The Use of Quantile Trees in the Prediction of the Diameter Distribution of a Stand

Lauri Mehtätalo, Matti Maltamo and Annika Kangas

Mehtätalo, L., Maltamo, M. & Kangas, A. 2006. The use of quantile trees in the prediction of the diameter distribution of a stand. Silva Fennica 40(3): 501–516.

This study deals with the prediction of the basal area diameter distribution of a stand without using a complete sample of diameters from the target stand. Traditionally, this problem has been solved by either the parameter recovery method or the parameter prediction method. This study uses the parameter prediction method and the percentile based diameter distribution with a recent development that makes it possible to improve these predictions by using sample order statistics. A sample order statistic is a tree whose diameter and rank at the plot are known, and is referred to in this paper as a quantile tree. This study tested 13 different strategies for selection of the quantile trees from among the trees of horizontal point sample plots, and compared them with respect to RMSE and the bias of four criterion variables in a dataset of 512 stands. The sample minimum was found to be the most promising alternative with respect to RMSE, even though it introduced a rather large amount of bias in the criterion variables. Other good and less biased alternatives are the second and third smallest trees and the tree closest to the plot centre. The use of minimum is recommended for practical inventories because its rank is probably easiest to determine correctly in the field.

Authors' addresses *Mehtätalo*, Yale School of Forestry and Environmental Studies, 205 Prospect Street, New Haven, CT 06511, USA; *Maltamo*, University of Joensuu, Faculty of Forestry, P.O. Box 111, FI-80101 Joensuu, Finland; *Kangas*, University of Helsinki, Department of Forest Resources Management, P.O.Box 27, FI-00014 University of Helsinki, Finland
E-mail lauri.mehtatalo@metla.fi

1 Introduction

The development of the forestry applications of horizontal point sampling, (i.e., relascope sampling, angle count sampling) has had a considerable effect on small area forest inventory in Finland. The measurement of basal area led, first, to the use of so-called relascope tables (stand volume models) and second, since the 1970s, basal area diameter distribution models were developed as the basis of the calculation of stand volume in a small area forest inventory system (Nyyssönen 1954, Kilkki and Siitonen 1975, Päivinen 1980, Kilkki et al. 1989). With a basal area diameter distribution, the frequency of a diameter class was expressed in terms of basal area instead of number of stems, being consistent with the horizontal point sampling, where each sampled tree corresponded to a fixed basal area instead of a fixed number of stems. The parameter prediction approach was used and predictor variables, such as basal area and basal area weighted median diameter (DGM), were assessed at the stand level in the field. In fact, the use of basal area diameter distributions has been a Finnish speciality and not until in the end of the 1990s had it been studied elsewhere (Gove and Patil 1998). Correspondingly, in the NFI of Finland a horizontal point sample plot (HPS plot) based inventory system was established in the beginning of 1960s (NFI4).

The advantage of the basal area weighted sampling of trees is clear when compared to the sampling without weighting and it leads to improved accuracy of both basal area and volume estimates with a given number of measured trees (Sukvong et al. 1971, Matérn 1972). The situation is also the same for diameter distribution models. For example, in the study by Maltamo and Mabvurira (1999), stem frequency and basal area diameter distributions led, in the same study material, to the relative RMSE's of volume of 10.56 % and 2.56 %, respectively. However, the difference resulted mainly from the variable used in the scaling of the distribution (stem number in the former and basal area in the latter case), not from the weighting itself.

The disadvantage of horizontal point sampling is imprecise estimates of small diameter class frequencies. The angle count sampling, as such, provides unbiased estimates of stand character-istics (Bitterlich 1984). However, according to Vuokila (1959), the estimates of trees smaller than approximately 10 cm are imprecise, i.e., have considerable variance. It also means that the description of stand structure, measured with total number of stems or energy wood volume, is imprecise, as the number of small trees has a large effect on them. Furthermore, predicting stand development and thinning removals using such highly uncertain information on small classes is imprecise and may even be unrealistic (Maltamo and Kangas 1998, Kangas and Maltamo 2003).

To improve the ability of a basal area diameter distribution to describe stand structure characteristics other than volume, Siipilehto (1999) proposed an additional measurement of stem number be used in the parameter prediction. Apparently, it considerably improved the obtained estimate of stem number. However, it also improved the estimate of volume. Another approach is to calibrate the predicted distribution with stem number by means of calibration estimation (Kangas and Maltamo 2000a, see also Mehtätalo 2004a) in order to obtain estimates which are compatible with both measured basal area and stem number. More recently, a parameter recovery approach based on Weibull distribution has also been developed for this situation (Mehtätalo and Nyblom, forthcoming). The problem with the above-mentioned studies is that the stem number used in the modelling is known without error, whereas with the practical applications it is measured with error. In practical work, stem number is often visually assessed, and practical studies have shown that the accuracy of such measurement is very low, including RMSE's as high as 80 % (Kangas et al. 2002). If such estimates are used in the calculation of stand variables the advantage of using stem number is only marginal (Haara and Korhonen 2004), even if the measurement errors are accounted for (Mehtätalo and Kangas 2005). However, some recent studies have shown that practical measurement of stem number may be highly useful in special situations, e.g., when predicting the amount of energy wood (Kaartinen 2005). In the study of Kaartinen (2005), RMSE's of energy wood assortments were decreased to less than half when calibrated with the field measured stem number of energy wood sized trees. However, the dataset of this study was quite small.

One special attempt to improve the description of stand structure by using basal area diameter distribution and without measuring the number of stems was presented by Maltamo et al. (2003), using a most similar neighbour (MSN) model. In the MSN approach, the prediction is based on one or a few stands which are selected from among a large set of measured stands. The selection is based on a similarity measure, which is obtained through a canonical correlation analysis. Maltamo et al. (2003) placed as much weight as possible upon the accurate prediction of stem number by using only characteristics of the stand structure as dependent variables in the canonical correlation analysis. As a result, the RMSE and bias of stem number decreased without any decrease in the accuracy of the estimates of stand volume. However, the practical application of this kind of model is questionable for many situations, since the availability of modelling data (or previously measured data) is necessary for the end user of the model.

Recently, Mehtätalo (2005) developed an approach that utilizes sample trees whose diameter and rank on the plot are known, in order to improve the prediction of the diameter distribution based on stand characteristics. These trees are called sample order statistics or quantile trees. Special cases of sample order statistics are sample minimum, maximum and median diameters, but any sample order statistics can be used as well, for example the diameter of the 4th smallest tree of a sample of 14 trees. Sample order statistics are interpreted as measured percentiles of the stand. In the quantile tree approach, a fixed set of diameter percentiles is first predicted by using the parameter prediction approach. These predictions are subsequently improved using the measured percentiles, i.e., the quantile trees. The idea is illustrated in Fig. 1 with one measured quantile tree, but the method can be generalized to several trees, which can be collected either from one or several plots. Mehtätalo and Kangas (2005) used this approach to find an optimal measurement strategy for a single stand with respect to the accuracy of the total and sawtimber volume. However, no attention was paid to the small trees, and the quantile trees were selected randomly from among the sample trees of the plot. However, it might be advantageous to measure the sample
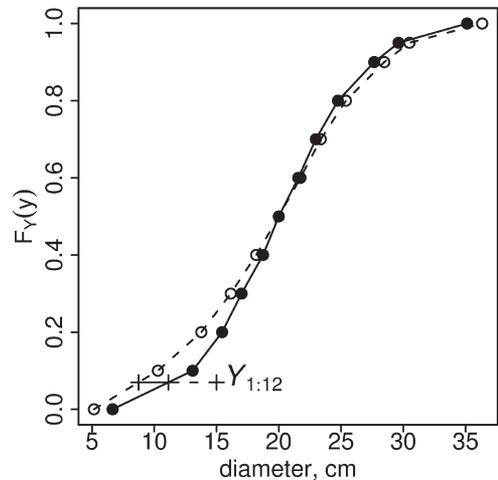

Fig. 1. The idea of the quantile tree approach. The dashed line connecting the open circles shows the cumulative distribution function based on predicting 12 diameters of predefined percentages using *a priori* estimated PPM models. The minimum diameter in a sample of 12 trees has been observed to be 15 cm. Based on the rank 1:12 and the predicted c.d.f., it is interpreted as a measurement of the 7th diameter percentile and plotted at point (15, 0.07). The horizontal difference between the observed diameter and the PPM prediction is an observed residual (horizontal line), which can be partitioned to a predicted stand effect (solid part) and a random residual (dashed part), according to the error variances of our PPM models and the sampling error in the quantile tree. Thus, the solid part is our prediction of how much the 7th percentile in this particular stand deviates from the PPM prediction. As the stand effect is correlated with diameters at other percentages, we can also predict stand effects for them to obtain an improved prediction for the diameter distribution (the solid line connecting the black circles).

order statistics in that part of the diameter distribution which is originally most inaccurate or which is important to be estimated accurately.

The aim of this study is to compare different strategies to choose the quantile trees to be measured. Special attention is paid to the description of small diameter classes and especially to the accuracy of stem number prediction in evaluat-

ing the strategies. In addition, the accuracy of obtained results is examined in relation to stand characteristics and the practical measurements of quantile trees are discussed.

2 Material

This study used a dataset collected from permanent sample plots by the Finnish Forest Research Institute between 1976 and 1992 (Gustafsen et al. 1988). The stands were a sample from those stands of the 6th and 7th national forest inventories that were located on mineral soils and on forest land. In each stand, 3 permanent circular plots were established. The plot size varied between stands and was determined so that at least 120 sample trees per stand were measured in southern Finland and 100 sample trees per stand in northern Finland. The tree species and the diameter at breast height were recorded from all trees belonging to the sample plot. This study used Scots pine trees from 512 stands. The same data were also used by Kangas and Maltamo (2000c).

For comparisons between predicted and actual diameter distributions of the stands, the assumed true total volume, energy wood volume and sawtimber volume were calculated for each stand using the volume functions of Laasasenaho (1982), based on tree diameter only. The energy wood consists of trees which will be used for energy production. The diameter of those trees varies between 4 and 10 cm (see Kaartinen 2005) and sawtimber trees were those with a diameter

of 16 cm or more. In addition, the true number of stems (1/ha), DGM, age and site fertility class were known from each stand. In the subsequent analysis, these estimates are assumed to be true values for the stand, even though they are estimated from the three circular plots. Table 1 shows a summary of the dataset.

In order to be able to simulate quantile trees, HPS plots were generated at the center of each circular sample plot of the data. When establishing HPS plots on small circular plots, one must take into account that some of the trees that should belong to the HPS plot may be outside the circular plot. Assuming that the diameters of trees around the circular plot of radius r do not exceed the maximum diameter of the plot, D_{max}, using a basal area factor q fulfilling the condition

$$q \geq \left(50\, D_{max}/r\right)^2$$

results in that all trees of the HPS plot are within the circular plot. As basal area factors smaller than one are unrealistic in practical inventories, the factor on each plot was determined as

$$q = \max\left[1, \left(50\, D_{max}/r\right)^2\right]$$

i.e. as small a value of the factor as possible was used but not values below 1. On some plots, this resulted in quite large basal area factors, the maximum value being 7.35. However, in order to compare the results to those of Kangas and Maltamo (2000c), these stands were retained in the data.

Table 1. Minimum, mean and maximum values of some variables in the dataset.

	Min	Mean	Max
DGM	5	15.6	35
Basal area	1.14	12.1	32.74
Stand age	15	67.2	183
Number of stems per ha	86.6	1193	4237
Energy wood volume per ha	0	8.4	43.5
Sawtimber volume per ha	0	38.7	181.4
Total volume per ha	4.2	66.3	201.4
Total number of measured trees per stand	15	86	154
Number of trees per HPS plot	1	10.2	24

3 Methods

3.1 Prediction of Diameter Distribution

Let us assume that $Y_1, Y_2, ..., Y_n$ is a random sample of tree diameters from a forest stand, where the sampling probability is proportional to the squared diameter of the tree. Assuming that trees are randomly located in the stand and no spatial autocorrelation exists in diameters, a sample of this kind is obtained from a HPS plot. The rth smallest observation of the sample, denoted by $Y_{r:n}$, is a random variable that is called the rth sample order statistic. For example, minimum, median and maximum diameters of a sample are special cases of sample order statistics.

A sample order statistic can be interpreted as an observed $100p$'th percentile of the diameter distribution F_Y. If the expected value of the order statistic is known, p' can be calculated by writing it into the distribution function as

$$p' = F_Y \Big[E \big(Y_{r:n} \big) \Big]$$

Thus, $Y_{r:n}$ can be interpreted as an unbiased measurement of the $100p$'th percentile of the basal area weighted diameter distribution. However, the measurement includes an error resulting from that it was measured from a sample. In order to calculate the expectation and variance of a sample order statistic, we need to know its probability distribution. If F_Y is the basal area weighted diameter distribution of the stand, the density of $Y_{r:n}$ is (Casella and Berger 2002: 229, Reiss 1989: 21)

$$
\begin{aligned}
&f_{r:n}(y) \\
&= \frac{n!}{(r-1)!(n-r)!} f_Y(y) \big[F_Y(y) \big]^{r-1} \big[1 - F_Y(y) \big]^{n-r}
\end{aligned}
\qquad (1)
$$

which can be used to calculate $E(Y_{r:n})$ and var($Y_{r:n}$), given that the diameter distribution $F_Y(y)$ is known (see Mehtätalo 2004b, 2005).

Borders et al. (1987) suggested a percentile-based approach in the description of stand structure. In their approach, the diameter distribution of a stand is described using percentiles $d_1, d_2, ..., d_k$ that correspond to fixed values $p_1, p_2, ... p_k$ of the cumulative diameter distribution and a continuous distribution function is approximated by interpolating between these percentiles. With the use of linear interpolation, the percentile based diameter distribution can be expressed as

$$
F_Y(y) \approx
\begin{cases}
0 & y < d_1 \\
a_i + b_i y & d_i \le y < d_{i+1}, i = 1, ..., k-1 \\
1 & y \ge d_k
\end{cases}
$$

where $b_i = (p_{i+1} - p_i)/(d_{i+1} - d_i)$ and $a_i = p_i - b_i d_i$. The percentile-based distribution can be interpreted either as a distribution-free method, where the distribution is obtained through interpolation between percentiles, which are interpreted as stand variables (Kangas and Maltamo 2000b). On the other hand, the distribution can be interpreted as a mixture of uniform distributions (Mehtätalo 2004b), which is a parametric distribution. The parameters of this distribution are the percentiles, which can be predicted, for example, by using the parameter prediction method (PPM).

By using PPM with the percentile based diameter distribution, the percentiles $d_1, d_2, ..., d_k$ of a target stand are predicted using field measurements of the predictors. Assuming that the model is correct and ignoring the estimation errors of the parameters, this approach offers expectations of the percentiles conditional to the values of stand characteristics.

As the observed sample order statistic can be interpreted as a measurement of the p'th percentile, it includes additional information about the percentiles of the stand. Mehtätalo (2005) showed how these two sources of information can be combined using linear prediction theory. The idea of the method is presented in Fig. 1. For a formal presentation, let us define the vectors $\mathbf{d} = (d_1, d_2, ..., d_k)'$ and $\mathbf{p} = (p_1, p_2, ..., p_k)'$. The PPM model of the percentiles can be written as

$$\mathbf{d} = E\big(\mathbf{d} \,|\, \mathbf{x} \big) + \mathbf{e} \qquad (2)$$

where $\mathbf{d}$ includes the $100\mathbf{p}$th percentiles of the stand, $E(\mathbf{d}|\mathbf{x})$ includes the PPM predictions of percentiles based on the vector of stand variables, $\mathbf{x}$, and $\mathbf{e}$ is the vector of stand effects (i.e. residuals) with expectation $\mathbf{0}$ and estimated variance-covariance matrix $\mathbf{D}$. In the traditional PPM approach, predictions $E(\mathbf{d}|\mathbf{x})$ are used as the predicted parameters of the distribution. In

our approach, we will also predict the vector of stand effects, $\mathbf{e}$.

For simplicity, the prediction method is described in the case of one measured sample order statistic per stand only. However, the method can be generalized to the case of several sample order statistics per stand, which can be measured either from one or several plots. When the measurements are collected from several plots, each sample plot is treated as a separate sample, and the ranks are determined only among the trees of that plot, not among all trees of the stand. If several quantile trees are measured from same plot, the correlation of sample order statistics from the same plots needs to be accounted for. For prediction of stand effects in this case, see Mehtätalo (2005).

Let us first assume that one of the predefined percentiles was measured from a sample plot, i.e. p' equals exactly the mth element of $\mathbf{p}$ as $p' = p_m$. In addition to the terms corresponding to those of model (2), the measured percentile includes a measurement error term, ε and the measured percentile can be written as:

$$Y_{r:n} = E\left(Y_{r:n}\,|\,\mathbf{x}\right) + e_m + \varepsilon$$

where $E(Y_{r:n}|\mathbf{x}) = E(d_m|\mathbf{x})$. Terms e_m and ε are the stand effect and measurement error, respectively. From a purely statistical point of view, they are just mutually independent random variables with expectations of zero and variances of σ_e^2 and $\sigma_\varepsilon^2 = \mathrm{var}(Y_{r:n}|\mathbf{x},\mathbf{e})$, respectively. Term σ_ε^2 is the error variance of the measured quantile tree. It is the variance of distribution (1), where the diameter distribution based on true stand effects of (2) is used as F_Y. As it is not known, we rely on predictions through an iterative approach described below and ignore the prediction errors. The expectations, variances and covariances of unobserved vector $\mathbf{d}$ and observed scalar $Y_{r:n}$ can be written as

$$
\begin{bmatrix} \mathbf{d}\,|\,\mathbf{x} \\ Y_{r:n}\,|\,\mathbf{x} \end{bmatrix} \sim \left(\begin{bmatrix} E\left(\mathbf{d}\,|\,\mathbf{x}\right) \\ E\left(Y_{r:n}\,|\,\mathbf{x}\right) \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{c} \\ \mathbf{c}' & \sigma_e^2 + \sigma_\varepsilon^2 \end{bmatrix} \right)
$$

where $\mathbf{c}$ is the mth row and σ_e^2 is the mth diagonal element of $\mathbf{D}$. The empirical best linear predictor of $\mathbf{d}$ is

$$\hat{\mathbf{d}} = E\left(\mathbf{d}\,|\,\mathbf{x}\right) + \frac{Y_{r:n} - E\left(Y_{r:n}\,|\,\mathbf{x}\right)}{\sigma_e^2 + \sigma_\varepsilon^2}\,\mathbf{c}$$

and the prediction variance is

$$\mathrm{var}\left(\hat{\mathbf{d}} - \mathbf{d}\right) = \mathbf{D} - \frac{1}{\sigma_e^2 + \sigma_\varepsilon^2}\,\mathbf{c}\mathbf{c}' \qquad (3)$$

However, as p' never equals any element of $\mathbf{p}$ exactly, interpolations are needed in the calculation of $E(Y_{r:n}|\mathbf{x})$, $\mathbf{c}$ and σ_e^2 (see Mehtätalo 2004). The interpolations are analogous to those needed when a stem curve based on fixed polar coordinates is predicted with the use of diameter measurements from arbitrary heights (Lappi 1986).

As the true diameter distribution F_Y would be already needed in the calculation of $E(Y_{r:n})$, $\mathrm{var}(Y_{r:n})$ and p' (Eq. 1) the prediction needs to be carried out iteratively. At the first iteration, the diameter distribution based on predicted percentiles $E(\mathbf{d}|\mathbf{x})$ is used as F_Y to calculate a prediction of the first step. The obtained prediction is used as F_Y and the prediction is carried out again. This is iterated until the iteration does not have a remarkable effect on the predictions. However, sometimes the prediction does not converge or the obtained set of percentiles is not monotonic after some iteration step. In these cases, either the PPM predictions $E(\mathbf{d}|\mathbf{x})$ or a prediction based on a reduced set of measurements can be used (Mehtätalo and Kangas 2005). In this study, the former approach was selected. The percentage of unsuccessful iterations varied from 1.8% with one sample tree per stand to 7.4% with six sample trees per stand. About 8% of the failures resulted from the iteration not being converged within 50 steps and the rest of the 92% resulted from the distribution not being monotonic after some iteration step.

In this study, the percentile models of Maltamo and Kangas (2000b) were used for prediction. These models predict the 0th,10th,… 80th,90th,95th and 100th percentiles of the diameter distribution by using basal area, basal area median diameter (i.e. the 50th percentile of the basal area weighted diameter distribution), stand age and site fertility class as predictors. The estimated variance-covariance matrix of stand effects, $\mathbf{D}$, is published in Mehtätalo (2004a).

All diameters, including the predicted percentiles and measured quantile trees, are used in the logarithmic scale until final predictions are obtained. Before transformation of these predictions to the arithmetic scale, half of the prediction variance, obtained from the diagonal of (3) is added to the predictions. To obtain a continuous distribution function between the final predictions, Späth's rational spline was used (Späth 1974, Lether 1984, see Kangas and Maltamo 2000c) using values 30 and 25 for the tautness parameters. This spline was used instead of linear interpolation because it resulted in a more accurate estimation of the number of stems and total volume and made it possible to compare our results against the study of Kangas and Maltamo (2000c).

3.2 Strategies for Selecting Quantile Trees

A total of 13 strategies were used in selecting the sample order statistics from the three HPS plots available from each stand. Each strategy in Table 2 describes how measurements from a single plot are selected. The selection rule was based either on the distance between the tree location and the plot center or on the rank of the tree among the trees of the sample plot. We assumed no more than six quantile trees per stand would be measured. Thus, those strategies requiring one or two trees per plot to be measured were carried out by using 1, 2 or 3 plots, resulting in 2, 4 or 6 trees per stand, and the strategies requiring three trees per plot were carried out by using one or two plots, resulting in either 3 or 6 trees per stand to be measured. With a given strategy and number of plots to be measured, each measurement strategy was replicated as many times as possible using the three plots available from each stand. Thus, for example, strategy 1S with one plot per stand was replicated three times but strategy 1S with three plots could be replicated only once. The strategies that used randomly selected quantile trees were replicated 50 times. In a few stands, the number of trees on the HPS plot was smaller than the strategy would have required (e.g., two trees on a plot where the strategy would require three trees to be measured). In these cases, all the trees of the plot were used.

3.3 Comparison of Predictions

The predicted diameter distributions were transformed to a stand table using one cm diameter classes and the predicted number of stems per ha, as well as predicted energy wood, saw timber and total volumes per ha were calculated in a similar manner to that which was used to calculate the assumed true values. For each of these criteria, the bias and relative root mean squared error were calculated using equations

Table 2. The strategies used.

Name	Description	Simulated numbers of sample trees, and corresponding numbers of replications per stand in parentheses
1C	The tree closest to the plot center	1 (3), 2 (3), 3 (1)
1S	The smallest tree of the plot	1 (3), 2 (3), 3 (1)
2S	The 2nd smallest tree of the plot	1 (3), 2 (3), 3 (1)
3S	The 3rd smallest tree of the plot	1 (3), 2 (3), 3 (1)
12C	The 1st and 2nd closest trees to the plot center.	2 (3), 4 (3), 6 (1)
12S	The 1st and 2nd smallest trees of the plot	2 (3), 4 (3), 6 (1)
13S	The 1st and 3rd smallest trees of the plot	2 (3), 4 (3), 6 (1)
23S	The 2nd and 3rd smallest trees of the plot	2 (3), 4 (3), 6 (1)
123C	Three closest trees to the plot center	3 (3), 6 (3)
123S	Three smallest trees of the plot	3 (3), 6 (3)
Sa	The tree closest to saw timber limit	1 (3), 2 (3), 3 (1)
La	The largest tree of the plot	1 (3), 2 (3), 3 (1)
Ra	Randomly selected trees from the three plots	1 (50) ,2 (50), 3 (50), 4 (50), 5 (50), 6 (50)

$$\text{bias} = \frac{\sum\limits_{i=1}^{n}\left(X_i - \hat{X}_i\right)}{n}$$

and

$$\text{RMSE}(\%) = 100 \sqrt{\frac{\sum\limits_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2}{n}} \Big/ \bar{X}$$

where $\hat{X}_i$ and X_i are the predicted and (assumed) true values of the criterion variable in stand i, respectively, and $\bar{X}$ is the mean of the true values over all stands.

4 Results

4.1 Overall Performance of the Strategies

In all the strategies, the RMSE of the number of stems decreased steadily as the number of quantile trees increased (Fig. 2). The absolute bias in the number of stems increased, but this increase did not override the effect of decreasing the variance. Selection of the closest or smallest trees of the plots (Strategies 1C and 1S) resulted in the lowest RMSE of the number of stems and of these two strategies, the former clearly gave a lower absolute bias. The accuracy of the energy wood prediction could also be clearly improved with the use of quantile trees. For example, measuring the second smallest tree from three plots reduced the relative RMSE of energy wood volume from 46.5% to 41%. Selection from among the smallest trees of the plots (Strategies 1S, 2S and 3S) resulted in the lowest RMSE and also decreased the absolute bias of energy wood volume.

With respect to the accuracy of volume prediction, the situation was not as clear. The volume prediction without quantile trees was already very accurate and the use of quantile trees resulted in only minor changes in the accuracy. In particular, the use of only one plot per stand usually increased the RMSE of volume regardless of the number of sample trees per plot and improvements occurred when the number of sample plots increased from 1 to 2 or 3. The situation is similar also with sawtimber volume and seems to result

mainly from an increase in the absolute bias when the number of plots is increased from 0 to 1. The reason for this phenomenon may be a trend or large scale heterogeneity in the total volume within a stand, which causes bias in estimates when measurements from one plot only are used. Selection from among the smallest or largest trees of the plots (Strategies 1S, 2S, 3S and La) resulted in the lowest RMSEs of volume and sawtimber volume, while strategies 1C, 12C, 123C, Sa, La and Ra resulted in the lowest absolute biases.

Kangas and Maltamo (2000c) used the same dataset (called INKA data in their study) and same criterion variables when they compared the methods of percentiles, Weibull, and k-nearest neighbours (k-nn) in the prediction of diameter distribution. They found the k-nn approach to be the best compromise between RMSE:s of the number of stems, volume, and sawtimber volume (Kangas and Maltamo 2000c). The relative RMSE:s were 23.20, 2.04 and 13.52 per cent and biases –23.20, –0.10 and –0.45 trees per ha, respectively. In this study, at least two quantile trees were needed to result in a corresponding accuracy of the number of stems (strategies 1C and 1S). An equally low RMSE of volume could be obtained only by using two closest trees from three sample plots and the RMSE of the sawtimber volume was lower only when the largest trees were used as quantile trees.

In relation to the RMSE's of all four criteria, the strategy using smallest trees of the plots (1S) performs best but it causes rather large increases in absolute biases. The lowest biases result from using strategies where quantile trees are not selected systematically from the extremes of the sample (Strategies 1C, 12C, 123C, Sa and Ra). However, the decrease of strategy 1S in the RMSE criterion is so high that it overrides the effect in the bias. Thus, if low absolute bias is not an issue, measuring the smallest tree of an angle count sample could be a good strategy in practical inventories.

4.2 Dependency of the Performance on Stand Properties

In the studies of Kangas and Maltamo (2002) and Mehtätalo and Kangas (2005), the optimal
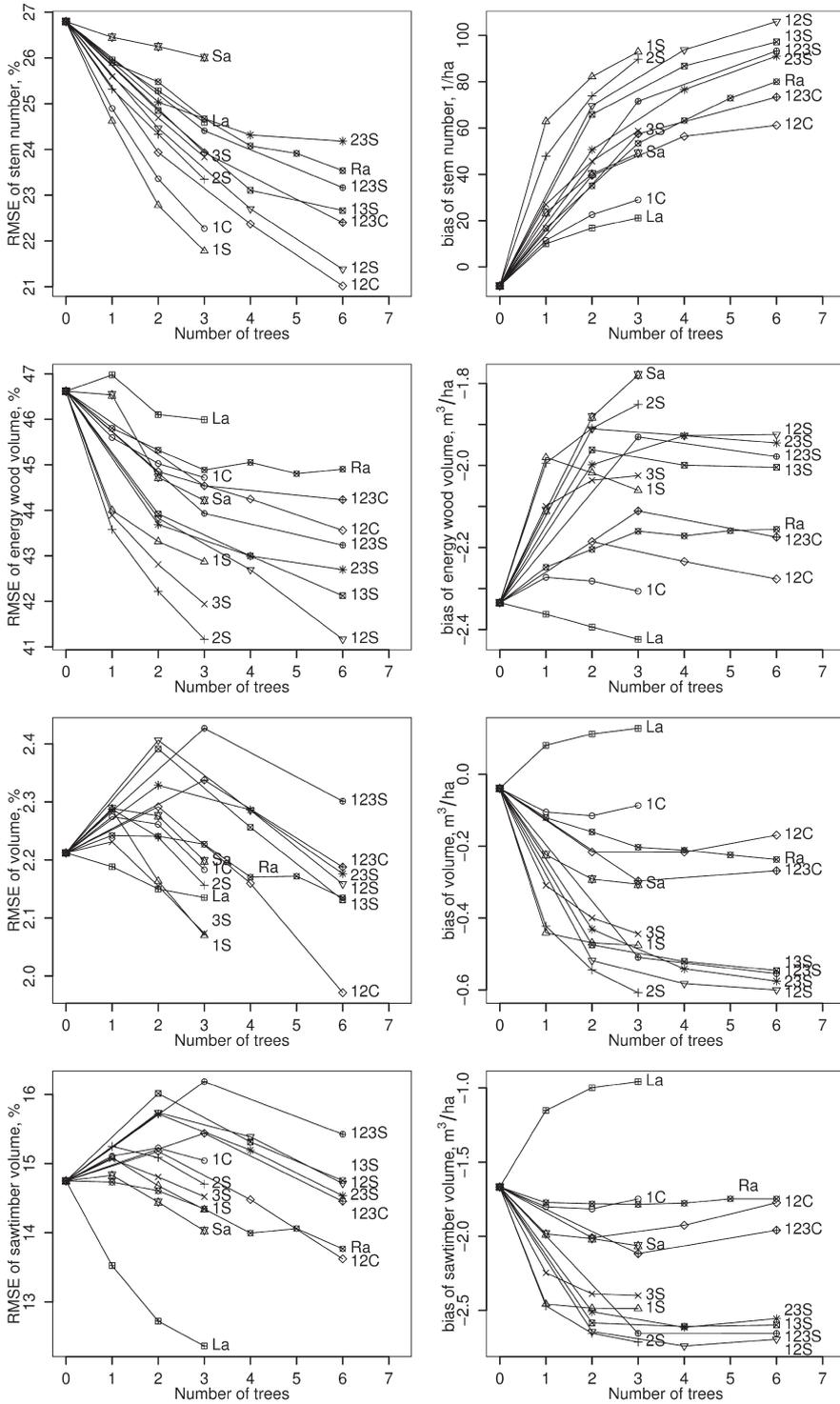
Fig. 2. The effect of increasing the number of quantile trees on the RMSE and bias of the number of stems, volume, sawtimber volume and energy wood volume using different strategies in selecting the quantile trees.

measurement strategy of a stand was found to depend on stand characteristics. Thus, also in this study, RMSEs of criterion variables were plotted against stand characteristics. This part of this study only reports on the results of strategies where the number of sample trees per plot was one (see Table 2). Lowess smoothers were fitted to the data of root mean squared errors using R function 'loess' with value 1 of the smoothing parameter. The value of the smoother using no quantile trees was subtracted from the obtained values to make the plot more easily readable. The strategy that gives the lowest RMSE varies

greatly with respect to stand characteristics, and in certain situations it seems profitable not to use quantile trees at all (Fig. 3).

As basal area and DGM are correlated, consideration of performance of strategies against DGM and basal area simultaneously gives a more realistic view of the performance than separate consideration of Fig. 3. Thus, the data was classified into six classes according to basal area and DGM. Fig. 3 and corresponding figures about the other criterion variables were used to determine appropriate class limits. The strategies were ranked in each class according to the RMSEs and
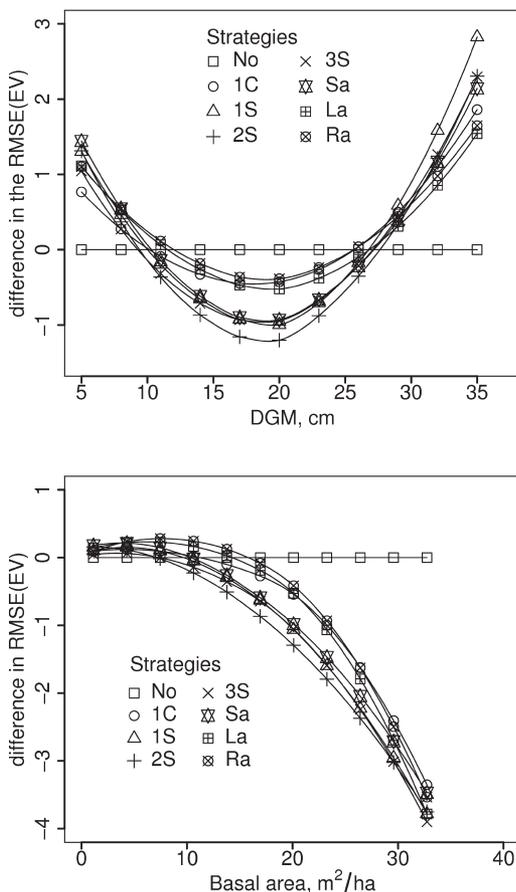
Fig. 3. Smoothed change in the RMSE of energy wood volume caused by the use of quantile trees according to strategies given in the figures with respect to DGM and basal area.
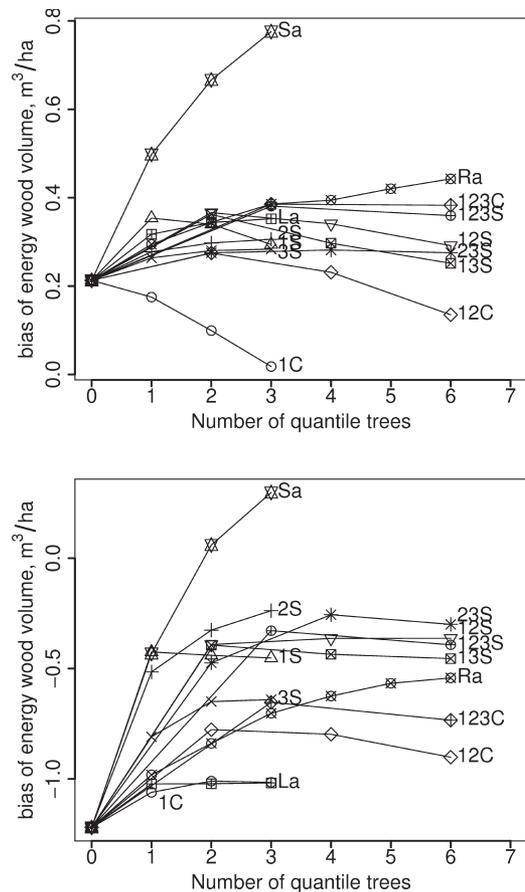
Fig. 4. The effect of increasing the number of quantile trees on the bias of energy wood volume using different strategies in strata $11 < DGM < 18$ cm, basal area > 12 m^2/ha (upper graph) and $DGM > 18$, basal area > 12 m^2/ha.

standard deviations of the criterion variables.

When one compares the RMSE and standard deviation of all criteria (Tables 3 and 4), one can see that the standard deviation behaves more logically than the RMSE, the reason being that quantile trees selected from the tails of the distribution have a remarkable effect on the biases. In such a subpopulation of stands where the percentile model is biased, a strategy that causes an opposite bias cancels the bias and performs very well, but in another subpopulation, the same strategy may perform very poorly because the biases are of the same sign. An example of this is strategy Sa in Fig. 4. This is why both RMSE and standard deviation are considered below.

The results of Tables 3 and 4 are not very easy to interpret because a large part of the variation is random. This effect is especially strong in the case of criterion variables where the differences between strategies are small. However, statistical tests that would have shown the significance of the differences could not be easily carried out because of interdependencies between observations of the data. Thus, in order to separate the essential differences from the unimportant ones, total order was calculated for each cell of Tables 3 and 4 as follows. First, to obtain a relative performance value for each strategy with respect to each criterion in a given class, each RMSE (or standard deviation) value was divided by the RMSE (or standard deviation) of the best strategy in that class. Secondly, these values were summed over the criterion variables in order to obtain a total performance for each strategy in a given class. Finally, the strategies were ordered with respect to this sum so as to obtain the total order.

The total order of strategies shows that the strategy that uses the smallest tree is among the two best strategies in 8 classes out of 12 classes in tables 3 and 4, and the performance seems not to depend on stand characteristics. Other good strategies include those that use the second and third smallest trees. Also the strategy using the closest tree performs well, except for the class of largest DGM and basal area, where this strategy is surprisingly the worst one. The strategy that uses the largest trees performs well in stands with small DGM.

Measuring the smallest or closest trees of a plot as quantile trees results in the largest improvements in the accuracy of the number of stems, especially in stands with a small basal area. The energy wood volume also seems to be most accurately estimated by using the smallest, second smallest or third smallest trees as quantile trees, even though in the class of smallest DGM and basal area the strategy of not using quantile trees performs best with respect to RMSE. With respect to the RMSE of the total volume and sawtimber volume, the strategy that uses the largest trees as quantile trees is among the best strategies, regardless of the stand properties but the strategies which select the smallest trees also perform well, especially in stands with large DGM. As the differences between strategies with respect to these two criteria are small, they do not contribute very much to the total order.

5   Discussion

This study examined how the choice of quantile trees affects the accuracy of stand characteristics derived from stand measurements and predicted diameter distribution. This was done by comparing different strategies to measure additional trees from different parts of the diameter distribution. The basic assumption was that the diameter of the basal area median tree (50 % percentile) was known and measured. Special attention was paid to those forest characteristics that depend on an accurate description of small diameter classes, such as the number of stems and the energy wood volume. However, stand total and sawtimber volumes were considered, as well. The results were slightly contradictory, but the main conclusion was that the best alternative is to carry out additional measurements from among the smallest trees of the HPS plot.

In general, the most accurate strategies were the ones where either some of the smallest trees of the HPS plot or the closest tree to the centre of the plot was measured. In many cases, the closest tree is among the smallest ones since small trees are measured in the horizontal point sampling only near to the plot centre. From among all possible quantile trees, the smallest tree is perhaps the easiest alternative to be correctly determined in the field without measuring the diameters of

Table 3. The order of strategies where one tree/plot was selected according to the RMSE of the four criterion variables. Abbreviation "No" means no quantile trees used, the other abbreviations are in Table 2. The table is based on simulated realizations where the number of quantile trees was equal to or less than 3.

		Basal area < 12	Basal area > 12
DGM < 11		n = 141	n = 23
	$RMSE_N$	1S Cl 2S Ra La 3S Sa No	2S 1S 3S Cl Ra La Sa No
	$RMSE_{Energy}$	No 3S Cl 2S 1S Ra Sa La	No Cl 3S 1S 2S Ra Sa La
	$RMSE_{Volume}$	La Sa No Ra 3S Cl 2S 1S	No Cl Ra Sa La 1S 3S 2S
	$RMSE_{Sawtimber}$	La Sa Ra 1S 2S Cl 3S No	La Sa Ra 3S 2S 1S Cl No
	Total order	Cl 1S La 3S 2S No Ra Sa	La 1S Sa 2S Cl 3S Ra No
11 < DGM < 18		n = 82	n = 96
	$RMSE_N$	Cl 2S 1S 3S Ra No La Sa	No La Cl 3S 1S Ra Sa 2S
	$RMSE_{Energy}$	3S Sa 2S Ra 1S La Cl No	2S 3S 1S Sa Ra La Cl No
	$RMSE_{Volume}$	La Cl Ra Sa 3S No 2S 1S	No La Cl 1S 3S Ra Sa 2S
	$RMSE_{Sawtimber}$	Sa La Ra Cl 3S 2S 1S No	La Cl Sa Ra No 1S 2S 3S
	Total order	La Sa Cl Ra 3S 2S 1S No	La Cl 1S No 3S Ra 2S Sa
DGM > 18		n = 54	n = 125
	$RMSE_N$	Cl 1S 2S Ra 3S No La Sa	3S La 1S Ra 2S No Cl Sa
	$RMSE_{Energy}$	2S Cl 1S La 3S Ra No Sa	2S 1S 3S Sa La Ra Cl No
	$RMSE_{Volume}$	Cl 2S 1S 3S No Ra La Sa	3S 1S 2S La Ra Sa No Cl
	$RMSE_{Sawtimber}$	2S Cl No 3S 1S Ra La Sa	La 3S No 1S Ra 2S Sa Cl
	Total order	Cl 2S 1S 3S Ra No La Sa	3S 1S 2S La Ra Sa No Cl

Table 4. The order of strategies where one tree/plot was selected according to the standard error of the four criterion variables. For clarification, see Table 3.

		Basal area < 12	Basal area > 12
DGM < 11		n = 141	n = 23
	se_N	1S Cl 2S La Ra Sa 3S No	2S 1S 3S Cl La Sa Ra No
	se_{Energy}	Cl 3S No 2S 1S Ra La Sa	Cl 2S 3S 1S Ra Sa La No
	se_{Volume}	La Sa 3S No Ra Cl 2S 1S	No 1S 2S Sa Cl 3S La Ra
	$se_{Sawtimber}$	La Sa Ra 1S 2S Cl 3S No	La Sa 3S 2S 1S Cl Ra No
	Total order	Cl 1S La Sa 2S 3S Ra No	2S 1S 3S Cl La Sa Ra No
11 < DGM < 18		n = 82	n = 96
	se_N	2S 3S Cl Ra 1S La Sa No	1S La 3S No 2S Cl Ra Sa
	se_{Energy}	3S Cl Ra 2S No La 1S Sa	La No Cl 3S 1S 2S Ra Sa
	se_{Volume}	3S La Ra Cl 2S No Sa 1S	1S La No 3S Cl 2S Ra Sa
	$se_{Sawtimber}$	La Sa Ra Cl 1S 2S 3S No	La Cl 2S 1S 3S No Ra Sa
	Total order	3S La Cl Ra 2S Sa 1S No	La 1S No 3S Cl 2S Ra Sa
DGM > 18		n = 54	n = 125
	se_N	1S Cl 2S Sa Ra 3S La No	2S 1S 3S Ra La Sa Cl No
	se_{Energy}	2S 1S Cl La Sa 3S Ra No	2S 1S 3S La Ra Sa Cl No
	se_{Volume}	1S Sa 2S Cl Ra 3S La No	2S 3S 1S La Sa Ra No Cl
	$se_{Sawtimber}$	Sa 2S 1S Ra Cl 3S La No	La 3S 1S No 2S Sa Ra Cl
	Total order	2S 1S Sa Cl Ra 3S La No	3S 1S 2S La Sa Ra No C

other trees of the HPS plot. This is because it is necessarily rather close to the plot center and all trees that are far from the center are necessarily larger than it. There is a tradition in Finland to assess minimum diameter at the stand level as a part of the inventory by compartments so as to support the prediction of basal area diameter distribution (see e.g. Päivinen 1980). However, finding a population minimum from a stand is not very easy; finding a minimum of a small sample plot is much easier. Thus, measuring the smallest trees of HPS plots as quantile trees is supported both by reduction in RMSE and by the practical issues discussed above.

The use of large trees as quantile trees did not lead to as satisfactory results as the use of smaller trees did, even though they were useful in the prediction of volume and sawtimber volume. The reason may be that basal area diameter distribution already gives much weight to larger trees. Thus, even though the largest trees reduce the standard deviation of volume prediction, the reduction is so small that an increase in bias cancels it. In addition, the basal area median tree was expected to be known, which also may have the effect that no additional information could be obtained from the largest end of the distribution by using quantile tree measurements.

When the results of this study are compared to those of the study by Maltamo et al. (2003) it can be seen that the use of quantile trees did not improve accuracy of stem number as much as did the MSN based non-parametric application where stem number was given special attention. In the study by Maltamo et al. (2003), the use of the same percentile models as in this study led to an RMSE of stem number of 26.77%, while with the MSN method, it was as low as 18.39% in a dataset that is a subset of the data of this study. On the other hand, in this study the use of quantile trees also improved the prediction of sawtimber volume, whereas in the case of MSN, the accuracy decreased. The reason for the worse results in the prediction of stem number may be that in this study, an existing model was calibrated, whereas in the study by Maltamo et al. (2003) a whole diameter distribution model was constructed. Thus, there are methods that produce higher accuracy with respect to some criterion variables, but the quantile tree approach, which only utilizes

the information measured from the target stand, seems to be a rather safe approach with respect to all criterion variables used, and seems to produce a prediction that is rather good for any part of the diameter distribution. In addition, paying more attention to the accuracy of the stem number in the modelling stage could probably lead to the percentile models providing higher accuracy in relation to the number of stems.

In Finland, there is increasing pressure to change the current small area forest inventory system, which is based on a field inventory of forest area by stands. This system has been found to be inefficient, expensive and inaccurate for modern information needs. The main ways to develop inventory have been the use of remote sensing methods and optimization of field measurements. In the case of remote sensing, lidar based applications, in particular, have provided very promising results (e.g. Maltamo et al. 2006). In fact, more accurate stand level variable estimates have been produced than in the case of the current field assessment based system. However, especially in stands with very valuable tree stock, there is also a need for a highly detailed description of stand structure and timber assortments. In such cases, the approach used in this study and in Mehtätalo and Kangas (2005) may be the optimal way to derive sufficient information.

The result that extreme quantiles improve predictions more than intermediate quantiles is also an interesting result from a theoretical point of view. The distributions of intermediate order statistics are more symmetric and closer to normal distribution than those of extreme order statistics, which are skewed. Asymptotically, the distribution of intermediate sample order statistics is a normal distribution, while the distributions of minima and maxima are asymptotically extreme value distributions (Reiss 1989). As the prediction method is based on an assumption about linearity, it should work best under normality. Because of the correlation structure of the percentile models, a measured minimum diameter larger than the corresponding PPM prediction tends to narrow the distribution from both ends, and a small minimum, correspondingly, tends to make it wider. Since the distribution of the minimum is skewed to the right, extremely large minimums are more probable than extremely small minimums. Corre-

spondingly, extremely small maximums are more probable than extremely large maximums. This is why the method more likely narrows the predicted distribution than makes it wider, especially when extreme order statistics are used. This explains why extreme order statistics increased the biases more than intermediate ones.

In the original percentile models, the standard errors increased steadily the more the predicted percentile differed from the known percentile, 50% (Kangas and Maltamo 2000b). The extreme percentiles were the most difficult ones to model. This is due to both the nature of the extreme statistics (above) and the fact that the percentiles closer to 50% also correlate more with it. Therefore, additional information from these parts of the distribution is likely to improve the results more than information from the intermediate parts of the distribution. Thus, even though the assumptions behind the method are better fulfilled by using intermediate order statistics, the greater information content of extreme order statistics make them more useful than the intermediate ones.

As measurements of order statistics from the same sample are correlated, two quantile trees from the same plot do not provide as much information as two order statistics from different plots. This can be seen, for example, in that with an equal number of quantile trees per stand, strategies 1S and 2S give lower RMSEs than strategy 12S (Fig. 2). In addition, a trend in stand structure within the stand may cause the quantile trees to decrease the accuracy when measured from one plot only, as was seen in the case of total and sawtimber volume (Fig 2). Thus, most information would be achieved by measuring one quantile tree from several plots.

It may sometimes be necessary to measure more trees than one per plot in order to sort them correctly. In this case, one would be willing to use all measured diameters as quantile trees. However, this would make the strategy used dependent on the sample obtained. For example, if we originally aim at using the smallest tree of the plot but need to measure also the second smallest one because of a small difference in diameter between these two trees, the measured second smallest tree is likely to be abnormally small. Using both the measured trees would then lead to purposive sampling, which should generally be avoided

(Gregoire 1998). Thus, we warn about using this strategy before studying its effects on accuracy.

Acknowledgements

References

Bitterlich, W. 1984. The relascope idea. Commonwealth Agricultural Bureaux. Farnham Royal. 242 p.

Borders, B.E, Souter, R.A., Bailey, R.L. & Ware, K.D. 1987. Percentile-based distributions characterize forest stand tables. Forest Science 33(2): 570–576.

Casella, G. & Berger, R.L. 2002. Statistical inference, second edition. Duxbury Advanced Series. Pacific Grove, USA. 660 p.

Gove, J.H. & Patil, G.P. 1998. Modelling basal area-size distribution of forest stands: a compatible approach. Forest Science 44(2): 285–297.

Gregoire, T.G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research 1998: 1429–1447.

Gustafsen, H.G., Roiko-Jokela, P. & Varmola, M. 1988. Kivennäismaiden talousmetsien pysyvät (INKA ja TINKA) kokeet. Suunnitelmat, mittausmenetelmät ja aineistojen rakenteet. Finnish Forest Research Institute, Research Papers 292. FFRI, Helsinki, Finland. 212 p.

Haara, A. & Korhonen, K.T. 2004. Kuvioittaisen arvioinnin luotettavuus. Metsätieteen aikakauskirja 4/2004: 489–508.

Kaartinen, K. 2005. Energiapuun määrän ennustaminen mesäsuunnittelutiedolla. Metsäsuunnittelun ja -ekonomian pro gradu. Joensuun yliopisto, metsätieteellinen tiedekunta. 64 p.

Kangas, A. & Maltamo, M. 2000a. Calibrating predicted diameter distribution with additional information. Forest Science 46(3): 390–396.

— & Maltamo, M. 2000b. Percentile-based basal

area diameter distribution models for Scots pine, Norway spruce and birch species. Silva Fennica 34(4): 371–380.

— & Maltamo, M. 2000c. Performance of percentile based diameter distribution prediction and Weibull method in independent data sets. Silva Fennica 34(4): 381–398.

— & Maltamo, M. 2002. Anticipating the variance of predicted stand volume and timber assortments with respect to stand characteristics and field measurements. Silva Fennica 36(4): 799–811.

— & Maltamo, M. 2003. Calibrating predicted diameter distribution with additional information in growth and yield predictions. Canadian Journal of Forest Research 33(3): 430–434.

— , Heikkinen, E. & Maltamo, M. 2002. Puustotunnusten maastoarvioinnin luotettavuus ja ajanmenekki. Metsätieteen aikakauskirja 3/2002: 425–440.

Kilkki, P. & Siitonen, M. 1975. Simulation of artificial stands and derivation of growing stocks models from this material. Acta Forestalia Fennica 145. 33 p.

— , Maltamo, M., Mykkänen, R. & Päivinen, R. 1989. Use of the Weibull function in estimating the basal area dbh-distribution. Silva Fennica 23(4): 311–318.

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

Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of Scots pine. Communicationes Instituti Forestalis Fenniae 134. 69 p.

Lether, F.G. 1984. A FORTRAN implementation of Späth's interpolatory rational spline: drawing a taut smooth curve through data points. Technical Report WL 22, Department of Mathematics, University of Georgia, Athens, Georgia. 15 p. + appendix.

Maltamo, M. & Kangas, A. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Canadian Journal of Forest Research 28(8): 1107–1115.

— & Mabvurira, D. 1999. Prediction of diameter distribution of Eucalyptus grandis in Zimbabwe using varying information. In: Pukkala, T. & Eerikäinen, K. (eds.). Growth and yield modelling of tree plantations in South and East Africa. Joensuun yliopisto, Metsätieteellinen tiedekunta tiedonantoja 97. p. 97–111.

— , Kangas, A., Uuttera, J., Torniainen, T. & Sara-

mäki, J. 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogenous Scots pine stands. Forest Ecology and Management 133(3): 263–274.

— , Malinen, J., Kangas, A., Härkönen, S. & Pasanen, A-M. 2003b. Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. Forestry 76(4): 449–464.

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

Matérn, B. 1972. The precision of basal area estimates. Forest Science 18: 123–125.

Mehtätalo, L. 2004a. An algorithm for ensuring compatibility between estimated percentiles of diameter distribution and measured stand variables. Forest Science 50(1): 20–32.

— 2004b. Predicting stand characteristics using limited measurements. Finnish Forest Research Institute, Research Papers 929. 39 p. + appendices I–V.

— 2005. Localizing a predicted diameter distribution with sample information. Forest Science 51(4): 292–303.

— & Kangas, A. 2005. An approach to optimizing field data collection in an inventory by compartments. Canadian Journal of Forest Research 35(1): 100–112.

— & Nyblom, J. forthcoming. Development and comparison of parameter prediction and recovery approaches for basal area weighted Weibull distribution. Submitted manuscript.

Nyyssönen, A. 1954. Estimation of stand volume by means of relascope. Communicationes Instituti Forestalis Fenniae 44. 131 p.

Päivinen, R. 1980. On the estimation of stem diameter distribution and stand characteristics. Folia Forestalia 442. 28 p.

Reiss, R.-D. 1989. Approximate distributions of order statistics with applications to nonparametric statistics. Springer Series in Statistics. Springer-Verlag, New York. 355 p.

Siipilehto, J. 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. Silva Fennica 33(4): 281–301.

Späth, H. 1974. Spline algorithms for curves and sur-

faces. Utilitas Mathematica Publishing. Inc., Winnipeg, Canada.

Sukvong, S., Frayer, W.E. & Mogren, E.W. 1971. Generalized comparisons of the precision of fixed-radius and variable-radius plots for basal-area estimates. Forest Science 17: 263–271.

Vuokila, Y. 1959. On the accuracy of the relascope method of cruising. Communicationes Instituti Forestalis Fenniae 51. 62 p.

Total of 37 references