

# Modelling Percentile Based Basal Area Weighted Diameter Distribution

Annika Kangas, Lauri Mehtätalo and Matti Maltamo

---

**Kangas, A., Mehtätalo, L. & Maltamo, M.** 2007. Modelling percentile based basal area weighted diameter distribution. *Silva Fennica* 41(4): 425–440.

In percentile method, percentiles of the diameter distribution are predicted with a system of models. The continuous empirical diameter distribution function is then obtained by interpolating between the predicted values of percentiles. In Finland, the distribution is typically modelled as a basal-area weighted distribution, which is transformed to a traditional density function for applications. In earlier studies it has been noted that when calculated from the basal-area weighted diameter distribution, the density function is decreasing in most stands, especially for Norway spruce. This behaviour is not supported by the data. In this paper, we investigate the reasons for the unsatisfactory performance and present possible solutions for the problem. Besides the predicted percentiles, the problems are due to implicit assumptions of diameter distribution in the system. The effect of these assumptions can be somewhat lessened with simple ad-hoc methods, like increasing new percentiles to the system. This approach does not, however, utilize all the available information in the estimation, namely the analytical relationships between basal area, stem number and diameter. Accounting for these, gives further possibilities for improving the results. The results show, however, that in order to achieve further improvements, it would be recommendable to make the implicit assumptions more realistic. Furthermore, height variation within stands seems to have an important contribution to the uncertainty of some forest characteristics, especially in the case of sawnwood volume.

**Keywords** diameter distribution, prediction, interpolation, stand structure

**Addresses** *Kangas*: Department of Forest Resources Management, P.O.Box 27, FI-00014 University of Helsinki, Finland; *Mehtätalo & Maltamo*: University of Joensuu, Faculty of Forestry, P.O. Box 111, FI-80101 Joensuu, Finland

**E-mail** annika.kangas@helsinki.fi, lauri.mehtatalo@joensuu.fi, matti.maltamo@joensuu.fi

**Received** 4 December 2006 **Revised** 7 June 2007 **Accepted** 17 August 2007

**Available at** <http://www.metla.fi/silvafennica/full/sf41/sf413425.pdf>

---

## 1 Introduction

Diameter distribution is one of the most descriptive and important stand characteristics. However, in forestry practice the empirical diameter distribution is seldom measured. For example, in Finnish compartmentwise inventory, the growing stock is described by partly visually assessed stand characteristics, such as mean diameter and basal area, for each tree species. In applications, the diameter distribution is predicted with models. The predicted distribution is used to compute stand volume characteristics with tree-wise height and volume models and as a basis for tree growth predictions (e.g. Päivinen 1980).

In Finland, basal-area weighted diameter distribution has been commonly used, since it can be easily scaled to observed basal area, and basal area is the most important forest characteristic assessed in practical field inventories. Scaling the distribution to observed basal area also ensures good estimates for the stand volume calculated from the distribution (e.g. Kangas and Maltamo 2000b). In applications, however, the basal-area weighted distribution is transformed to a frequency distribution. This is done in order to be able to utilize single-tree growth models, for example. Therefore, in addition to obtaining good estimates of volume, also good estimates of the frequency distribution are needed.

Borders et al. (1987) developed the percentile based diameter distribution prediction method. This method characterises an empirical distribution function with 12 percentiles defined with respect to number of stems in a stand. The number of stems in desired diameter classes was calculated by linear interpolation between the predicted percentiles. Maltamo et al. (2000) used the percentile based approach to predict irregular diameter distributions of stands in a natural state. Gobakken and Næsset (2005) and Maltamo et al. (2006) expanded the use of percentile based distributions to applications where diameter distributions are predicted by using airborne laser scanning data.

Kangas and Maltamo (2000a) estimated percentile based basal-area weighted diameter distribution models for the three most common tree species in Finland. Two sets of models were

estimated: one set with and another without number of stems as a predictor. Transforming the basal-area weighted diameter distribution to the traditional frequency distribution has not, however, produced satisfactory results. The predicted frequency distribution has been decreasing in most stands, especially for spruce (Kangas and Maltamo 2000b, Bollandsås and Næsset 2007). This problem has been particularly evident for the model set without stem number as independent variable.

Overall, the accuracy of percentile based diameter distributions has proved to be quite similar to diameter distributions based on probability distributions (e.g. Kangas and Maltamo 2000b). More than on estimation method, the accuracy depends on the amount of information available from the stand. Inclusion of the number of stems as a predictor improved the total volume and saw timber volume estimates for all species, but the improvements were especially large for number of stems estimates obtained from the predicted distribution. It means that the predicted stem numbers are not correct, even if the stem number were known and used as an independent variable, but that knowing the stem number improves the behaviour of the model system. Thus, it can be assumed that the stem number carries information about the shape of the diameter distribution that is useful.

Stem number is not, however, measured in the usual field work so that the models including stem number as predictor cannot be applied in most cases. Using a predicted stem number in the models does not improve the model behavior: the important information in the stem number is obviously just the variation that cannot be explained with the basal area and mean diameter or other forest characteristics. Siipilehto (2006) presented an approach where a group of stand characteristics including stem number were predicted simultaneously. Measured value of any of these characteristics, e.g. basal area, can then be used to calibrate the estimate of stem number based on the correlations of the errors across the models. This kind of approach could possibly also improve the usability of predicted (and calibrated) stem number in diameter distribution models.

It is, however, also possible to utilize stem number indirectly, by accounting for the ana-

lytical relationships between basal-area weighted diameter distribution and frequency distribution, i.e. between diameter, stem number and basal area. This analytical relationship can be accounted for using a set of models where the stem number is included as a soft constraint in the modelling process. The relations between the errors of the models are then utilized in the Seemingly Unrelated Regression (SUR) or Three Stage Least Square (3SLS) estimation. This enables striving for coefficients optimal for both estimating the basal-area weighted and the frequency distribution.

Furthermore, the models of Kangas and Maltamo (2000a) were estimated from an angle-count sampling data, which is unreliable for the frequencies in the smallest diameter classes (e.g. Schreuder et al. 1993). Estimating the models from a fixed-size sample plot data may also improve the model behaviour in the smallest diameter classes, since such data has smaller variances of frequency in these classes.

The aim of this paper is to investigate the reasons for the unsatisfactory performance of percentile methods without stem number as a predictor and present possible solutions for the problem. We analyze how much the problem can be lessened with simple ad hoc methods like increasing new percentiles into the system. We also present a method accounting for the analytical relationships between basal area, stem number and diameter in the modelling. Our assumption is that taking the analytical relationships into account in modelling is the best approach to alleviate the problems.

## 2 Material

The data set includes the permanent sample plots (INKA sample plots) measured by the Finnish Forest Research Institute (FFRI), originally for growth modelling purposes (Gustavsen et al. 1988). The sample plots were established on mineral soils across Finland. The data includes clusters of three circular plots located systematically within a stand avoiding stand edges. For estimating basal-area weighted diameter distribution, the information of these circular plots was combined. Altogether 100–120 trees were measured in each

stand for diameter at breast height to the nearest 0.1 cm. Tree height was measured from about 30 sample trees to the nearest 0.1 meters. Data were selected according to the following criteria: the basal area of spruce in the stand had to be over 1 m<sup>2</sup>/ha and the number of spruce stems over 50 per hectare; the number of measured trees in the three sample plots had to be at least 10; the basal area median diameter had to be over 5 cm; and the range between both minimum and median and maximum and median diameter had to be over 2 cm. Altogether, the data included 328 stands.

Näslund's height model (1937) was constructed separately for each stand using sample tree measurements. The height of each tally tree was then predicted with these models. A random component was added to the predictions from a normal distribution using the estimated standard deviation of each height model. This was done in order to retain a realistic height variation in the data set. Total, sawnwood and pulpwood volumes were calculated for each tree using taper curve functions presented by Laasasenaho (1982). Basal-area weighted diameter distributions were formed by using basal areas of individual trees. Finally, stand characteristics were calculated as averages and sums of tallied trees (Table 1).

**Table 1.** Range, mean and standard deviation (SD) of main characteristics of the study material. A denotes age,  $d_{gM}$  basal area median diameter,  $h_{gM}$  height of basal area median tree, N number of stems, G basal area, V volume and  $V_s$  sawnwood volume.

Variable	Min	Max	Mean	SD
A, years	13.00	170.00	76.01	32.46
$d_{gM}$ , cm	3.00	36.20	17.12	6.03
$h_{gM}$ , m	2.80	28.73	14.33	4.98
N, ha <sup>-1</sup>	67.45	3186.12	914.30	626.51
G, m <sup>2</sup> ha <sup>-1</sup>	1.01	34.14	12.78	8.96
V, m <sup>3</sup> ha <sup>-1</sup>	2.89	383.23	97.92	82.72
$V_s$ , cm	0.00	319.51	44.92	58.18

### 3 Percentile Based Basal-Area Weighted Diameter Distribution Models

#### 3.1 The Original Models

In the original models estimated by Kangas and Maltamo (2000a), the empirical basal area diameter distribution was described with the aid of percentiles of stand basal area (0, 10, ..., 90, 95 and 100%), denoted by  $d_0, d_{10}, \dots, d_{100}$ . The 5th percentile was not used in this system, since  $d_0$  and  $d_{10}$  were deemed to be quite close in most stands. The logarithms of these 12 diameters were modelled using measured stand variables as predictors using the seemingly unrelated regression (SUR) (Zellner 1962). The median of the distribution (50th percentile) is commonly assessed in compartmentwise inventory in Finland and was thus assumed to be known.

To be able to construct the diameter distribution using the predicted diameter percentiles, all the diameters must be positive. Logarithmic models were used in order to meet this requirement. The diameters are also required to be monotonic with  $d_0 < d_{10} < \dots < d_{100}$ , in order to produce a monotone distribution function and nonnegative frequencies for the diameter classes. Excluding the 5th percentile was assumed to help in producing monotonic distributions. However, to meet this requirement, an additional model was needed. This additional model was used to model the difference between  $d_{10}$  and  $d_0$  with an intercept term. Since SUR estimation minimises the variance with respect to each model considered, the additional model worked like an ad-hoc constraint in the estimation process. This procedure ensured the monotonicity in the estimation data set, but it does not guarantee it in all conditions.

In the application stage, the estimate of the relative basal area in each 1-cm diameter class  $[d, d+1]$  was calculated from the cumulative distribution of diameters  $F$  as  $F(d+1) - F(d)$ . The value of the empirical distribution  $F$  was obtained by interpolating between the predicted percentiles with Späth's rational spline interpolation (Späth 1974, Lether 1984, Maltamo et al. 2000) with parameters  $q_i$  and  $p_i$  having fixed values 25 and 30 for each interval  $i$ . When  $q_i$  and  $p_i$  approach

infinity, the rational spline degenerates to a piecewise linear function, and making  $q_i$  and  $p_i$  zero produces a cubic spline. The used parameter values (thus) produced a nearly piecewise linear interpolation.

#### 3.2 The Implicit Assumptions in the System

The analytical relationship between the basal-area weighted and unweighted diameter distributions can be presented with formulas

$$f_G(d) = \frac{d^2 f_N(d)}{\int_0^\infty u^2 f_N(u) du} \quad (1)$$

and

$$f_N(d) = \frac{d^{-2} f_G(d)}{\int_0^\infty u^{-2} f_G(u) du} \quad (2)$$

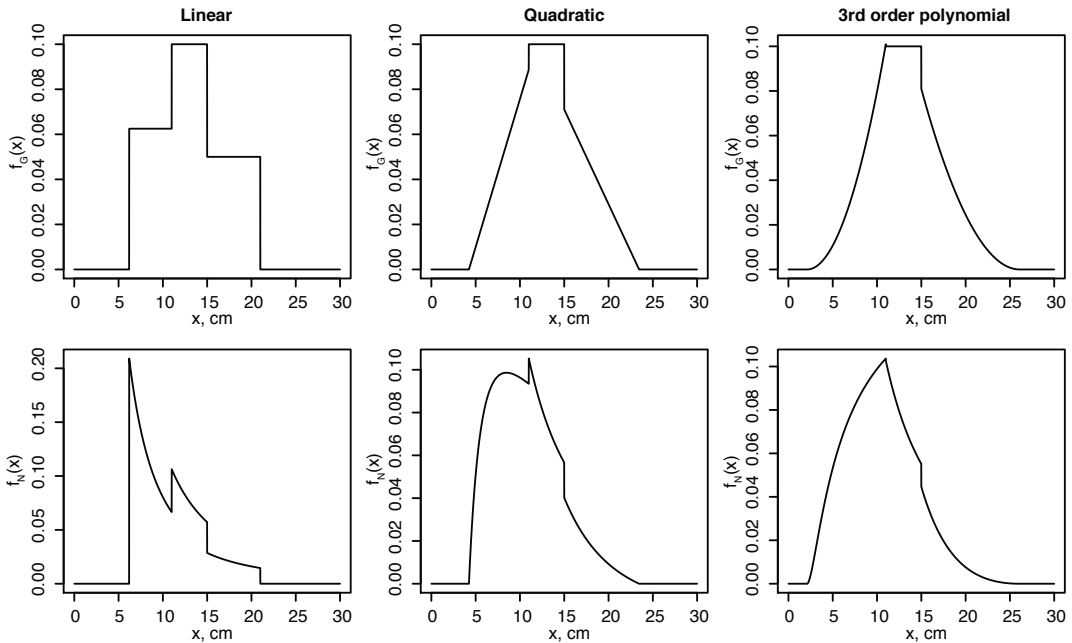
where  $f_G$  denotes the density of basal-area weighted distribution and  $f_N$  the frequency distribution, and the nominator scales the density to unity (e.g. Gove and Patil 1998). With basal-area weighted diameter distribution, the stem number between diameters  $L$  and  $U$  can be estimated from

$$N_{L,U} = \frac{40000G}{\pi} \int_L^U f_G(u) u^{-2} du \quad (3)$$

and the stand stem number is obtained by having  $L = d_0$  and  $U = d_{100}$ .

Assuming a linear interpolation between the predicted percentiles means that the density of basal area is assumed to be uniform within each interval. This means that a decreasing stem number in the diameter classes within this interval is implicitly assumed (Fig. 1). This assumption may be realistic for most of the distribution, but not all. Thus, partly the unsatisfactory results may be due to these implicit assumptions in the linear interpolation, not in the percentile estimates.

It would, however, be possible to utilise special assumptions for tails only. For instance, it could be assumed that the tails were estimated with second- or third-order polynomials (David and Nagaraja 2003, Mehtätalo et al. 2007). In the lower tail, this would cause the unweighted density to be increasing from  $d = d_0$  upwards, but



**Fig. 1.** Densities of basal-area weighted (up) and unweighted diameter distribution where percentiles of basal-area weighted diameter distribution are  $d_5 = 7$ ,  $d_{30} = 11$ ,  $d_{70} = 15$  and  $d_{95} = 20$  and the tails are interpolated using 1st (left), 2nd (middle) and 3rd (right) order polynomials.

not necessarily for the whole interval (Fig. 1). For example, with quadratic interpolation it would turn to a decreasing function at  $d = 2d_0$ . In the upper tail, higher order polynomials would make the tail lighter. Thus, higher order polynomials could be more realistic assumptions than linear interpolation, especially for the lower tail. However, they are also more difficult to parameterize into the model, as the analytic functions for stem number become more complicated.

In the case of linear interpolation, the density of the basal area weighted diameter distribution is constant within each interval  $[d_i, d_{i+1}]$  (Mehtätalo 2004)

$$f_G(d) = \frac{1}{100} \frac{p_{i+1} - p_i}{d_{i+1} - d_i} \tag{4}$$

where  $p_i$  is the cumulative distribution value at percentile  $d_i$ . The stem number for each interval can be obtained from Eq. 3 as

$$n_i = \frac{40000}{\pi} G \left( \frac{p_{i+1} - p_i}{d_{i+1} - d_i} \right) \left( \frac{1}{d_i} - \frac{1}{d_{i+1}} \right) \tag{5}$$

Using second-order polynomials for tails, the density of the basal area weighted diameter distribution within each interval would be

$$f_G(d) = \begin{cases} 0, & d < d_0 \\ 2a_l^2 d + 2a_l b_l, & d_0 < d < d_2 \\ \frac{p_{i+1} - p_i}{d_{i+1} - d_i}, & d_2 < d < d_{90} \\ -2a_u^2 - 2a_u b_u, & d_{90} < d < d_{100} \\ 0, & d > d_{100} \end{cases} \tag{6}$$

where the tails are specified so that the global minimum of the polynomial used in the lower tail is 0 and the global maximum of the polynomial used in upper tail is 1. Parameters  $a_l$ ,  $b_l$ ,  $a_u$  and  $b_u$  can be solved by forcing the interpolated distribution function to pass through predicted 1st and 2nd percentiles in the lower tail and through 90th and 95th percentiles in the upper tail, for instance. This would lead to estimates

$$b_l = \frac{d_1 \sqrt{p_2} - d_2 \sqrt{p_1}}{d_1 - d_2} \tag{7}$$

$$a_l = \frac{-b_l + \sqrt{p_1}}{d_1} \tag{8}$$

for lower tail and

$$b_u = \frac{d_{90}\sqrt{1-p_{95}} - d_{95}\sqrt{1-p_{90}}}{d_{90} - d_{95}} \tag{9}$$

and

$$a_u = \frac{-b_u + \sqrt{1-p_{95}}}{d_{95}} \tag{10}$$

for upper tail. This, in turn, would lead to estimates of minimum and maximum diameters as

$$\hat{d}_0 = -b_l / a_l = d_1 - \frac{\sqrt{p_1}(d_1 - d_2)}{\sqrt{p_1} - \sqrt{p_2}} \tag{11}$$

and

$$\hat{d}_{100} = -b_u / a_u = d_{95} + \frac{\sqrt{1-p_{95}}(d_{90} - d_{95})}{\sqrt{1-p_{95}} - \sqrt{1-p_{90}}} \tag{12}$$

Using the tail model, stem number for the interval  $[d_0, d_2]$  can be obtained from Eq. 3 as

$$n_i = 2 \frac{40000}{\pi} G \frac{\sqrt{p_1} - \sqrt{p_2}}{d_1 - d_2} \left\{ \frac{\sqrt{p_1} - \sqrt{p_2}}{d_1 - d_2} \left[ \ln d_2 - \ln \left( d_1 - \frac{\sqrt{p_1}(d_1 - d_2)}{\sqrt{p_1} - \sqrt{p_2}} \right) - 1 \right] - \frac{d_1\sqrt{p_2} - d_2\sqrt{p_1}}{(d_1 - d_2)d_2} \right\} \tag{13}$$

and that for the interval  $[d_{90}, d_{100}]$  as

$$n_i = 2 \frac{40000}{\pi} G \frac{\sqrt{1-p_{95}} - \sqrt{1-p_{90}}}{d_{95} - d_{90}} \left\{ \frac{\sqrt{1-p_{95}} - \sqrt{1-p_{90}}}{d_{95} - d_{90}} \left[ \ln d_{90} - \ln \left( d_{95} - \frac{\sqrt{1-p_{95}}(d_{95} - d_{90})}{\sqrt{1-p_{95}} - \sqrt{1-p_{90}}} \right) - 1 \right] - \frac{d_{95}\sqrt{1-p_{90}} - d_{90}\sqrt{1-p_{95}}}{(d_{95} - d_{90})d_{90}} \right\} \tag{14}$$

Thus, minimum and maximum diameters need not to be estimated with separate models.

## 4 Approaches for Improving the Results

### 4.1 Methods

If several separate but interrelated models are estimated, simultaneous estimation is needed. Assuming the models to be independent will lead to biased coefficients or, at least, to inefficient estimation (e.g. Zellner 1962, Zellner and Theil 1962). In forest modelling, simultaneous estimation has been used in some growth and yield models (e.g. Borders and Bailey 1986, Zhang et al. 1997, Hasenauer et al. 1998, Eerikäinen 2002, and Siipilehto 2006), and also for diameter distribution models (Borders et al. 1997, Maltamo et al. 2000, Kangas and Maltamo 2000a, Robinson 2004, Maltamo et al. 2006).

Simultaneous equations may be seemingly unrelated or directly related. In the first case, none of the independent variables are estimated with an equation in the system, but the errors of the separate models may be correlated. In the second case, some of the independent variables

in the equations are estimated with another equation (endogenous variables), and some are not (exogenous variables). The errors of the models may or may not be correlated.

In Seemingly Unrelated Regression (SUR), the correlations between the errors of models are accounted for in estimating the coefficients of the models. First, the coefficients are estimated with Ordinary Least Squere (OLS) separately for each model in the group, and the correlations between the errors are estimated. Then, these correlations are used in estimating a final set of parameters. In this case, OLS models are unbiased for each separate model, but modelling efficiency can be improved if the correlations are accounted for, and the independent variables are not the same in each model. If the models are directly related, two-stage (2SLS) or three-stage least squares (3SLS) methods are used. In 2SLS, the equations for the endogenous variables are first estimated, and predicted values of these variables are then used as independent variables when estimating the final set of parameters. This is to ensure the unbiasedness of the approach. In 3SLS, it is also assumed that the errors of the models are correlated, so that after 2SLS the estimated correlations of the model errors are used in the same way as in SUR for estimating the final set of parameters.

It is also possible to use the SUR and/or 3SLS estimation as a sort of “soft constraint”. For instance, Zhang et al. (1997) used the sum of treewise growth estimates to estimate the stand growth simultaneously. This helped to constrain the treewise growth models so that the estimates of standwise growth were more precise. Similar ideas are utilized in this study: the estimates of stem numbers in different diameter classes, based on the estimates of the percentiles, were used to constrain the percentile models.

## 4.2 Modelling Approaches

The first attempt to improve the original models was to re-estimate the models from the fixed-area INKA sample plots, and including three new percentiles, namely 1%, 2% and 5%, into the system. This approach may improve the results in a sense that the intervals in the smaller tail of the distribution would be smaller, so that the implicit

assumption of decreasing density of number of stems within each interval does not have so big effect. It also means that the produced distributions should have less heavy tails overall. On the other hand, this approach is assumed to produce more problems due to non-monotonicity than the original model.

The modelling technique was the same as in original models, SUR. In the old models, a dummy variable for mesic and poorer mineral soils was included, but not in the new ones, and in the new ones the temperature sum (TS) was included unlike in the old ones. These models are later called re-estimated models. The models were re-estimated in a logarithmic form requiring bias correction when transforming the estimates back to arithmetic scale.

In the second attempt to improve the model behaviour, the relationship between basal area and stem number in a diameter class  $[d_i, d_{i+1}]$  was used to constrain the model. Assuming linear interpolation, the model for stem number between any two percentiles is obtained from Eq. 5

$$n_i = b_i G \left( \frac{1}{d_{i+1} - d_i} \right) \left( \frac{1}{d_i} - \frac{1}{d_{i+1}} \right) + \varepsilon_i \quad (15)$$

Thus, the strict analytical relationship (Eq. 5) is loosened by including parameter  $b_i$  and an error term for each class. Then, if the estimated parameters  $b_i$  differ from their theoretical values  $12732,4(p_{i+1} - p_i)$ , it would indicate that linear interpolation assumption does not fit. If the estimated values are near to their theoretical values, linear interpolation assumption is suitable, and the imposed restrictions can be assumed to improve the parameters of interest, namely the coefficients for the percentile models.

As the equations are not linear with respect to percentile diameters, a nonlinear modeling approach is needed. The models were fitted using MODEL procedure of SAS, using nonlinear 3SLS method. The diameter percentiles were estimated with a model of form

$$d_i = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p) + \varepsilon_i \quad (16)$$

where  $\beta_0 - \beta_p$  are parameters to be estimated,  $x_1 - x_p$  are the independent variables. The percentiles were assumed to be endogenous variables and

the exogenous variables were stand basal area (G), logarithm of stand age (t), logarithm of the stand age divided by basal area, the basal area of spruce divided by the stand total basal area, temperature sum (TS) and the logarithm of the basal area median diameter  $d_{50}$ . These independent variables are the same that were used in the re-estimated models. Later these models are referred to as new models. In the new models, diameters were directly estimated so that bias corrections were not needed.

Finally, improvements were attempted using the second-order polynomials for the tails. In this case, models for the minimum and maximum diameter were not estimated at all, but they were estimated using Formulas 11 and 12. The estimates for the stem numbers in interval  $[d_0, d_2]$  was calculated with Eqs. 13 and 14, and in interval  $[d_1, d_2]$  with Eq. 5. The stem number in interval from  $[d_1, d_2]$  could also have been calculated exactly from the second order polynomial, but this approximation seemed to work well enough. The stem number in interval  $[d_0, d_1]$  was obtained from subtraction. Eqs. 13 and 14 were also used as soft constraints in the modelling phase, so that  $80000/\pi$  was replaced with parameters  $b_1$  and  $b_{14}$  and an error term was included as in Eq. 15.

### 4.3 Comparison of the Models

First, the modelling approaches were compared based on the standard errors of the estimated models. The problems due to non-monotonicity were also considered for each case. Stem numbers for each interval were calculated using the analytical relationships presented, and their accuracy was analysed.

Then, the models were compared in an application stage. In this stage, rational spline was used for interpolation, with the same parameters as in the original study (Kangas and Maltamo 2000a). The height and volume models were applied and the accuracy of resulting stand characteristics was calculated. The basic performance of the models was examined by calculating the root mean square errors and biases of stand volume estimates ( $\text{m}^3/\text{ha}$ ) obtained with these methods. Tree total and sawnwood volumes for each diameter class were calculated with Laasasenaho's taper curve models

(1982), using diameter at breast height and tree height as a predictors. Tree height was predicted by using models of Siipilehto (1999). In this approach, the parameters of Näslund's height model are predicted for each stand so that the height of the mean tree (tree with  $d_{50}$ ) coincides with the observed value.

In both stages, the results were compared to the results obtained by using the true percentiles instead of estimated ones. This was done in order to find out how much of the problems were due to percentile estimates, how much were due to other reasons.

The absolute root mean square error (RMSE) was calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (V_i - \hat{V}_i)^2}{n}} \quad (17)$$

where  $n$  is the number of sample stands,  $V_i$  is the true volume of stand  $i$  and  $\hat{V}_i$  is the volume of stand  $i$  estimated from the predicted distribution. The relative RMSE of the volume estimate was calculated by dividing the absolute RMSE by the true mean volume  $\bar{V}$  of the stands. The bias of the predictions was calculated as

$$bias = \frac{\sum_{i=1}^n (V_i - \hat{V}_i)}{n} \quad (18)$$

In addition to stand total volume, the RMSE and bias of sawnwood volume and number of stems were considered.

Finally, an error index proposed by Reynolds et al. (1988) was used in the comparisons as a measure of the goodness-of-fit of the distributions. The error index was calculated in 1-cm diameter classes for stem numbers. Thus, the error index of a given stand was the sum of the absolute differences between the actual and predicted stem frequencies of the diameter classes

$$e = \sum_{i=1}^K \left| \hat{f}_i - f_i \right| \quad (19)$$

where  $\hat{f}_i$  and  $f_i$  are the predicted and true frequency of diameter class  $i$ , respectively, and  $K$  is the number of diameter classes.



## 5 Results

The parameters of original models (Kangas and Maltamo 2000a) are presented in Table 2, those of the re-estimated models including the models for 1%, 2% and 5% percentiles in Table 3 and those of the new models in Table 4. It can be noted that in the re-estimated models, contrary to the prior beliefs, the model for minimum diameter model had a larger standard error than that of the original models, so that using fixed sample plots did not improve the models in this respect. With respect to standard errors of other common percentiles,

those of the re-estimated models were a little smaller than those of the original models in 5 cases out of 10. The re-estimated model required three additional restricting models with only an intercept term in order to produce monotonic distributions, between 2% and 0%, between 100% and 95% and also between 5% and 2%. This is according to prior beliefs.

The RMSEs of the new models are not directly comparable with the other, logarithmic models, but relative RMSEs provide a suitable basis for comparison. The standard errors of the logarithmic models can be interpreted as approximate relative RMSEs for the diameters in an arithmetic

**Table 2.** The coefficients (SUR) of the original model, estimated from angle count plot data (Kangas and Maltamo 2000a). Median point ( $d_{gM}$ ) is expected to be known. Clarifications of variable codes:  $d_0, \dots, d_{100}$  diameter percentiles, Soil = dummy variable for stands on mesic and poorer mineral soil. For other variable codes, see Table 1.

	Intercept	Ln( $d_{gM}$ )	ln(A)	ln(A/G)	ln(G)	Soil	RMSE
Ln( $d_0$ )	-0.3561	0.8351	–	-0.1178	-0.1261		0.4021
Ln( $d_{10}$ )	-0.2120	0.8830	–	-0.0736	–		0.2732
Ln( $d_{20}$ )	-0.1667	0.9679	–	-0.0789	–		0.1711
Ln( $d_{30}$ )	-0.3199	1.0528	–	-0.0379	–		0.1191
Ln( $d_{40}$ )	-0.1315	1.0266	–	-0.0313	–		0.0990
Ln( $d_{60}$ )	0.1766	0.9688	–	–	–		0.0564
Ln( $d_{70}$ )	0.3237	0.8964	0.0348	–	–		0.0770
Ln( $d_{80}$ )	0.4768	0.8381	0.0603	–	–		0.0932
Ln( $d_{90}$ )	0.7771	0.7502	0.0792	–	–	-0.0392	0.1051
Ln( $d_{95}$ )	0.9005	0.7016	0.1014	–	–	-0.0409	0.1164
Ln( $d_{100}$ )	1.3823	0.6241	0.0832	–	–	-0.0682	0.1580

**Table 3.** Re-estimated models (SUR) for different percentile diameters of Norway spruce from INKA data, including the new percentiles (1, 2, and 5%).

	Intercept	Ln( $d_{gM}$ )	ln(A)	ln(A/G)	ln(G/G <sub>tot</sub> )	TS	RMSE
Ln( $d_0$ )	-1.5345	0.6244	0.4001	-0.1624	-0.4964	-0.2072	0.6365
Ln( $d_1$ )	-0.9141	0.6860	0.2156	-0.1070	-0.1622	0.0724	0.2448
Ln( $d_2$ )	-1.0213	0.7675	0.1774	-0.0822	-0.0651	0.1449	0.2339
Ln( $d_5$ )	-0.8715	0.7966	0.1497	-0.0680	–	0.1913	0.1981
Ln( $d_{10}$ )	-0.6844	0.8865	0.1017	-0.0604	–	0.1434	0.1547
Ln( $d_{20}$ )	-0.3571	0.9480	0.0471	-0.0472	–	0.0648	0.1208
Ln( $d_{30}$ )	-0.1803	0.9848	0.0197	-0.0281	–	–	0.0932
Ln( $d_{40}$ )	-0.1115	1.0098	–	–	–	–	0.0535
Ln( $d_{60}$ )	0.1984	0.9591	–	–	–	–	0.0638
Ln( $d_{70}$ )	0.3521	0.9346	–	–	–	–	0.0921
Ln( $d_{80}$ )	0.5970	0.8776	–	–	–	–	0.1172
Ln( $d_{90}$ )	0.7680	0.8530	–	–	–	–	0.1355
Ln( $d_{95}$ )	0.8247	0.8526	–	–	–	–	0.1421
Ln( $d_{100}$ )	0.8443	0.8554	–	–	–	–	0.1412

**Table 4.** New models estimated with 3SLS for different percentile diameters of Norway spruce including the constraining models of stem numbers in different diameter classes. The standard errors of the coefficients are presented in brackets.

	Intercept	ln(A)	ln(A/G)	Ln(G/G <sub>tot</sub> )	ln(d <sub>gM</sub> )	TS	RMSE	RMSE%	R <sup>2</sup>
<i>d</i> <sub>0</sub>	-1.3866 (0.3883)	0.3985 (0.0885)	-0.1969 (0.0574)	-0.6699 (0.1461)	0.7055 (0.0950)	-0.2154 (0.0880)	2.00	53.97	0.23
<i>d</i> <sub>1</sub>	-0.4271 (0.1814)	0.1794 (0.0407)	-0.1290 (0.0297)	-0.2714 (0.0695)	0.6461 (0.0537)	-	1.62	28.44	0.52
<i>d</i> <sub>2</sub>	-0.4033 (0.1695)	0.1129 (0.0358)	-0.1097 (0.0261)	-0.2034 (0.0582)	0.7593 (0.0505)	-	1.78	27.05	0.59
<i>d</i> <sub>5</sub>	-0.5860 (0.1347)	0.0456 (0.0182)	-0.0549 (0.0134)	-	0.9212 (0.0405)	-	1.86	22.35	0.72
<i>d</i> <sub>10</sub>	-0.4700 (0.0982)	-	-0.0432 (0.0095)	-	1.0094 (0.0304)	-	1.74	17.18	0.83
<i>d</i> <sub>20</sub>	-0.3747 (0.0690)	-	-0.0308 (0.0062)	-	1.0403 (0.0215)	-	1.50	12.07	0.91
<i>d</i> <sub>30</sub>	-0.2126 (0.0515)	-	-0.0213 (0.0043)	-	1.0231 (0.0162)	-	1.30	9.13	0.94
<i>d</i> <sub>40</sub>	-0.1240 (0.0251)	-	-	-	1.0149 (0.00829)	-	0.77	4.86	0.98
<i>d</i> <sub>60</sub>	0.2458 (0.0297)	-	-	-	0.9431 (0.00984)	-	1.08	5.82	0.97
<i>d</i> <sub>70</sub>	0.3808 (0.0451)	-	-	-	0.9252 (0.0150)	-	1.80	8.51	0.93
<i>d</i> <sub>80</sub>	0.6128 (0.0520)	-	-	-	0.8735 (0.0173)	-	2.26	10.32	0.90
<i>d</i> <sub>90</sub>	0.8184 (0.0605)	-	-	-	0.8364 (0.0202)	-	2.95	12.16	0.86
<i>d</i> <sub>95</sub>	0.9486 (0.0625)	-	-	-	0.8106 (0.0209)	-	3.25	12.67	0.85
<i>d</i> <sub>100</sub>	0.9372 (0.0621)	-	-	-	0.8261 (0.0207)	-	3.32	12.55	0.86

scale, so that RMSE of 0.63 for logarithm of minimum diameter in re-estimated models corresponds approximately to 63% RMSE of diameter (e.g. Lappi 1986). Interpreted in this way, the relative RMSEs of the new models were better than those of the original models in most of the percentiles (excluding minimum diameter) and also better than those of the re-estimated models, except for percentiles 1, 2 and 5%. This may be partly due to smaller amount of independent variables that were used in the 3SLS approach for those percentiles (all variables not significant with 5% risk level were excluded), and partly due to constraining stem number models. The new models required one additional model between 1% and 0% in order to produce monotonic results. Thus, the information concerning the stem numbers enhanced more satisfactory behaviour of the

model in this sense.

In the stem number models, the parameters *b<sub>i</sub>* differed from their theoretical values less than 1% in 7 cases out of 14, and less than 6.5% in 12 cases out of 14 (Table 5). In the first class, from 0% to 1%, however, the parameter was about 42% smaller than theoretical value and in the last interval 28% smaller than the theoretical value. This indicates that in these two intervals, the stem numbers are consistently overestimated if linear interpolation is used. The small value of *b<sub>i</sub>* thus compensates for the overestimation.

This can also be seen from the estimates of stem number for the 14 intervals obtained from the analytical relationships. Using true percentiles and the estimated percentiles from the re-estimated models, the estimates of stem number were obtained using Eq. 5. Correspondingly, the

**Table 5.** The parameters  $b_i$  of the stem number models, the theoretical parameter values and the standard error of the estimates.

Model	Estimate	Theoretical value	s.e.
$b_1$	74.22165	127.3	2.0682
$b_2$	128.39	127.3	0.7135
$b_3$	380.382	382	1.3758
$b_4$	638.1402	636.6	2.0986
$b_5$	1272.963	1273	3.2281
$b_6$	1282.755	1273	4.0486
$b_7$	1302.129	1273	4.6813
$b_8$	1277.789	1273	5.0712
$b_9$	1299.457	1273	6.4981
$b_{10}$	1296.145	1273	7.0526
$b_{11}$	1336.275	1273	8.4306
$b_{12}$	1278.992	1273	11.4362
$b_{13}$	675.6775	636.6	9.5500
$b_{14}$	457.4792	636.6	4.0518

estimates of stem number were obtained using the estimated percentiles and model (Eq. 15) in the case of new models. The stem number estimates obtained from new models were in three intervals better than those estimated from true percentiles, and in 11 intervals out of 14 better than those estimated from the re-estimated models. Thus, accounting for the analytical relationships improved the estimates (slightly) in 11 cases.

In the first interval, the best estimates were obtained from re-estimated model (RMSE 70.21), and worst with true percentiles (RMSE 165.94). Thus, in this particular interval, the linear interpolation produced the greatest errors. In the case of re-estimated models, the errors in estimated percentiles compensated for that error, which reduced the RMSE. The most probable reason is that using a model shortens the tail, and therefore lessens the effect of implicit assumptions. In the new models, both the small value of  $b_i$  in the model and the shortening tail compensated for the errors due to implicit assumptions, and the result was almost as large a bias as in the case of true percentiles, but to a different direction. If the theoretical value of  $b_i$  had been used in this interval instead of the estimated one, the results would have been better: bias 10.5 and RMSE 108.8. The error due to linear interpolation covers for a large part of the uncertainty involved in the stem number estimates. The proportion of RMSE

based on true percentiles from that of the new models varies from about 20% to 135%, being on average about 75%. Roughly three fourths of the uncertainty in stem number estimates in each interval is thus due to linear interpolation (ignoring the effect of possible compensation).

When the new models were used so that the minimum and maximum diameters and the stem numbers in the first and last interval were predicted with tail estimators (Eqs. 11–14) and the corresponding models were excluded from the modelling phase, the results were bias 38.32 and RMSE 80.91 in the first interval and bias  $-5.57$  and 7.83 in the last interval. Thus, tail estimator improved the result in the first interval, but slightly worsened in the last one. This also indicates that linear interpolation does not fit to the first interval. In last interval, the second order polynomial seems to produce too light tail, and linear interpolation seems to be better. If models based on (13) and (14) were also included as soft constraints in the system, the results were bias  $-43.1$  and RMSE 86.83 for the first interval, and  $-5.32$  and 6.62 for the last one. Thus, this constraint did not improve the fit in the first interval but did so in the last one. In this case, the parameters were 29932 (17.5% greater than the theoretical value) and 22176 (12.9% smaller than the theoretical value). Thus, the constraint fitted clearly better than the one based on linear interpolation, but yet a better assumption would be required to obtain a truly useful constraint. On the other hand, using the second order polynomial for tail produced monotonous distributions in all stands without any ad-hoc constraints.

When the models were implemented into an application, and the resulting stand characteristics from all the three different models (original, re-estimated and new) were compared to the corresponding results obtained by using true percentiles, the results were quite surprising. The accuracy of forest characteristics obtained in the application phase was fairly similar in all these cases, except for the error index (Table 7).

The error index, which describes how well the distribution fits, was clearly better with true percentiles. The error index could also be improved from 10.808 to 9.461 by introducing three new percentiles, but it could not be further improved by using the analytical stem number information,

even though the class-wise stem number estimates in most classes could be improved (Table 6). True percentiles did produce only slightly better results than the percentiles estimated with new models for volume (RMSEs 11.828% and 12.062%, respectively) and sawnwood volume (RMSEs 18.336% and 22.195%, respectively) (Table 7). In stem number estimates, reduction in the relative RMSE of stem number was from 22.116% (original models) to 20.587% (new models).

As the stem number estimate is likely to be effected with the long and heavy tails, an ad-hoc shortening of the tail was also tested. This was carried out by using 1% diameter as a “true minimum” in the estimation. With this value, the RMSE of stem number could be reduced to 6.059

and bias to 5.212 (Table 7). However, the use of 1% diameter as a minimum in true percentile values produced worse diameter distributions than the re-estimated or new models, when the distributions were visually inspected (Fig. 2).

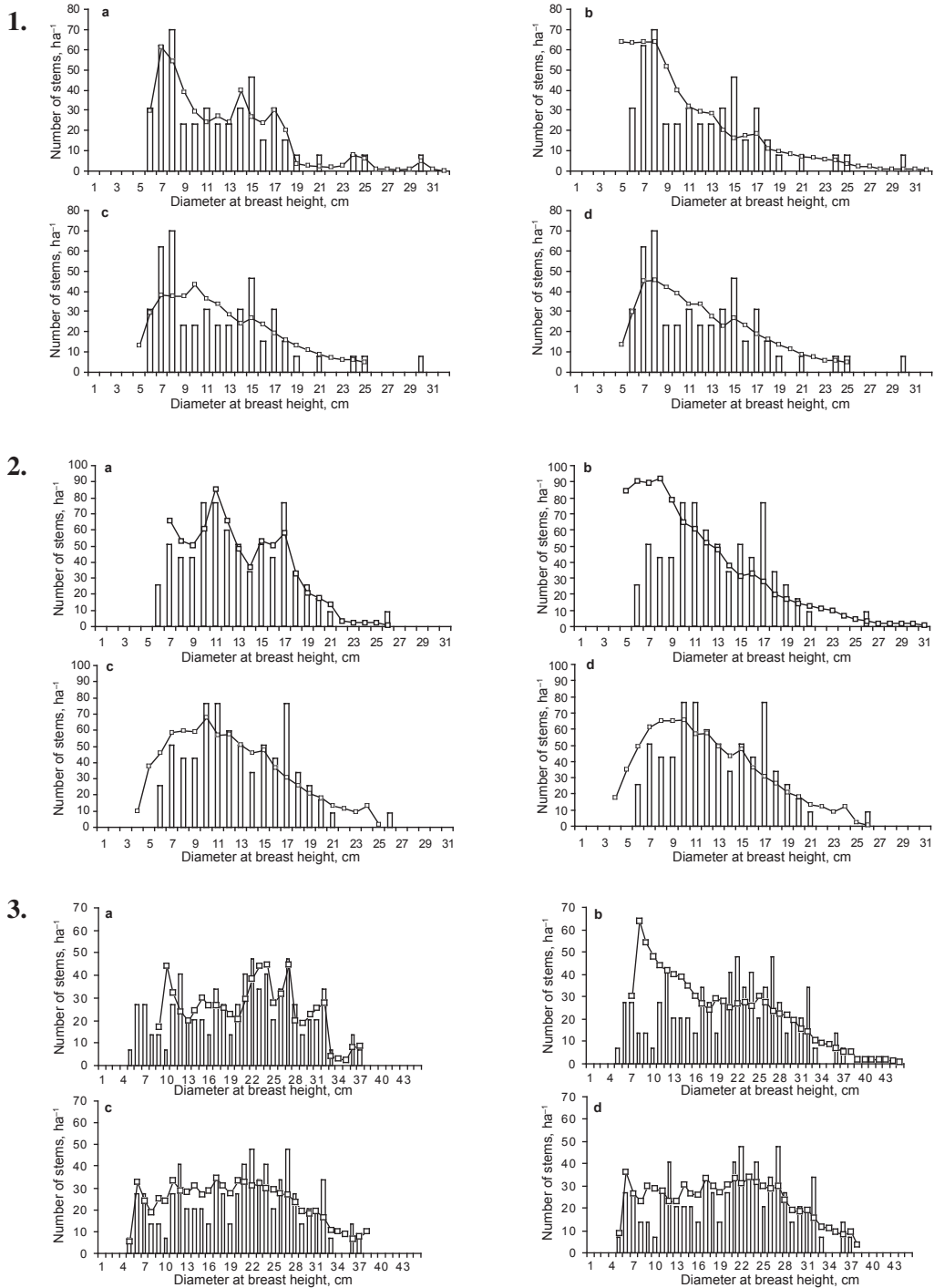
Irrespective of the seemingly minor improvements, using the stem number information for constraining the percentile models seems to force the models to behave visually more satisfactorily. In Fig. 2 are shown three example stands, for which the diameter distributions with different models are presented. In these examples, the density functions obtained with new models are no more decreasing, as they were with the original models.

**Table 6.** The accuracy of class stem number estimates (i.e. stem numbers at percentiles 1–100 minus the stem number at the preceding percentile, denoted by  $ir_1 - ir_{100}$ ), estimated with true percentiles, with the re-estimated models and Eq. 5, and estimated percentiles and the stem number models included in the new model system.

	True percentiles			Re-estimated INKA models			New models		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Ir <sub>1</sub>	-53.72	157.00	165.94	24.40	65.83	70.21	44.45	114.10	122.45
Ir <sub>2</sub>	1.08	6.07	6.16	6.23	22.08	22.94	6.37	22.55	23.43
Ir <sub>5</sub>	-0.45	7.94	7.96	9.45	39.32	40.44	10.63	39.15	40.57
Ir <sub>10</sub>	0.11	7.65	7.65	7.29	35.68	36.42	6.30	33.58	34.17
Ir <sub>20</sub>	-0.27	7.72	7.73	5.24	33.98	34.39	3.18	32.33	32.49
Ir <sub>30</sub>	0.16	6.94	6.94	2.12	18.33	18.45	0.95	17.68	17.71
Ir <sub>40</sub>	1.74	6.35	6.58	1.85	12.45	12.58	0.43	9.84	9.85
Ir <sub>50</sub>	0.39	5.59	5.60	-0.54	7.14	7.16	-0.36	4.26	4.27
Ir <sub>60</sub>	1.03	6.07	6.15	0.42	6.06	6.08	-0.09	3.40	3.40
Ir <sub>70</sub>	0.81	5.56	5.62	0.76	8.14	8.17	0.33	6.15	6.16
Ir <sub>80</sub>	1.91	5.61	5.92	2.25	9.20	9.47	0.46	7.61	7.62
Ir <sub>90</sub>	0.74	6.30	6.35	0.97	9.66	9.71	0.19	7.92	7.92
Ir <sub>95</sub>	0.84	4.31	4.39	-0.49	6.44	6.46	-1.37	5.27	5.44
Ir <sub>100</sub>	-1.99	3.33	3.88	-6.29	5.88	8.61	-2.91	4.96	5.75

**Table 7.** The relative RMSE and absolute biases of volume, sawnwood volume, stand stem number and error index, calculated with true percentiles, with true percentiles modified so that 1% diameter was used as a minimum diameter, with original model percentiles, with re-estimated models, and with the new models.

		True	True modified	Original	Re-estimated INKA	New
V	RMSE%	11.828	11.911	11.637	12.062	12.069
	Bias	-0.067	-0.147	-1.126	-0.100	-0.059
V <sub>sawnwood</sub>	RMSE%	18.336	18.582	21.893	22.062	22.195
	Bias	-0.999	-1.046	-0.551	-0.356	-0.453
N	RMSE%	21.412	6.059	22.116	21.139	20.587
	Bias	-69.672	5.212	-12.049	50.235	40.717
Error index		6.833	7.086	10.808	9.461	9.462



**Fig. 2.** Examples of predicted density functions in three example stands (1–3): a = true percentile values (1% diameter used as a true minimum), b = original percentile models of Kangas & Maltamo (2000b), c = re-estimated percentile models and d = new model. True distribution is presented with bars, and the estimate with lines.

## 6 Discussion

In this paper, we investigated the reasons for the unsatisfactory performance of basal area weighted percentile-based diameter distribution models and presented possible solutions for the problem. The improvements were attempted by including new percentiles into the system and using analytical relationships between basal area, stem number and diameter. The results were indeed better than those obtained using the original model in the sense that the basal-area weighted distributions transformed to traditional density functions were not decreasing (Fig. 2). However, otherwise the improvements were minor, except in the error index, which could be decreased by 12.4%.

When the results were compared to those obtained using true percentiles in the application phase, it can be seen that the results cannot be much improved by improving the percentile models. Small improvements are still possible with respect to error index, but in other respects further improvements in the results seem very difficult. Therefore, there are other sources of uncertainty involved, which seem to be more important than the errors in percentile estimates.

Besides the (nearly linear) spline interpolation used, another likely source of error is the height variation involved in the data set. This is supported by the fact that the original percentile models produced RMSE of about 3.29% (Kangas and Maltamo 2000b) for volume in the same INKA dataset, while in this study the RMSE was 11.83% even with true percentiles. In the former case, the volumes of the trees were calculated using diameter as the only independent variable, both for the measured trees used as reference information and for the trees sampled from the predicted distribution. Thus, height variation did not affect the results. In this study, a volume model with both height and diameter as independent variables was used, and the height variation of the reference data was accounted for.

For sawnwood volume, the RMSE was 18.34% with true percentiles and 21.89% for the original models. The effect of percentile estimates in this study seems thus to be fairly low, and that of height variation and linear interpolation assumption cover for most of the error (roughly

four fifths). In the study by Kangas and Maltamo (2000b) the original models produced RMSE 11.85% for sawnwood volume in the same INKA data set that was used for the current study. Partly the difference can be explained with the fact that height variation was not accounted for in the original study, but partly the difference may be due to the differences in the definition of sawnwood volume. In the original study, all trees greater than 17 cm at breast height were considered as sawnwood trees, and their volume formed the sawnwood volume. Because of the definition, most of the errors were likely to be due to the errors in percentile estimates. In this study, the sawnwood volume was calculated as the proportion of tree volume fulfilling the minimum diameter requirements for each tree sampled from the distribution. Consequently, since the RMSE observed in the original study was fairly high, it can be assumed that some of the errors in percentiles were compensated with some other error sources in the case of new models, and the true effect is more than one fifth of the errors.

Another interesting phenomena in this study is that the other set of original percentile models with stem number as an independent variable (Kangas and Maltamo 2000a), produced clearly better estimates for stem number than the true percentiles in this study in the same data set, namely RMSE 6.88% and bias -40.69. It can be assumed that the models with stem number adjusted the estimate of minimum percentiles so that the implicit bias could be partly corrected. The most reasonable explanation for this is that those models shortened the tails, and thus lessened the effect of implicit assumptions.

Therefore, it seems that the bias and problems in the stem number are mostly due to the (nearly) linear interpolation used in the application. It may be that there are single trees in the stand that have a diameter much smaller than other trees, and these cases may produce problems to estimation, as linear interpolation of basal-area weighted diameter percentiles implicitly assumed a decreasing density between 0% and 1% percentiles.

Thus, the most significant remaining error sources are other than percentile estimates. It seems that further improvements would require that the implicit assumptions in the models are

made more realistic. It would also mean that the height models should be improved. Adjusting the shape of the height model in addition to level could help. This means that more than one height sample tree should be measured and the height sample trees should not be too similar to each other in diameter. It could also be useful to utilize the whole conditional distribution of tree heights given the tree diameter, instead of using only conditional expectations. This would require integration of the volume estimates over the distribution of heights. Simpler approaches would be simulation of heights from the conditional distribution or a two-point distribution approach as used by Lappi et al. (2006).

## 7 Conclusions

The percentiles models of Kangas and Maltamo (2000a) have been widely used in forest planning calculations, e.g. in MELA simulator. Although the test results have shown the accuracy of these models being comparable with other models, the behaviour of percentile based models has been unsatisfactory, especially in Norway spruce stands. The work concerning the improvement of percentile models of this study has improved the behaviour of the models in the sense that frequency functions are not always decreasing any more. This study concentrated on the prediction of diameter distribution using current stand characteristics but it is expected that the usability of new percentiles models will be better also in growth simulations. This is due to the fact that the number of small trees will now be more realistic. The future research work will include the modelling of new percentile models also for other tree species such as Scots pine and birch species.

## Acknowledgements

The research work reported in this paper was funded by the Academy of Finland (decision numbers 200775, 212680), and the Universities of Helsinki and Joensuu.

## References

- Bollandsås, O.M. & Næsset, E. 2007. Estimating percentile-based diameter distributions in uneven-aged Norway spruce stands using airborne laser scanner data. *Scandinavian Journal of Forest Research* 22: 33–47.
- Borders, B.E. & Bailey, R.L. 1986. A compatible system of growth and yield equations for slash pine fitted with restricted three-stage least squares. *Forest Science* 32: 185–201.
- , Souter, R.A., Bailey, R.L. & Ware, K.D. 1987. Percentile-based distributions characterize forest stand tables. *Forest Science* 33: 570–576.
- David, H.A. & Nagaraja, H.N. 2003. *Order statistics*. Third edition. Wiley. Hoboken, New Jersey, USA. 458 p.
- Erikäinen, K. 2002. A site dependent simultaneous growth projection model for *Pinus kesiya* plantations in Zambia and Zimbabwe. *Forest Science* 48: 518–529.
- Gobakken, T. & Næsset, E. 2005. Weibull and percentile models for lidar-based estimation of basal area distribution. *Scandinavian Journal of Forest Research* 20: 490–502.
- Gove, J.H. & Patil, G.P. 1998. Modeling basal area-size distribution of forest stands: a compatible approach. *Forest Science* 44: 285–297.
- Gustavsen, H.G., Roiko-Jokela, P. & Varmola, M. 1988. Kivennäismaiden talousmetsien pysyvät (INKA ja TINKA) kokeet: Suunnitelmat, mittausmenetelmät ja aineistojen rakenteet. *Metsäntutkimuslaitoksen tiedonantoja* 292. 212 p. (In Finnish.)
- Hasenauer, H., Monserud, R.A. & Gregoire, T.G. 1998. Using simultaneous regression techniques with individual-tree growth models. *Forest Science* 44: 87–95.
- Kangas, A. & Maltamo, M. 2000a. Percentile based basal area diameter distribution models for Scots pine, Norway spruce and birch species. *Silva Fennica* 34: 371–380.
- & Maltamo, M. 2000b. Performance of percentile based diameter distribution prediction method and Weibull method in independent data sets. *Silva Fennica* 34: 381–398.
- Laasasenaho, J. 1982. Taper curve and volume function for pine, spruce and birch. *Communications 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.
- , Mehtätalo, L. & Korhonen, K.T. 2006. Generalizing sample tree information. In: Kangas, A. & Maltamo, M. (eds). *Forest inventory: methodology and applications*. Springer. p. 85–106.
- 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. 15 p + app.
- Maltamo, M., Kangas, A., Uuttera, J., Torniaainen, T. & Saramäki, J. 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogeneous Scots pine stands. *Forest Ecology and Management* 133: 263–274.
- , Eerikäinen, K., Packalén, P. & Hyypä, J. 2006. Estimation of stem volume using laser scanning based canopy height metrics. *Forestry* 79: 217–229.
- Mehtätalo, L. 2004. An algorithm for ensuring compatibility between estimated percentiles of diameter distribution and measured stand variables. *Forest Science* 50: 20–32.
- , Gregoire, T.G. & Burkhart, H.E. 2007. Comparing strategies for modelling tree diameter percentiles from remeasured plots. Revised manuscript.
- Näslund, M. 1937. Skogsförsöksanstaltens gallringsförsök i tallskog. Meddelanden från Statens Skogsförsöksanstalt 29. 169 p. (In Swedish.)
- Päivinen, R. 1980. On the estimation of stem diameter distribution and stand characteristics. *Folia Forestalia* 442. 28 p.
- 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: 373–399.
- Robinson, A. 2004. Preserving correlation while modeling diameter distributions. *Canadian Journal of Forest Research*. 34: 221–232.
- Schreuder, H.T., Gregoire, T.G. & Wood, G.B. 1993. *Sampling methods for multiresource forest inventory*. John Wiley & Sons. 446 p.
- Siipilehto, J. 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. *Silva Fennica* 33: 281–301.
- 2006. Linear prediction application for modelling the relationships between a large number of stand characteristics of Norway spruce stands. *Silva Fennica* 40: 517–530.
- Späth, H. 1974. *Spline algorithms for curves and surfaces*. Utilitas Mathematica Publishing, Winnipeg, Canada.
- Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of American Statistical Association* 57: 348–368.
- & Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* 30: 54–78.
- Zhang, S., Amateis, R.L. & Burkhart, H. 1997. Constraining individual tree diameter increment and survival models for loblolly pine plantations. *Forest Science* 43: 414–423.

*Total of 30 references*