposes, we observed that the response sets that comprised solely of volume attributes do not describe diameter distributions as well as the other response configurations studied. However, the use of diameter percentiles as responses with stem number did work in separate NN imputation by tree species with total volume calibration for coniferous species. For NN methods, the percentiles were first presented as response variables in ground truth data analyses by Malinen et al. (2001). Maltamo et al. (2009) included diameter percentiles in response configurations in ALS-based diameter distribution imputation but no significant benefit was achieved. Although the responses of SET<sub>sep3</sub> were not particularly successful in this study, we nevertheless deemed the observation interesting and valuable to report.

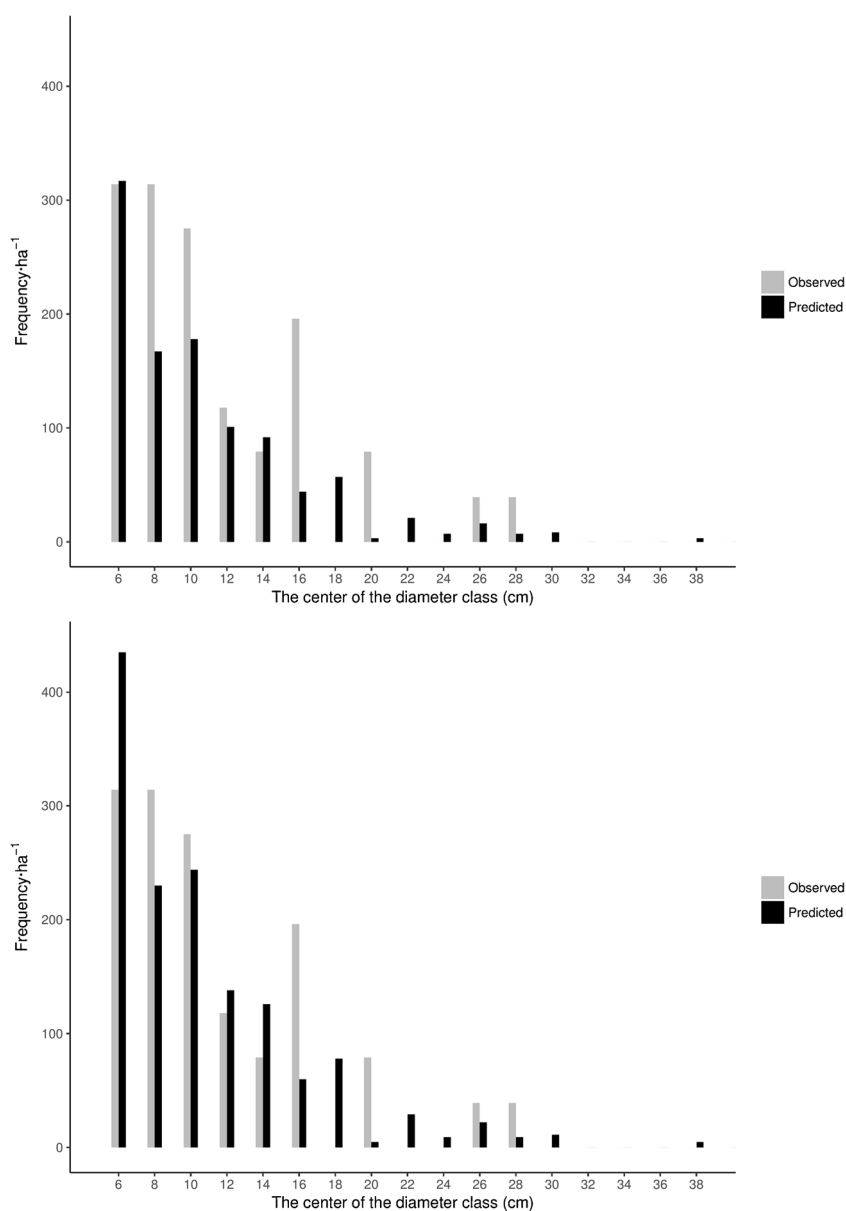
The application of the total volume calibration may not guarantee improvements in error rates or in the ErrorIndex values, as also reported by Maltamo et al. (2007). Overall, the advantage of the calibration (according to the error rates of predicted timber assortments) is evident from the point of view of operational forest inventories. Since the calibration is carried out by means of total volume, it may increase the error rates of the other variables, i.e., it causes trade-offs among the predicted variables. Our results showed that calibration usually negatively affects the bias values and the error rates of predicted timber assortment volumes, especially for deciduous species (Tables 3, 4, and 5). In contrast, the calibration always positively affected the error rates of predicted coniferous logwood volumes (Table 3), a significant variable when the commercial value of a forest is evaluated.

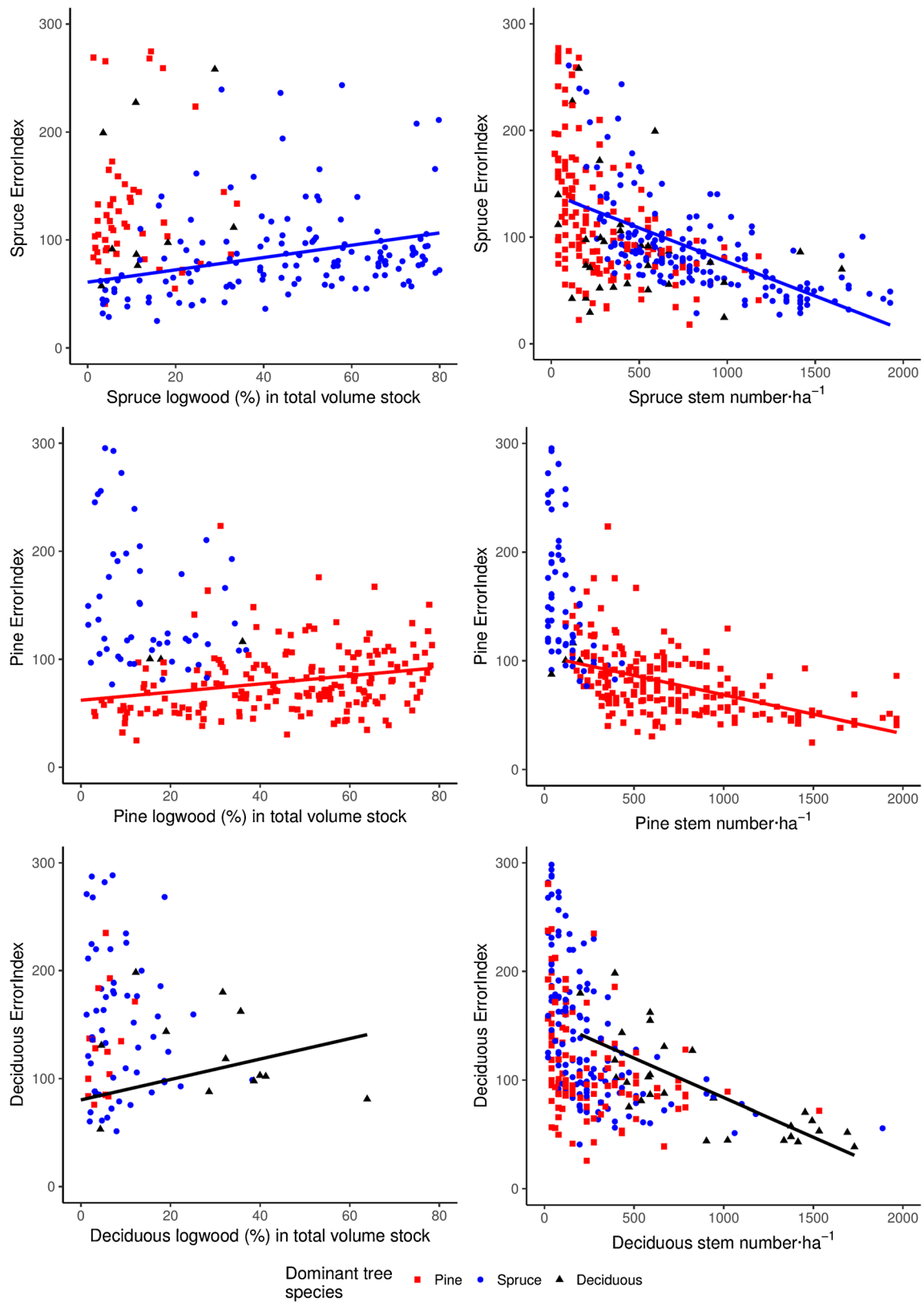
The findings of this study suggest that it might be reasonable to construct diameter distributions by separate NN imputation by tree species. One disadvantage of predicting diameter distributions by tree species is that selected neighbors may include tree species that do not occur in the target plot or vice versa. Principally, this will negatively affect the error rates of the predictions for minor species. Although the prediction error rates for deciduous species remain high, the error rates shown in Tables 3 and 4 indicate that separate NN imputation by tree species (SET<sub>sep1</sub> and SET<sub>sep2</sub>) might be better for species-specific prediction of diameter distributions compared to simultaneous NN imputation. Nevertheless, the ErrorIndex values for the deciduous species (Fig. 4 or Table 6) contradict that conclusion and provide evidence that the RMSEs of

predicted timber assortments might not always describe the shape of the diameter distribution as precisely as ErrorIndex. Despite the improvements in error rates for predicted timber assortments, the separation of minor species with current methods and remote sensing devices seems to be extremely difficult (e.g., Packalén et al. 2009; Packalén and Maltamo 2008). Tree species recognition has been extensively studied, and high-density ALS data or multispectral ALS, for example, have been suggested to aid in the separation of tree species (e.g., Vauhkonen et al. 2009; Yu et al. 2017; Budei et al. 2017). High-density ALS enables the detection of individual trees and applications that utilize individual tree approaches with diameter distributions predicted by NN imputation have been published (e.g., Hou et al. 2016).

The diameter distributions of coniferous species were generally predicted with better ErrorIndex values for separate NN imputation by tree species, compared to simultaneous NN imputation (Fig. 4 and Table 6). An examination of ErrorIndex values at the plot level revealed that ErrorIndex values degraded if the tree species in question was not a dominant species (see Fig. 5). Interestingly, when only the dominant species were considered, the ErrorIndex values were better in denser and young forests (Fig. 5). Improved minor tree species recognition is needed to curb the decreasing trend in ErrorIndex values (Fig. 5) in terms of stem number. However, it must be noted that some of the high-ErrorIndex values in Fig. 5 may be a result of a small number of trees in a plot, which usually means that the tree species in question has a

**Fig. 5** The illustration of plot level ErrorIndex values ( $\text{ErrorIndex} \leq 300$ ) versus logwood proportions (logwood proportion  $> 0\%$ ) and stem numbers when calibrated  $\text{SET}_{\text{sep1}}$  was applied. The colors and shapes describe the dominant species of the plot in question. The dominant species is determined according to the volume proportions





**Fig. 6** An example of the total volume calibration for a young spruce-dominated plot. The topmost figure describes the diameter distribution for Norway spruce when calibration is not applied. The lower figure shows the same case when calibration is applied. A bin width of 2 cm is used

minor role from the point of view of forest management. The pulpwood tail of the diameter distribution could be predicted with better ErrorIndex values than logwood, if only the dominant tree species are considered in the performance assessments (Fig. 5).

We would like to point out that there are some aspects to consider when extrapolating the results to other inventory areas. Firstly, the low-density leaf-off ALS data utilized were acquired at an optimal leaf-off period without snow. Previous studies have suggested that leaf-off data might be convenient for forestry purposes (Villikka et al. 2012). However, the acquisition of such data may be extremely difficult under Finnish conditions due to the narrow and unpredictable period of suitable weather and the growing season conditions in the spring (Villikka et al. 2012). Secondly, the NN imputations and predictor selections were carried out for the sake of computational efficiency, in a train mode in which the distance metrics are solved only once. According to the study of Packalén et al. (2012), the optimism due to the train mode of NN imputation increases when the amount of predictor variables increases. However, the optimism is highest when only one response is used. Here, we used 11 predictor variables in all analyses with a varying number of responses, but always with more than one, and, therefore, the optimism could be anticipated as being only very minor. Thirdly, the use of a heuristic optimization algorithm in the selection of predictor variables may not produce the best set of variables although a substantial number of iterations were used. For example, the selection of predictor variables by minimization of mean RMSEs of the 15 responses may result in slightly different predictions of diameter distributions between several optimizations processes. Fourthly, the number of sample plots must be considered; the data set consisted of 424 plots and can be regarded as comprehensive enough for species-specific forest inventories implemented by means of NN imputation. In our data set, only plots located completely within the borders of a forest stand were included in the training data. The exclusion of border plots might cause a simplifying effect on the training data due to the increase in homogeneity of forest structures in the plots. However, this exclusion is justified by the way the data is used in forest inventories in Finland.

Unfortunately, the data acquisition was not carried out in such a way that stand level calculations could be implemented. In the study by Haara and Korhonen (2004), RMSEs of traditional stand wise assessment were 52.0, 62.3, and 135.6% for the logwood volume of pine, spruce, and deciduous trees, respectively. In our study, the logwood volume prediction for spruce and for pine (to a lesser extent) performed well, although the RMSEs were calculated at the plot level. However, it should be noted that the prediction of diameter distributions by NN imputation requires DBH measurements for all trees in the sample plots. In addition, NN imputations are sensitive to the comprehensiveness of the data set, and all

of the various species compositions and tree dimensions should be included in the training data.

## 5 Conclusions

We applied NN imputation to predict species-specific diameter distribution by means of low-density leaf-off ALS data and aerial images. The results indicated that the careful selection of responses for NN imputation and the calibration of the diameter distribution to the total volume estimate can improve the error rates associated with volume predictions and the ErrorIndex values of diameter distributions. The separate NN imputation by tree species performed better, with lower error rates, for the predictions of logwood, pulpwood, and tree species volumes, compared to those predicted by the simultaneous NN imputation. At best, the error rate predictions for logwood volumes for coniferous species are satisfactory, taking into account that the predictions were carried out at the plot level. However, tree species recognition still poses a bottleneck in the prediction of diameter distributions by tree species with ALS and aerial images, and this can be seen in the poor error rates associated within minor species predictions. From the point of view of forest management and procurement, the most important attributes are, however, related to the dominant tree species.

**Acknowledgements** We would like to thank Prof. Heli Peltola and Prof. Jyrki Kangas for the acquisition of the financial support for the field measurements.

**Funding** This research is a contribution to the project *Comparative test to predict species-specific diameter distributions and distributions in forest information systems* funded by the Finnish Forest Centre (proj. 30033). This research was also supported by the project *Sustainable, climate-neutral, and resource-efficient forest-based bioeconomy* (FORBIO, proj. 14970), funded by the Strategic Research Council at the Academy of Finland.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Axelsson P (2000) DEM generation from laser scanner data using adaptive TIN models. *IAPRS*, Amsterdam, The Netherlands 33:111–118
- Bollandsås O, Maltamo M, Gobakken T, Næsset E (2013) Comparing parametric and non-parametric modelling of diameter distributions on independent data using airborne laser scanning in a boreal conifer forest. *Forestry* 86:493–501. <https://doi.org/10.1093/forestry/cpt020>
- Budei BC, St-Onge B, Hopkinson C, Audet F (2017) Identifying the genus or species of individual trees using a three-wavelength airborne lidar system. *Remote Sens Environ* 204:632–647. <https://doi.org/10.1016/j.rse.2017.09.037>

- Breidenbach J, Gläser C, Schmidt M (2008) Estimation of diameter distributions by means of airborne laser scanner data. *Can J For Res* 38: 1611–1620. <https://doi.org/10.1139/x07-237>
- Gobakken T, Næsset E (2004) Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. *Scan J For Res* 19:529–542. <https://doi.org/10.1080/02827580410019454>
- Gorgoso J, Álvarez González J, Rojo A, Grandas-Arias J (2007) Modelling diameter distributions of *Betula alba* L. stands in north-west Spain with the two-parameter Weibull function. *Forest Syst* 16: 113–123. <https://doi.org/10.5424/srf/2007162-01002>
- Haara A, Korhonen K (2004) Kuvioittaisen arvioinnin luotettavuus. *Metsätieteen aikakauskirja* 2004. <https://doi.org/10.14214/ma.5667>
- Hou Z, Xu Q, Vauhkonen J, Maltamo M, Tokola T (2016) Species-specific combination and calibration between area-based and tree-based diameter distributions using airborne laser scanning. *Can J For Res* 46(6):753–765
- Kangas A, Maltamo M (2000) Performance of percentile based diameter distribution prediction and Weibull method in independent data sets. *Silva Fenn* 34:381–398. <https://doi.org/10.14214/sf.620>
- Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. *Science* 220:671–680
- Koivuniemi J, Korhonen KT (2006) Inventory by compartments. *For Invent*:271–278. [https://doi.org/10.1007/1-4020-4381-3\\_16](https://doi.org/10.1007/1-4020-4381-3_16)
- Korpela I, Ørka H, Maltamo M, Tokola T, Hyyppä J (2010) Tree species classification using airborne LiDAR—effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. *Silva Fenn* 44:319–339. <https://doi.org/10.14214/sf.156>
- Laasasenaho J (1982) Taper curve and volume functions for pine, spruce and birch. *Commun Inst For Fenn* 108:1–74
- Lamb SM, MacLean DA, Hennigar CR, Pitt DG (2017) Imputing tree lists for New Brunswick spruce plantations through nearest-neighbor matching of airborne laser scan and inventory plot data. *Can J Rem Sens* 43:269–285. <https://doi.org/10.1080/07038992.2017.1324288>
- Lappi J (1993) *Metsäbiometrian menetelmiä*. University of Joensuu, Faculty of Forest Sciences
- Malinen J, Maltamo M, Harstela P (2001) Application of most similar neighbor inference for estimating marked stand characteristics using harvester and inventory generated stem databases. *IJFE* 12:33–41
- Maltamo M, Malinen J, Kangas A, Härkönen S, Pasanen A (2003) Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. *Forestry* 76:449–464. <https://doi.org/10.1093/forestry/76.4.449>
- Maltamo M, Eerikäinen K, Packalén P, Hyyppä J (2006) Estimation of stem volume using laser scanning-based canopy height metrics. *Forestry* 79:217–229. <https://doi.org/10.1093/forestry/cpl007>
- Maltamo M, Suvanto A, Packalén P (2007) Comparison of basal area and stem frequency diameter distribution modelling using airborne laser scanner data and calibration estimation. *For Ecol Manag* 247:26–34. <https://doi.org/10.1016/j.foreco.2007.04.031>
- Maltamo M, Næsset E, Bollandsås OM, Gobakken T, Packalén P (2009) Non-parametric prediction of diameter distributions using airborne laser scanner data. *Scand J For Res* 24:541–553. <https://doi.org/10.1080/02827580903362497>
- Maltamo M, Packalén P (2014) Species-specific management inventory in Finland. In: Maltamo M, Næsset E, Vauhkonen J (eds) *Forestry applications of airborne laser scanning*. Springer Netherlands, Dordrecht, pp 241–252
- Maltamo M, Mehtätalo L, Valbuena R, Vauhkonen J, Packalén P (2017) Airborne laser scanning for tree diameter distribution modelling: a comparison of different modelling alternatives in a tropical single-species plantation. *Forestry*, pp 1–11. <https://doi.org/10.1093/forestry/cpx041>
- Moeur M, Stage AR (1995) Most similar neighbor: an improved sampling inference procedure for natural resource planning. *For Sci* 41: 337–359
- Næsset E (1997) Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sens Environ* 61:246–253. [https://doi.org/10.1016/S0034-4257\(97\)00041-2](https://doi.org/10.1016/S0034-4257(97)00041-2)
- Næsset E (2004) Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. *Scand J For Res* 19:164–179. <https://doi.org/10.1080/02827580310019257>
- Næsset E (2014) Area-based inventory in Norway—from innovation to an operational reality. In: Maltamo M, Næsset E, Vauhkonen J (eds) *Forestry applications of airborne laser scanning*. Springer Netherlands, Dordrecht, pp 215–240
- Packalén P, Maltamo M (2007) The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sens Environ* 109:328–341. <https://doi.org/10.1016/j.rse.2007.01.005>
- Packalén P, Maltamo M (2008) Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Can J Res* 38:1750–1760. <https://doi.org/10.1139/X08-037>
- Packalén P, Suvanto A, Maltamo M (2009) A two stage method to estimate species-specific growing stock. *Photogramm Eng Remote Sens* 75:1451–1460. <https://doi.org/10.14358/pers.75.12.1451>
- Packalén P, Temesgen H, Maltamo M (2012) Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Can J Remote Sens* 38:557–569. <https://doi.org/10.5589/m12-046>
- 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 Fenn* 42. <https://doi.org/10.14214/sf.237>
- Poudel KP, Cao QV (2013) Evaluation of methods to predict Weibull parameters for characterizing diameter distributions. *For Sci* 59: 243–252. <https://doi.org/10.5849/forsci.12-001>
- R Core Team (2017) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing
- Reynolds M, Burk T, Huang W (1988) Goodness-of-fit tests and model selection procedures for diameter distribution models. *For Sci* 34: 373–399
- Saad R, Wallerman J, Lämäs T (2015) Estimating stem diameter distributions from airborne laser scanning data and their effects on long term forest management planning. *Scand J For Res* 30:186–196. <https://doi.org/10.1080/02827581.2014.978888>
- Shang C, Treitz P, Caspersen J, Jones T (2017) Estimating stem diameter distributions in a management context for a tolerant hardwood Forest using ALS height and intensity data. *Can J remote Sens* 43: 79–94. <https://doi.org/10.1080/07038992.2017.1263152>
- Siipilehto J (1999) Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. *Silva Fenn* 33. <https://doi.org/10.14214/sf.650>
- Strunk JL, Gould PJ, Packalén P, Poudel KP, Andersen H, Temesgen H (2017) An examination of diameter density prediction with k-NN and airborne Lidar. *Forests* 8:444. <https://doi.org/10.3390/f8110444>
- Thomas V, Oliver RD, Lim K, & Woods M (2008). LiDAR and Weibull modeling of diameter and basal area. *For Chron*, 84(6), 866–875. <https://doi.org/10.5558/ffc84866-6>
- Tuominen S, Haapanen R (2013) Estimation of forest biomass by means of genetic algorithm-based optimization of airborne laser scanning and digital aerial photograph features. *Silva Fenn* 47. <https://doi.org/10.14214/sf.902>
- 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. *For Sci* 55:37–47. <https://doi.org/10.5589/m08-052>

- Villikka M, Packalén P, Maltamo M (2012) The suitability of leaf-off airborne laser scanning data in an area-based forest inventory of coniferous and deciduous trees. *Silva Fenn* 46:99–110. <https://doi.org/10.14214/sf.68>
- White JC, Wulder MA, Varhola A, Vastaranta M, Coops NC, Cook BD, Pitt D, Woods M (2013) A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. *For Chron* 89:722–723. <https://doi.org/10.5558/tfc2013-132>
- Yu X, Hyypä J, Litkey P, Kaartinen H, Vastaranta M, Holopainen M (2017) Single-sensor solution to tree species classification using multispectral airborne laser scanning. *Remote Sens* 9:108. <https://doi.org/10.3390/rs9020108>