SILVA FENNICA

Silva Fennica 42(4) research articles

www.metla.fi/silvafennica · ISSN 0037-5330
The Finnish Society of Forest Science · The Finnish Forest Research Institute

Estimating Species-Specific Diameter Distributions and Saw Log Recoveries of Boreal Forests from Airborne Laser Scanning Data and Aerial Photographs: a Distribution-Based Approach

Jussi Peuhkurinen, Matti Maltamo and Jukka Malinen

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.

The low-density airborne laser scanning (ALS) data based estimation methods have been shown to produce accurate estimates of mean forest characteristics and diameter distributions, according to several studies. The used estimation methods have been based on the laser canopy height distribution approach, where various laser pulse height distribution -derived predictors are related to the stand characteristics of interest. This approach requires very delicate selection methods for selecting the suitable predictor variables. In this study, we introduce a new nearest neighbor search method that requires no complicated selection algorithm for choosing the predictor variables and can be utilized in multipurpose situations. The proposed search method is based on Minkowski distances between the distributions extracted from low density ALS data and aerial photographs. Apart from the introduction of a new search method, the aims of this study were: 1) to produce accurate species-specific diameter distributions and 2) to estimate factual saw log recovery, using the estimated height-diameter distributions and a stem data bank. The results indicate that the proposed method is suitable for producing species-specific diameter distributions and volumes at the stand level. However, it is proposed, that the utilization of more extensive and locally emphasized reference data and auxiliary variables could yield more accurate saw log recoveries.

Keywords diameter distribution, non-parametric estimation, pre-harvest measurement, stem data bank
Addresses *Peuhkurinen* and *Maltamo*, University of Joensuu, Faculty of Forest Sciences, P.O. Box 111, FI-80101 Joensuu, Finland; *Malinen*, Finnish Forest Research Institute, Joensuu Research Unit, P.O. Box 68, FI-80101 Joensuu, Finland
Received 26 June 2007 **Revised** 10 April 2008 **Accepted** 23 May 2008
Available at http://www.metla.fi/silvafennica/full/sf42/sf424625.pdf

1 Introduction

Methods based on low-density airborne laser scanning (ALS) data have been shown to produce accurate estimates of forest characteristics on an operational scale (estimation unit from hundreds of square meters to hectares). These methods have been based on the laser canopy height distribution approach, where various laser pulse height distribution -derived predictors are related to the stand characteristics of interest. The most commonly used methods have been that of linear regression (e.g. Means et al. 2000, Næsset 2002, Lim et al. 2003, Holmgren 2004, Jensen et al. 2006), and non-parametric estimation (Maltamo et al. 2006, Packalén and Maltamo 2006, Packalén and Maltamo 2007).

Although regression methods are widely used and they tend to provide reliable results, non-parametric nearest neighbor (NN) methods have certain advantages over parametric methods, especially, when simultaneously considering the estimation of various characteristics of interest. According to LeMay and Temesgen (2005) NN methods:

a) retain the variance structure of the data;
b) will produce realistic results, since estimates are within the bounds of real data;
c) are distribution-free, therefore distributional assumptions are not required for predictor variables nor for the variables of interest; and
d) can be easily used in multivariate situations.

However, advantages a) and b) cannot be guaranteed if more than one nearest neighbors are used for the estimation.

The studies by Maltamo et al. (2006), Packalén and Maltamo (2006), and Packalén and Maltamo (2007) have demonstrated that the non-parametric k Most Similar Neighbor (k-MSN) method is very suitable for estimating mean stand characteristics from the ALS data. Furthermore, by combining aerial photographs and ALS data, the overall estimation accuracy can be improved and it is possible to obtain species-specific estimates (Packalén and Maltamo 2006, Packalén and Maltamo 2007).

There are some issues that need to be considered when applying NN method: 1) a distance metric used in the selecting the neighbors, 2) the number of neighbors, and 3) the weighting of the neighbors (LeMay and Temesgen 2005). Important issue is also how the selection of the predictor variables is carried out. Usually, in the canopy height distribution method, various predictor variables are calculated from the distribution of laser pulse heights, and then the most significant and least cross-correlated ones are chosen for the model. If some optical remote sensing material (e.g. aerial photographs) is additionally used in the estimation, the number of candidate predictor variables can be vast, so the selection procedure for finding the best predictors or predictor combinations becomes complicated. This relates to both of the methods: linear regression and non-parametric estimation. In the case of NN methods, another problem is the "curse of dimensionality" (Bellman 1961). This means that when the number of predictor variables increases (the number of dimensions increase), the solution of the estimation task rapidly becomes more complex.

The natural way of describing the ALS data are laser pulse height, and intensity, distributions. The spectral information of the aerial photographs can be represented in the form of distributions, as well. Thus, the same methodology could be used in analyzing both of the data sources. To avoid the above problems considering the selection of the predictor variables and to keep in mind the distributional characteristics of the laser height, intensity, and spectral distributions, a search method for finding the nearest neighbors should satisfy two conditions: 1) it should take into consideration the variability and the heterogeneous structure of the laser height, intensity, and spectral distributions; and 2) it should be based on as few predictor variables as possible. To realize this, a new nearest neighbor search method will be introduced which is based on Minkowski distance of order one (later: Minkowski distance).

In this study we test the proposed search method by estimating species-specific volumes and diameter distributions using laser pulse height distributions and spectral histograms derived from aerial photographs. The estimation of species-specific diameter distributions is routine operation in field measurement based forest inventories (e.g. Siipilehto 1999, Newton et al. 2004) whereas in the case of remote sensing data only diameter

distributions for the tree stock as a whole has been considered so far (Gobakken and Næsset 2004, Maltamo et al. 2007).

Additional goal of the study is to test the estimation method in producing estimates of the factual saw log recovery (i.e. the saw log recovery which includes reduction due to bucking constraints (allowable length and diameter combinations) and external technical defects) using estimated species-specific height-diameter distributions and stem data bank. The estimation of the factual saw log recovery is based on the assumption that the height-diameter distribution and tree species composition can be used as predictor variables in estimating the external quality of trees, which, in other hand, could be used in predicting the quality of wood.

The result of the estimation chain (i.e. species-specific diameter distributions and saw log recoveries) can be considered as pre-harvest measurement information. Currently, pre-harvest measurement is usually carried out by field work, but different computational methods have also been developed in Nordic countries. The most promising ones has been based on nonparametric k-MSN and k-Nearest Neighbor (k-NN) regression (e.g. Tommola et al. 1999, Malinen et al. 2001).

To sum up the aims of this study are 1) to introduce a new search method for producing nearest neighbor estimates from remote sensing material and 2), to produce accurate species-specific diameter distributions through the use of ALS data and aerial photographs. Additionally, the proposed method is tested in estimating factual saw log recoveries with the use of estimated height-diameter distributions and the stem data bank.

2 Materials

2.1 Study Area

The study was conducted at the Matalansalo study area, located in the municipality of Varkaus in eastern Finland (about 28°29´E, 62°18´N) and owned by the forest industry company UPM Kymmene Ltd. The Matalansalo study area can be considered as a typical managed boreal forest area in Finland. The total area is approximately 1200 hectares and it is dominated by coniferous species (Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.)); deciduous species are found mainly as a minority in mixed species stands.

2.2 Field Data Collection

2.2.1 Plot Data

In summer 2004, a total of 474 circular plots each with a 9 meter radius were systematically placed over the 67 forest stands and positioned with Trimble Pro XRS GPS that used real time differential correction. Diameter at breast height (dbh), species, tree storey and tree class (dead, alive) were all recorded for every tree inside the plot with a dbh at least 5 cm. In addition, one tree from every tree storey by species per plot was chosen as a height sample tree. Heights for the rest of the trees were calculated using the species-specific height models of Veltheim (1987), which were calibrated by plots with the use of sample tree measurements. The plot measurements did not include the assessment of technical quality of the trees, and therefore, it was not possible to extract factual saw log recoveries from the plot data. Thus, the circular plot data (Table 1) was used only as a reference in the estimation of the species-specific volumes and diameter and height-diameter distributions.

2.2.2 Harvester-acquired Data

A total of 14 marked stands located in the study area were clear cut after acquisition of the remote sensing material. These test stands were delineated using Trimble GeoXT GPS and differential correction in a post-processing mode. To avoid the effect of trees left in the clear cut area, retention tree groups of more than 2 trees were also delineated and left out of the analysis. The harvester data of the stands consisted of STM files, which include the position of the harvester at the time of the felling, harvester measured diameters of the stem in 10-cm intervals from the felling point to the last cutting point, length of the usable trunk part, species, and bucking parameters and buck-

Table 1. Mean characteristics of 474 reference plots. Only the living trees are included.

	HGM, m	DGM, cm	Stock, n ha^{-1}	Total volume, m^3 ha^{-1}	Volume of pine, m^3 ha^{-1} (% of total volume)	Volume of spruce, m^3 ha^{-1} (% of total volume)	Volume of deciduous species, m^3 ha^{-1} (% of total volume)
Minimum	6.0	7.6	275	14.5	0 (0)	0 (0)	0 (0)
Maximum	30.6	43.6	4048	601.7	422.1 (100)	500.3 (100)	216.8 (100)
Mean	17.0	19.8	1505	199.7	95.4 (53.4)	82.8 (34.2)	21.5 (12.4)
Standard deviation	5.1	6.4	691	101.0	86.4 (39.5)	110.6 (35.8)	35.9 (20.7)

Table 2. Mean characteristics of 14 test stands based on STM files.

	Area, ha	Pine		Spruce		Volume of deciduous species, m^3 ha^{-1}	Total volume, m^3 ha^{-1}
		Total volume, m^3 ha^{-1}	Volume of saw log, m^3 ha^{-1}	Total volume, m^3 ha^{-1}	Volume of saw log, m^3 ha^{-1}		
Minimum	0.3	0.3	0.2	15.9	5.6	1.8	134.4
Maximum	8.0	208.0	186.3	279.1	203.3	45.9	333.5
Mean	3.1	68.3	53.9	172.4	110.6	14.0	254.7
Standard deviation	2.6	69.2	62.3	84.3	65.6	14.1	69.5

Table 3. Proportions of tree species and saw log proportions in the test and stem data bank stands. Number of stands in the test data were 14 and in the stem data bank 35.

	Test stands				Stem data bank stands			
	Minimum	Maximum	Mean	Standard deviation	Minimum	Maximum	Mean	Standard deviation
Percentage of spruce	10.3	98.0	67.3	26.1	0.0	100.0	70.3	25.4
Percentage of pine	0.1	86.8	27.5	24.9	0.0	100.0	24.4	25.0
Percentage of deciduous species	1.4	14.9	5.2	4.3	0.0	17.1	4.8	4.8
Saw log proportion of spruce	15.8	76.4	61.2	16.7	30.9	88.3	72.4	13.0
Saw log proportion of pine	40.4	89.6	72.3	14.3	52.8	92.3	75.4	10.4

ing results (e.g. factual saw wood recovery) for every harvested tree, according to the Standard for Forest Data and Communication (StanForD 2006). Harvested trees were assigned to the correct stand according to the coordinates of the harvester provided in the STM files, and the diameter distributions and bucking results extracted from the stand-wise harvester data were used as test data (Table 2).

The circular sample plot data and the harvester data were collected independently. Eight of the total 14 test stands were such that they contained sample plots or portions of the sample plots. A total of 45 complete plots were located inside the test stands.

2.3 Stem Data Bank

The harvester-acquired data from the study area consisted only of 14 test stands. Thus, we used separate stem data bank for estimating factual saw

log recoveries for the test stands to have more extensive reference data. The stem data bank data consisted of 35 mainly spruce dominated clear cut stands originally collected for the Finnish Forest Research Institute's research project "Value formation of timber stand when targeting for alternative end-products in timber harvesting". The stem data bank stands were located outside the laser scanned study area approximately 10 to 250 kilometers north-east, south and west from the test stands. The stem data bank included tree data extracted from STM files (i.e. species and bucking results for every harvested tree). Unlike the test stands, stem data bank stands were not delineated on the ground, therefore per hectare information could not be extracted from them. Nevertheless, test stands and stem data bank stands can be compared using tree species and saw log proportions (Table 3). According to this comparison, the stem data bank data was extensive despite insufficient stands with low saw log proportions.

2.4 Remote Sensing Material

2.4.1 ALS Data and Processing

The ALS data covering the study area were collected on the 4th of August, 2004 using an Optech ALTM 2033 laser scanning system. The operating altitude was 1,500 meters above ground and the field of view was 30 degrees. This resulted in a nominal sampling density of about 0.7 measurements per square meter and a footprint of 45 cm at the ground level. A digital terrain model (DTM) was processed from last pulse data with TerraScan software (see www.terrasolid.fi). At first, the ground points were separated from other points using the method explained in Axelsson (2000). Then a raster DTM was created from the classified ground points by calculating their mean values within each one-metre raster cell. Values for raster cells with no data were derived by Delaunay triangulation and the bilinear interpolation method. The ALS data was further processed by subtracting DTM from the laser pulses to produce a point cloud with x, y and dz coordinates, where dz is the height above the ground. Only the first pulse data was utilized in the estimation procedure of the forest characteristics.

2.4.2 Aerial Photography Data and Processing

In addition to ALS data, color-infrared photographs at a scale of 1:30,000 were acquired on the 22nd of August 2004 with a Leica RC30 camera having a UAGA-F 13158 objective of focal length 163.18 mm and an anti-vignetting filter (AV525 nm). The films were digitized at a resolution of 14 μm, orthorectified using the DTM generated from the ALS data, and resampled to a pixel size of 50 cm. Since the study area was covered by three aerial photographs, radiometric calibration was required in order to make the images comparable. The correction was performed by the method presented by Tuominen and Pekkarinen (2004), using a Landsat 7 ETM satellite image taken in June 2002 of the study area. The radiometric resolution of the final images was 8 bits.

3 Methods

The study followed the structure briefly described below. The estimation procedure itself is presented schematically in a flowchart (Fig. 1).

At first, we formulated a suitable distance metric to be used with ALS data and spectral histograms in k-NN estimation. Then the estimation method was applied to the plot level in order to study the effectiveness and suitability of the method and to discover the optimal estimation parameters. After this, the species-specific volumes, diameter, and height-diameter distributions were predicted for the test stands.

Remote sensing material covered only the study area, not the stem data bank data area. Thus, the two-stage KNN approach was used in estimation of the saw log recoveries. The estimated species-specific height-diameter distributions were used to find k nearest reference stands from the stem data bank in order to produce saw log volume estimates. Finally, the results were evaluated using root mean square error of the estimates (RMSE) and Reynolds' error index (Reynolds et al. 1988). The following sections give a more detailed description of the methods used.

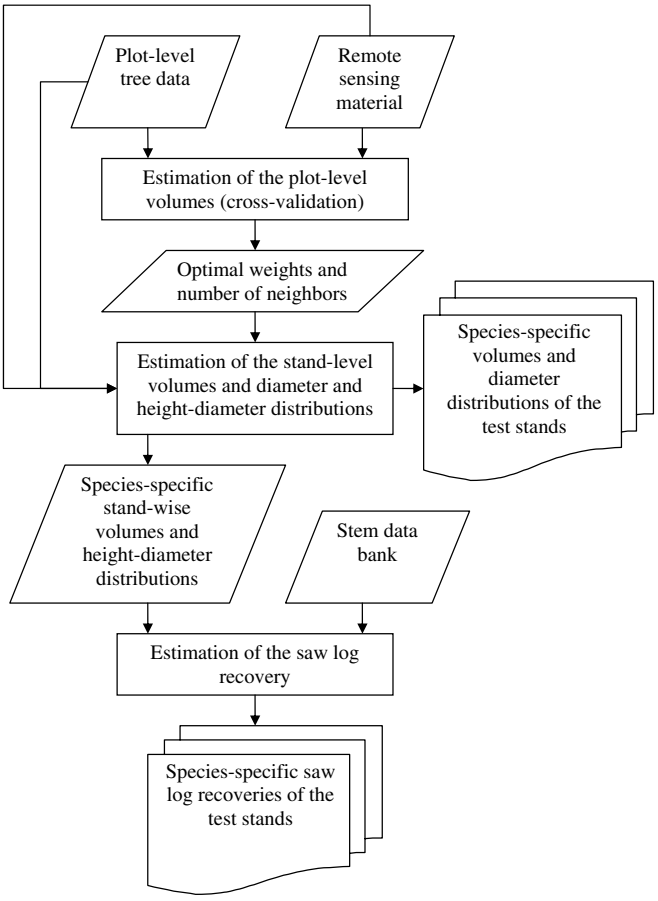

Fig. 1. A schematic presentation of the estimation procedure.

3.1 Estimation Method

The chosen method for estimation was a nonparametric k-NN method. The first step in the NN method is to define the distances between the target and reference objects. The distance metric used to fulfill the requirements mentioned earlier in this paper (it should take into consideration the whole variability and the nonparametric nature of the studied distributions, and it should be based on as few predictor variables as possible) was the Minkowski distance of order one, applied to classified distributions, and it was of the form:

$$D_{pq} = \sum_{i=1}^{n} \left| p_i - q_i \right| \quad (1)$$

where D_{pq} is distance between compared distributions, p_i is the proportion of observations of the test distribution in class i, q_i is the proportion of observations of the reference distribution in class i, and n is the number of classes in distributions. This measure yields one value that relates to the absolute difference of classified distributions. The value of D_{pq} varies from 0 (compared distributions are the same) to 2 (compared distributions have no observations in same classes). After computing the distances, the neighbors are sorted, and the nearest neighbors are assigned for every target object. There is no need to select among several variables, or to find the optimal weights for them, because the only information that is extracted from the difference of the distributions is given in one explicable numeric value.

3.2 Combining Distances from ALS Data and Several Bands

In the case of using more than one predictor variable, weights for them have to be somehow decided. The weights could not be searched for using multidimensional feature space, because the predictor variable (D_{pq}) used in this study is a distance metric by definition. Thus, the distances were calculated separately for ALS data and the aerial photograph's bands, and the weights for them were searched for afterwards using a brute-force search algorithm. The algorithm formulated weight combinations systemically so that every data set was given a weight from 0 to 1 with an interval of 0.05 and all weight combinations which summed up to 1 were examined. The optimization task was then to find the weight combination, which minimized the estimation error.

3.3 Plot-wise Volume Estimates

To examine the functionality of the proposed estimation method, it was at first tested in the estimation of the species-specific volumes for the measured reference plots. Because of the relatively small plot size (254.5 m^2), and therefore, a small number of laser pulses per plot, ALS data was classified using a 1 meter class size in order to avoid a prevalence of empty classes. Reclassification resulted, on average, in 5–6 laser observations for every class per plot. The spectral values of the aerial photographs were not reclassified; consequently, the processing was based on the original 256 classes, with an average of 4 observations in each of them at the plot-level. The volume estimates were produced from the leave-one-out cross validation in such a manner that the plots within the same stand as target plot were excluded from the reference data. The number of nearest neighbors was restricted to ten and the neighbors were weighted by 2 minus the distance measure D_{pq}. This estimation was done using either ALS data or aerial photographs, or both. The optimal weights for the ALS data and different bands of the aerial photographs were searched for by minimizing the RMSE of the total volume or, alternatively, the sum of the RMSEs of the species-specific volumes.

3.4 Estimating Stand Level Species-specific Diameter and Height-diameter Distributions

Estimating diameter and height-diameter distributions for the test stands followed a two-phase grid approach. At the first step, a grid with a 16x16 meter cell size (area corresponding to the size of the reference plot) was laid over the test stands, and the nearest reference plot for every grid cell was sought based on distances between spectral (in the case of aerial photographs) or height (in the case of ALS data) distributions by using the method described here. At the second step, the cell estimates were added up for the stand level. To produce the diameter and height-diameter distributions from the reference plot data, trees of the chosen (nearest) reference plots were added to a table containing stand wise height-diameter classes by species. Since all the cells were not complete (cells in the border of the stands), the number of each added tree was weighted by the ratio of the cell area and reference plot area. As a result of this procedure, we obtained three height-diameter distributions (for pine, spruce, and deciduous species) at the chosen height and diameter classification.

For the estimation of diameter distributions, we decided to use two centimeters even classification, which we supposed would give a detailed enough description of the stand structure. To enable sensible comparison of height-diameter distributions (to ensure enough observations in height-diameter classes), diameters were classified by two centimeters and heights in one meter classes. In the study area, this theoretically corresponds to 2-dimensional discrete distribution for about a size of 30x30 classes.

Because the STM-files of the test stands only included the length of the usable trunk part (i.e. the part of the trunk from the stump height to the final cutting point) the actual tree heights could not be used to produce height-diameter distributions. Thus, the lengths and volumes of the usable trunk parts were calculated for the trees of the reference plots with the use of the taper curve functions designed by Laasasenaho (1982).

The prediction of diameter distributions for the test stands was conducted in a similar manner as the prediction of plot-wise volumes with the

use of either ALS data or aerial photographs, or both. The weights used for the bands were those found in the optimization of plot-wise volumes, because the optimization of minimal errors in diameter distributions in small area plots with few trees is unstable and prone to wrong decisions. The number of the nearest neighbors was also reasoned according to the plot-wise results and the optimal number was found to be three. However, deciding how many neighbors to use is not a clear issue (LeMay and Temesgen 2005) and it will be discussed more thoroughly later on in this paper. The estimation of the diameter distributions produced also stand-level volume estimates as an additional variable of interest.

3.5 Estimating Saw Log Recovery

Factual saw log recoveries were considered to be the amount of actually harvested saw log volume extracted from the STM files. Apart from the factual saw log recovery, the total volume of the usable trunk part and the saw log proportion were also examined. The saw log proportion here is considered to be the percent of the saw log recovery from the volume of the usable trunk part.

The estimates for the usable trunk part can be produced directly from the trees of the reference plots, or indirectly by using the stem data bank as reference data. In the first case, the species specific volume estimates can be provided simultaneously with the estimation of the height-diameter distributions, by summing-up the volumes of the trees from the chosen (nearest) reference plots.

The reference plots did not provide information about the factual saw log recoveries because no technical quality attributes were measured. Thus, the estimation of the saw log recovery and saw log proportion was performed by searching for stands from the stem data bank that have the most similar height-diameter distributions with the test stand under consideration. This procedure is analogous with the method used for estimating the species-specific height-diameter distributions. At first, the distances between the height-diameter distribution of a test stand and the reference stands (stem data bank stands) were calculated by using the equation:

$$D_{rs} = \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{k}\left|r_{ijk} - s_{ijk}\right| \quad (2)$$

where m is the number of the tree species, n is the number of the diameter classes and k is the number of the height classes in the species-specific height-diameter distributions. Proportions r and s are calculated from the number of all trees in the stand. This equation is a generalization of the (Eq. 1) to a 3-dimensional distribution.

We intend to test the effect of using more than one nearest neighbor in estimation, and therefore, the selected neighbors were weighted by 2 minus D_{rs}. The species-specific estimates of the variables of interest were then calculated with the equation:

$$\mathrm{est}_v = \sum_{i=1}^{k}\left[\frac{V_{\mathrm{test}}}{V_{\mathrm{ref}_i}}\left(w_i \,/\, \sum_{j=1}^{k} wj\right)\mathrm{ref}_{v_i}\right] \quad (3)$$

where k is the number of the nearest neighbors, V_{test} is the estimated species-specific volume of the test stand and V_{ref} is the correct volume of the reference stand, w is the weight for the reference stand $(2 - D_{rs})$, and ref_v is the value of the variable of interest for the reference stand. Species-specific volumes were used in the equation to scale the estimate in the test stand level. This was done because the reference stands (stem data bank stands) had no precise information about the stand area, and therefore, we had to use absolute values, which needed to be scaled to the test stand level. The optimal number of the nearest neighbors was decided heuristically based on the estimation results achieved by using harvester measured height-diameter distributions and by taking into account the estimation error and bias, i.e. the number of neighbors was increased until the estimation error did not decrease or the bias started to increase significantly.

The small number of stands in the stem data bank and the geographical differences between the test and stem data bank stands could plausibly affect the results, therefore we decided to test the alternative where the test stands were included in the stem data bank to provide more locally emphasized data. In this case the estimation was conducted with the leave-one-out cross-validation, excluding the test stand for which the attributes were predicted from the stem data bank.

3.6 Reliability Characteristics

The diameter and height-diameter distributions which were estimated with the use of different datasets were compared by calculating Reynolds' error indices (Reynolds et al. 1988):

$$e = \sum_{i=1}^{k} w_i \left| f_i - \hat{f}_i \right| \quad (4)$$

where f_i and $\hat{f}_i$ are the numbers of stems in class i to be compared, and w_i is the weight of class i. It is noticeable, that Reynolds' error index is similar to Minkowski distance of order one with the exception that Reynolds' error index use weighting. In this case, basal area was used as a weight; therefore, the number of large trees had a major influence on the result. However, the basal area weighting is justified, since the large trees are more interesting from a wood procurement point of view, and the small trees (commercially invaluable trees) are not exactly described in the harvester data. In the case of the diameter distributions, the index i stands for diameter class and in the case of height-diameter distributions it stands for height-diameter class.

The reliabilities of the saw log recoveries and other volume estimates were investigated using root mean square error:

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

where n is the number of the stands, y_i and $\hat{y}_i$ are the observed and predicted values of variable y in stand i. The estimation biases were calculated as

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

The relative RMSEs were computed by dividing absolute RMSE values (5) by the observed mean value of the variable of interest.

4 Results

4.1 Plot-level Volume Estimates

Plot-wise species-specific volume estimates were at their best when both ALS data and aerial photographs were used together. The RMSE of the total volume was, at its lowest, about 20%, whereas that of coniferous species was from 60 to 80% and that of deciduous about 150% (Table 4).

The estimation error of the total volume decreased significantly as the weight of the ALS data increased (Fig. 2). The optimal weight of the ALS data was between 0.5 and 0.7, depending on the optimization task. The more the species-specific volumes were emphasized, the more important the influence of the spectral information became. However, using only aerial photographs in the estimation gave notably worse results and the use of the ALS data alone gave relatively good results compared to the ones

Table 4. The relative RMSE's of plot-level volume estimates, optimal weights, and optimal numbers of nearest neighbors.

	Laser data	Aerial photograph [a)]	Laser + aerial photograph [a)]	Laser + aerial photograph [b)]
Volume of pine	67.8	85.4	66.8	64.5
Volume of spruce	81.8	97.6	77.9	77.8
Volume of deciduous species	153.2	158.8	150.7	145.4
Total volume	22.4	43.4	21.8	22.8
Weight of laser data	1.00	0.00	0.70	0.50
Weight of near-infrared band	0.00	0.30	0.10	0.35
Weight of red band	0.00	0.70	0.10	0.10
Weight of green band	0.00	0.00	0.10	0.05
Number of neighbors	10	9	10	5

[a)] Optimization task is to minimize RMSE of total volume.
[b)] Optimization task is to minimize the sum of RMSE's of species specific volumes.

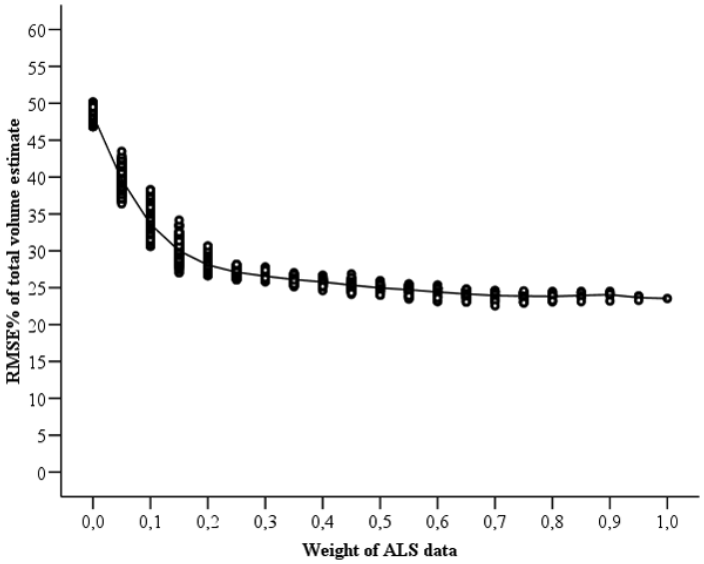

Fig. 2. The effect of the weight of the laser data on the estimation error of total volume at the plot-level. Interpolation line describes the average error.

achieved by utilizing both data sources with the optimal weights. The near-infrared band seemed to provide more auxiliary information than red or green bands, but no evident difference in relation to which bands of the aerial photograph should be used in the estimation besides the ALS data was noticed. The number of neighbors considerably affected the results up to 3–5 nearest neighbors. After that the estimation accuracy improved only slightly. Nevertheless, we noticed no increasing trend for the biases even when we used as many as 10 neighbors.

4.2 Stand-level Results

Species-specific diameter and height-diameter distributions were predicted using the optimal

Table 5. The main stand level results of diameter distribution, height-diameter distribution, and volume estimation by using the sample plots as reference data. The weights of the ALS data and bands of aerial photographs are the optimal ones found in the prediction of plot-wise volumes and the number of used neighbors is three.

Data source(s)	Reynolds' error index of diameter distribution	Reynolds' error index of height-diameter distribution	RMSE%				Bias%			
			Volume of pine	Volume of spruce	Volume of deciduous species	Total volume	Volume of pine	Volume of spruce	Volume of deciduous species	Total volume
Laser data	451.4	1093.1	52.2	24.2	78.6	6.7	4.4	5.2	–25.7	3.3
Aerial photograph [a)]	605.8	1228.8	77.8	20.7	103.0	18.1	–4.8	1.1	–61.1	–3.9
Laser + aerial photograph [a)]	435.0	1114.2	47.7	20.3	72.9	7.0	5.8	–2.0	–13.6	–0.6
Laser + aerial photograph [b)]	450.8	1161.8	52.4	23.0	68.8	8.9	7.5	–4.5	–26.3	–2.5

[a)] Optimization task is to minimize RMSE of total volume.
[b)] Optimization task is to minimize the sum of RMSE's of species specific volumes.

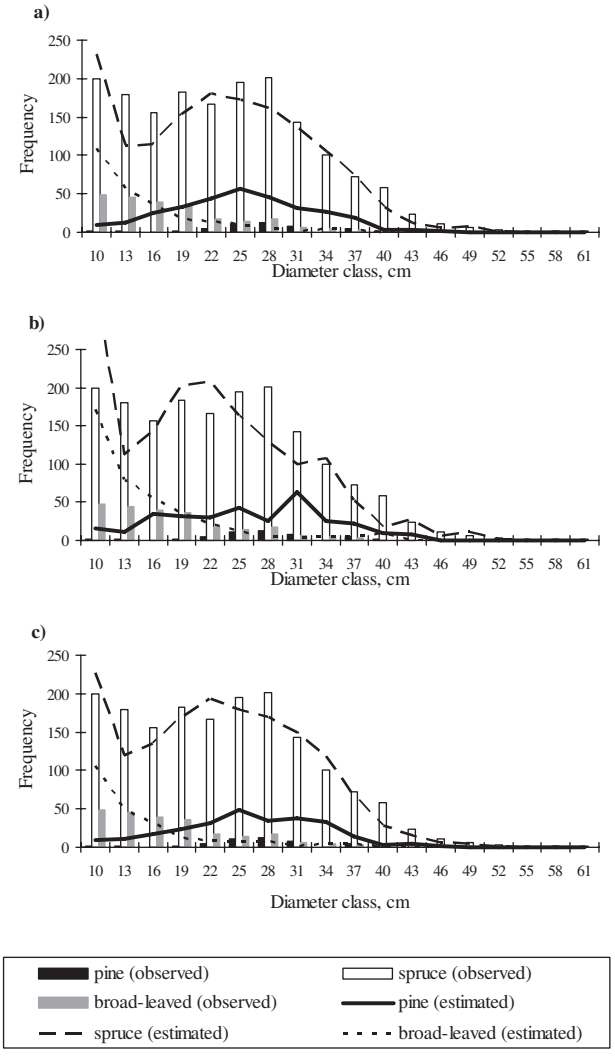

Fig. 3. An example of estimated diameter distributions for a spruce dominated stand. Distribution a) is estimated by using laser data, b) by using aerial photograph, and c) by using laser data and aerial photographs simultaneously.

weights found in the optimization of the plot-level volume estimates and three nearest neighbors. By using aerial photographs exclusively, the estimation error was significantly higher than if the ALS data were included in the analysis (Table 5).

Visual examination of the diameter distributions established the conclusion that ALS data performs better than aerial photographs and that the additional use of the aerial photographs slightly enhances the estimation result (Fig. 3 and Fig. 4). Visual examination also revealed that the diameter distributions of the dominant tree species are predicted much more accurately than those of the minor species; the estimated diameter distributions for the dominant tree species can be considered accurate also in the cases where the test stand is a mixed-species stand, if ALS data and aerial photographs are both used in the estimation

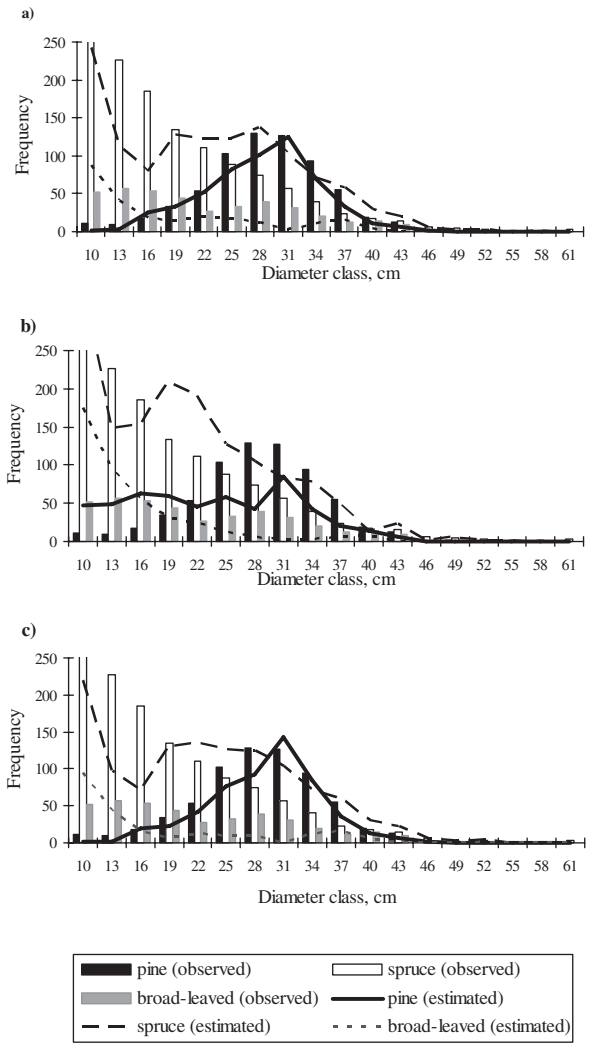

Fig. 4. An example of estimated diameter distributions for a pine dominated stand. Distribution a) is estimated by using laser data, b) by using aerial photograph, and c) by using laser data and aerial photographs simultaneously.

(Fig. 4). However, utilizing only the ALS data produced the most precise height-diameter distributions, obviously as a result of a direct physical relationship between the tree height distribution and the height distribution of the laser points. The species-specific volume estimates were at their best when both data sources were used in the estimation. By using the ALS data, the total volume was predicted with a RMSE of approximately 7%. Inclusion of the aerial photograph improved the prediction of the species-specific estimates, especially when deciduous species were considered. By using both data sources, the RMSE of the volume of the dominant tree species was less than 20%, whereas for the minority species it was approximately 70%.

Table 6. The main results of saw log recovery and saw log proportion estimation using the stem data bank as reference data. The values in parentheses are the corresponding results when test stands are included in the reference data. The weights of the ALS data and bands of aerial photographs are the optimal ones found in the prediction of plot-wise volumes and the number of used neighbors is 4 and (5).

Data source(s)	RMSE%				Bias%			
	Saw log volume of pine	Saw log volume of spruce	Saw log proportion of pine	Saw log proportion of spruce	Saw log volume of pine	Saw log volume of spruce	Saw log proportion of pine	Saw log proportion of spruce
Laser data	61.9 (61.5)	32.1 (33.1)	28.3 (20.3)	27.2 (29.7)	–23.9 (–20.7)	3.3 (2.0)	–2.3 (–0.5)	11.4 (12.3)
Aerial photograph [a)]	95.7 (65.1)	26.5 (30.9)	29.6 (21.1)	22.1 (24.9)	–15.0 (–16.9)	5.3 (–1.4)	–3.4 (0.7)	9.1 (7.8)
Laser + aerial photograph [a)]	66.0 (61.8)	31.8 (31.5)	32.3 (23.4)	27.7 (29.3)	–29.6 (–21.2)	11.1 (8.7)	–7.5 (–1.4)	11.6 (10.2)
Laser + aerial photograph [b)]	69.7 (63.7)	34.8 (34.2)	31.7 (23.2)	27.8 (28.7)	–30.3 (–20.3)	13.6 (11.3)	–4.5 (–0.2)	11.5 (9.8)
STM data	20.2 (12.5)	21.7 (21.3)	25.9 (24.4)	26.4 (25.7)	–9.2 (–3.1)	11.6 (4.4)	–0.3 (–3.3)	13.0 (5.5)

[a)] Optimization task is to minimize RMSE of total volume.
[b)] Optimization task is to minimize the sum of RMSE's of species specific volumes.

4.3 Saw Log Recoveries

Saw log recoveries and saw log proportions were predicted by utilizing the stem data bank and the estimated height-diameter distributions. At first, the estimation was done using the correct harvester measured height-diameter distributions. It was noticed that even though the height-diameter distributions were correct, the estimated saw log recoveries and proportions had rather large errors (Table 6). The final predictions were carried out with and without the inclusion of the test stands in the stem data bank. The results suggest that more comprehensive or locally representative data improves the estimation accuracy and diminishes the biases. The most suitable number of neighbors was 4 before the inclusion of the test stands in the stem data bank and 5 if the test stands were included. After these numbers, the errors did not diminish significantly or the biases began to increase. The harvester measured height-diameter distribution (the correct distribution) gave the most reliable saw log volume estimates, whereas in the case of the saw log proportions, the estimation errors were substantial despite the used height-diameter distribution.

5 Discussion

This study concerned itself with the prediction of the species-specific diameter distributions and saw log volumes by utilizing ALS data, aerial photographs, and stem data bank. The estimation was carried out by a non-parametric k-NN estimation method for which a new nearest neighbor search method was developed, based on the Minkowski distances between the distributions of the predictor variables. The estimation procedure is suitable for the current situation regarding data availability, i.e. harvester data and remote sensing material are collected independently and from the different areas. Thus, the results of the study should be seen not as the final accuracy for utilization of the ALS and harvester data, but as the accuracy level that is currently achievable.

The proposed estimation method was supported by the plot level results of the species-specific volumes of the growing stock, although, the estimation accuracies were not as good as those achieved by Packalén and Maltamo (2006) and Packalén and Maltamo (2007). However, the estimates remained fairly unbiased regardless the number of neighbors used. It is also noticeable that even though the smallest RMSE's were achieved when

the number of nearest neighbors was 5–10, the accuracies did not significantly increase after 3–5 neighbors. As mentioned earlier, the choice of the optimal number of neighbors can be problematic. The use of several neighbors can improve the estimation accuracy but also diminish the variation, and increase bias as a result of the averaging. According to Tuominen et al. (2003), the optimal number of neighbors is a trade-off between the estimation accuracy and the variation retained in the estimates. Since the estimation accuracy of the plot-wise volumes improved significantly up to three neighbors and the biases remained stable, it was justified to use three nearest neighbors in the estimation of the stand-wise diameter and height-diameter distributions and volumes. Moreover, the use of more neighbors could have a negative impact on the estimation results because of the averaging. The weights for the different data sources were optimized by minimizing the RMSE of the total volume estimate or alternatively the sum of the RMSEs of the species-specific volumes. In the latter case, the spectral data were more influential, which supports the conclusion that the spectral attributes contain information on the tree species composition. Because the target of the estimation was to predict stand-level diameter distributions, and furthermore saw log volume estimates by utilizing the species-specific height-diameter distributions, the optimization could have been done by minimizing, for example, Reynolds' error index of species-specific diameter distributions at the plot-level. However, this alternative is problematic when the plot size is relatively small, and therefore, the number of trees per plot also remains low, so it is difficult to estimate accurate diameter distributions; a few trees in the wrong diameter class can significantly affect the calculation of the error indices.

The stand-level volumes and species-specific diameter distributions were predicted by using the parameters optimized at the plot-level. Although plausible, the results would have been better if the optimization had been carried out at the stand-level. On the other hand, this could have ended in overly optimistic conclusions, since in the practical situation accurate stand-level data (i.e. harvester data collected immediately after the acquisition of the remote sensing material) is still absent nowadays. The stand-wise species-specific volume estimates were at the same level as those achieved by Packalén and Maltamo (2007). However, these numbers should be compared with caution because of the differences in the ground truth reference data; in our case, the ground truth data was collected independently from the plot data and included every tree of the test stand and the stand-level estimates were calculated by summing the estimation results of the cells of the grid laid over the whole stand, whereas Packalén and Maltamo (2007) used the method where the plots inside the test stand represented the stand-level data and the estimates were calculated as the average of the plot-level estimates. Thus, the true values included sampling errors.

The predicted diameter distributions were rather accurate considering the dominant tree species, and we were also able to estimate multi-modal distributions. The estimation of the diameter distributions for the minor species was, nevertheless, not so successful and the method tended to produce overestimates for these. This is probably a consequence of the method's averaging effect and the fact that the plot data was pine dominated, whereas the test stands were spruce dominated.

The estimation of saw log recovery by using predicted height-diameter distributions and the stem data bank was not successful. Even if the harvester measured height-diameter distribution was used, the estimation errors of the saw log proportions were substantial. There are several reasons for this. First of all, the similarity of the height-diameter distributions and tree species proportions does not automatically mean there is similarity between the tree quality attributes. Therefore some auxiliary variables, like forest site class and geographical location, could be useful in explaining the saw log proportion. Secondly, the stem data bank data was not very extensive; thorough examination revealed that three of the test stands were outside the boundaries of the stem data bank's species-specific saw log proportion range.

Similarly, the stem data bank may have been more locally emphasized, although the results improved only slightly when the test stands were included in the reference data. Furthermore, besides the actual quality of the trees, the saw log recovery is dependent on the bucking parameters used; the saw log recovery can differ substantially

if the optimization parameters, such as allowable log lengths and diameters or harvested timber assortments used in bucking, are changed. Finally, information is always lost and errors accumulate when the variables of interest are estimated by using long model chains. In the optimal situation, the saw log recoveries are estimated by searching for the nearest neighbors directly from the stem data bank by using ALS and spectral data predictors. This means that the stands of the stem data bank also need to be laser scanned and photographed in a relatively short time before the cuttings. Such data requires large laser scanning areas and is perhaps not available until ALS based forest inventory is in operational use.

Saw log recovery can also be predicted by using direct ALS based regression models, which was demonstrated by Korhonen et al. (2008). At first, they calculated the factual saw log recoveries for the trees in the modeling data using the saw log reduction factor models presented by Mehtätalo (2002) and then the saw log volume estimation models were formulated based on low density ALS data. The study was established at the Matalansalo test area and the models were tested with the test stands that were also used in this study. The results indicated that direct ALS based models were capable of producing rather accurate theoretical and factual saw wood volume estimates (RMSE of the factual saw log recovery was 18%). However, they did not consider the tree species proportions; on the contrary, the dominant tree species was taken as a priori information and the stands were understood to be single species stands.

Apart from the use of low density ALS data in the canopy height distribution approach, diameter distributions and volume estimates can be produced through individual tree delineation, which was demonstrated by Peuhkurinen et al. (2007), who managed to predict very accurate diameter distribution for mature spruce stands. However, they did not consider the tree quality or species recognition issues, though Holmgren and Persson (2004) have proved that the identification of tree species is possible from ALS data. The main issue which seems to prefer the canopy height distribution approach instead of the individual tree delineation is that the latter requires higher pulse density (4–10 pulses/m^3) to separate individual trees under boreal conditions (Hyyppä and Inkinen 1999, Persson et al. 2002), which incurs more data acquisition expenses.

In comparison with the results achieved with computational methods of pre-harvest measurements, the results presented here are not encouraging. Tommola et al. (1999) reported RMSE values of 14.7 and 11.3 for saw log/pulp wood ratios of pine and spruce stands, respectively. Malinen et al. (2001) achieved saw log ratios similar to our findings but more accurate volume estimates by using harvester measured or, alternatively, forest inventory based data. Furthermore, Malinen (2003) reported an RMSE for saw timber ratio of only 7.1% in spruce stands by utilizing harvester data. However, these numbers do not tell the whole truth. First of all, they used accurately measured stand information, not predicted stand variables. Tommola et al. (1999) had an extensive study material of 716 stands and they used some accurately measured variables, like the size of the logs in the modeling phase; the estimation itself was carried out using cross-validation. In those studies by Malinen et al. (2001) and Malinen (2003) the quality of the trees was not considered, thus, the saw timber ratios were calculated based on the dimensions of the trees and not the actual saw log recoveries. However, the estimation of defects that affect saw log recovery has been found troublesome (Malinen et al. 2007). The problem is more severe in spruce dominated stands, where the risk of butt rot (*heterobasidion)* damage is difficult to predict (Mattila and Nuutinen 2007). These notions do not diminish the fact that the results achieved here are not at a sufficient level, rather they confirm the matter that there is no soft option to predict the quality attributes of forest stands.

In this study, only the first pulse data was utilized in the estimation procedure. Even though it is plausible that laser canopy height distribution of first pulse data provides most of the volume information on the growing stock, other pulse types, i.e. last and intermediate pulses, or then intensity, could contain auxiliary information about forest structure and tree species composition, which may enhance the estimation results.

Here we considered the prediction of the stand characteristics from the wood procurement's point of view. However, we do not see any reason as to

why the proposed estimation method could not be utilized in other areas of interest, too. The method can be directly adapted to different estimation tasks, including classification procedures. Furthermore, because of the multivariate approach of the method, several continuous and discrete variables can be estimated simultaneously. It is only a question of which variables are used in the search for the optimal estimation parameters.

The obvious weakness of the method is that it looses some information contained in the laser canopy height distribution because of the classification of laser pulses into height classes. This is a problem that may occur if low density data and small plot size is used. Another problem is that the Minkowski distance tells us only the magnitude of the difference between two distributions, not where the difference is. Thus two very different distributions can give the same values of Minkowski distance when they are compared with some reference distribution.

Another weakness of the introduced nearest neighbor search method is that it is not a straightforward procedure to implement any other predictor variables than that of the form of distribution in the estimation procedure. For example, by considering the estimation of the saw log recovery by utilizing the stem data bank, the forest site class or geographical distance to the reference stand besides of the height-diameter distributions could improve the estimation accuracy. In other words, the problem is how to measure distance between these new variables of non-distributional form so that it is compatible with the Minkowski distance of order one applied to classified distributions.

References

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

Bellman, R.E. 1961. Adaptive control processes. Princeton University Press, Princeton, NJ.

Gobakken, T. & Næsset, E. 2004. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. Scandinavian Journal of Forest Research 19: 529–542.

Holmgren, J. 2004. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. Scandinavian Journal of Forest Research 09: 543–553.

— & Persson, Å. 2004. Identifying species of individual trees using airborne laser scanning. Remote Sensing of Environment 90: 415–423.

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

Jensen, J.L.R., Humes, K.S., Conner, T., Williams, C.J. & DeGroot, J. 2006. Estimation of biophysical characteristics for highly variable mixed- conifer stands using small-footprint lidar. Canadian Journal of Forest Research 36: 1129–1138.

Korhonen, L., Peuhkurinen, J., Malinen, J., Suvanto, A., Maltamo, M., Packalén, P. & Kangas, J. 2008. The use of airborne laser scanning to estimate sawlog volumes. Forestry 81(4): 499–510.

Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae 108. The Finnish Forest Research Institute, Helsinki, Finland. 74 p.

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

Lim, K., Treitz, P., Baldwin, K., Morrison, I. & Green J. 2003. Lidar remote sensing of biophysical properties of tolerant northern hardwood forests. Canadian Journal of Remote Sensing 29: 648–678.

Malinen, J. 2003. Locally adaptable non-parametric methods for estimating stand characteristics for wood-procurement planning. Silva Fennica 37(1): 109–120.

— , Maltamo, M. & Harstela, P. 2001. Application of most similar neighbor inference for estimating marked stand characteristics using harvester and inventory generated stem databases. International Journal of Forest Engineering 12(2): 33–41.

— , Kilpeläinen, H., Piira, T., Redsven, V., Wall, T. & Nuutinen, T. 2007. Comparing model-based approaches with bucking simulation based approach in the prediction of timber assortment recovery. Forestry 80(3): 309–321.

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.

— , Suvanto, A. & Packalén, P. 2007. Comparison of basal area and stem frequency diameter distribution modelling using airborne laser scanner data and calibration estimation. Forest Ecology and Management 247: 26–34.

Mattila, U. & Nuutinen, T. 2007 Assessing the incidence of butt rot in Norway spruce in southern Finland. Silva Fennica 41(1): 29–43.

Means, J.E., Acker, S.A., Brandon, J.F., Renslow, M., Emerson, L. & Hendrix, C.J. 2000. Predicting forest stand characteristics with airborne scanning lidar. Photogrammetric Engineering & Remote Sensing 66: 1367–1371.

Mehtätalo, L. 2002. Valtakunnalliset puukohtaiset tukkivähennysmallit männylle, kuuselle, koivulle ja haavalle. Metsätieteen aikakauskirja 4/2002: 575–591. [In Finnish].

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

Newton, P.F., Lei, Y. & Zhang, S.Y. 2004. A parameter recovery model for estimating black spruce diameter distributions within the context of a stand density management diagram. The Forestry Chronicle 80: 349–358.

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

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

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

Peuhkurinen, J., Maltamo, M., Malinen, J., Pitkänen, J. & Packalén, P. 2007. Pre-harvest measurement of marked stands using airborne laser scanning. Forest Science 53(6): 653–661.

Reynolds, M.R., Jr., Burk, T.E. & Huang, W.-C. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. Forest Science 34(2): 373–399.

Siipilehto, J. 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. Silva Fennica 33: 281–301

StanForD. 2006. Standard for Forest Data and communications. 4 April 2006. [Internet site]. Skogsforsk. Available at: http://www.skogforsk.se/templates/sf_ProjectStartPage____3967.aspx?sm=2&cri=17833&lipm=1. [Cited 7 May 2007].

Tommola, M., Tynkkynen, M., Lemmetty, J., Harstela, P. & Sikanen, L. 1999. Estimating the characteristics of a marked stand using k-nearest-neighbour regression. International Journal of Forest Engineering 10(2): 75–81.

Tuominen, S., Fish, S. & Poso, S. 2003. Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. Canadian Journal of Forest Research 33: 624–634.

— & Pekkarinen, A. 2004. Local radiometric correction of digital aerial photographs for multi source forest inventory. Remote Sensing of Environment 89: 72–82.

Veltheim, T. 1987. Pituusmallit männylle, kuuselle ja koivulle. M.Sc. thesis, Faculty of Agriculture and Forestry, University of Helsinki, Finland. [In Finnish].

Total of 32 references