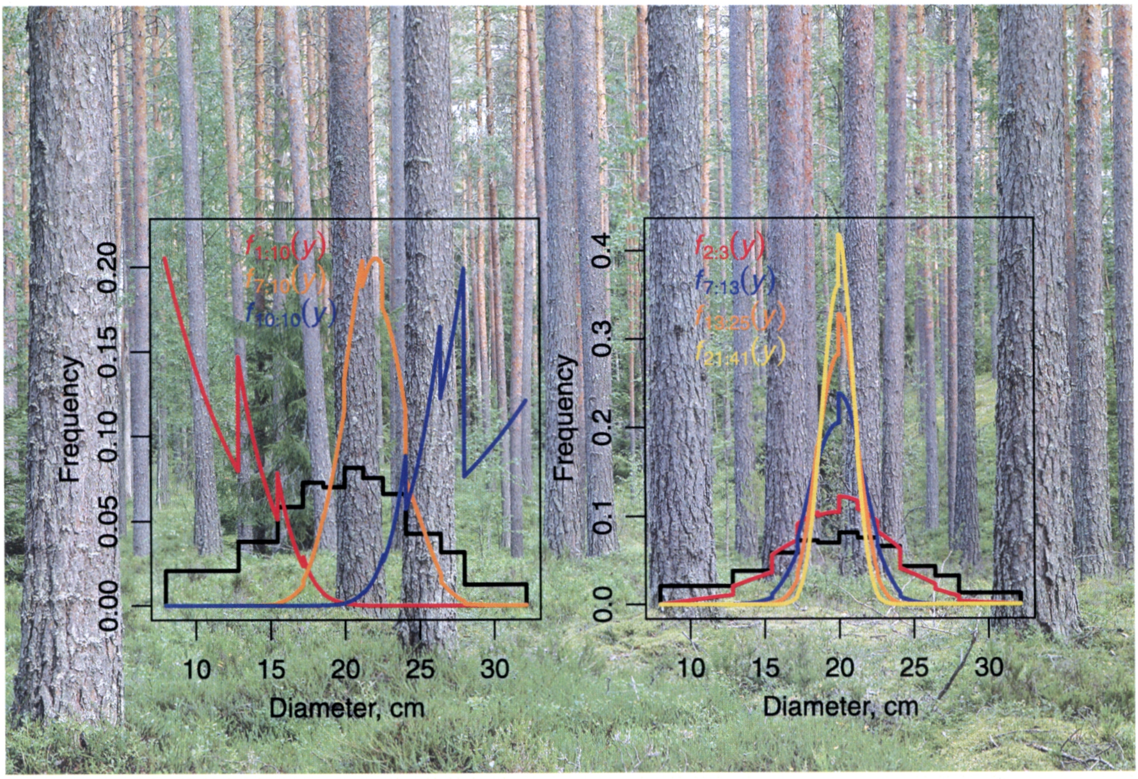
METSÄNTUTKIMUSLAITOKSEN TIEDONANTOJA 929, 2004
FINNISH FOREST RESEARCH INSTITUTE, RESEARCH PAPERS 929, 2004



Predicting Stand Characteristics Using Limited Measurements

Lauri Mehtätalo

JOENSUUN TUTKIMUSKESKUS
JOENSUU RESEARCH CENTRE

METLA

Predicting Stand Characteristics Using Limited Measurements

Lauri Mehtätalo

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Forestry of the University of Joensuu, for public criticism in auditorium K1 of the University, Yliopistokatu 2, Joensuu, on 24th September, 2004 at 12 o'clock noon.

Joensuu 2004

Mehtätalo, Lauri. 2004. Predicting stand characteristics using limited measurements. Metsäntutkimuslaitoksen tiedonantoja 929. Finnish Forest Research Institute, Research Papers 929, 39+72p. ISBN 951-40-1934-2, ISSN 0358-4283.

Publisher The Finnish Forest Research Institute
Joensuu Research Centre

Accepted by Kari Mielikäinen, Research Director, in September 2004

Supervisors Prof. Annika Kangas
University of Helsinki

Dr Juha Lappi
Finnish Forest Research Institute

Prof. Matti Maltamo
University of Joensuu

Reviewers Dr Jeffrey H. Gove
USDA Forest Service

Prof. Timothy G. Gregoire
Yale University

Examiner Dr Risto Ojansuu
Finnish Forest Research Institute

Front cover Distributions of sample order statistics (coloured lines) when the sample was drawn from a percentile-based diameter distribution (black line).
Photo: Erkki Oksanen/FFRI

Author's contact information

Address: Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, Fin-80101 Joensuu, Finland
Phone: +358-10-211 3051
email: Lauri.Mehtatalo@metla.fi
Fax: +358-10-211 3113

Yliopistopaino
Joensuu 2004

Abstract

This thesis reports a new system for the production of static stand description in an inventory by compartments. The stand description includes stock density, diameter distribution and height-diameter (*H-D*) curve.

The diameter distribution of the stand is expressed with percentiles. Firstly, expected percentiles are predicted with regression models using measurements of stand variables. Secondly, the predicted percentiles are localized for the stand using order statistics of horizontal point sample plots (HPS-plots) (i.e. quantile trees), which are interpreted as measured percentiles of the stand. Thirdly, the obtained localized percentiles are adjusted in order to ensure compatibility with the measured stem number. The expected *H-D* curve of the stand is predicted using the measured stand variables. Furthermore, it is localized for the stand using height sample trees. The longitudinal character of the model makes it possible to use measurements from several points in time. Both the localizations of diameter distribution and of the *H-D* curve are based on the prediction of random effects of the models with the best linear unbiased predictor using sample measurements of the response.

The individual components are combined as a new system for the prediction of stand description. A key feature of the system is its ability to utilize different amounts of input information. Furthermore, measurement errors of stand variables are utilized to some degree. The minimum input of the system, which can be obtained from one HPS-plot, consists of measurements of the basal area, basal area median diameter (*DGM*), stand age and site fertility class. Additional HPS-plots, stem number measurement(s) from fixed plot(s), old or new height sample trees and quantile trees can also be utilized. The system makes it possible to take more measurements from stands with a high accuracy requirement than from stands with a low accuracy requirement.

The system was utilized in estimating a model of expected errors of predicted volume and saw timber volume using different measurement strategies in different stands. The prediction error depended on the basal area and *DGM* of the stand and on the number of HPS-plots, height sample trees and quantile trees. Furthermore, a height measurement from a previous inventory decreased the prediction errors slightly. The measurements of stem number did not significantly improve the accuracy of volume and saw timber volume predictions. The estimated models were used as objective functions in a constrained optimization problem, where the object was to find an optimal measurement strategy for a single stand in an inventory where measurement time is limited.

Key words: stand structure, diameter distribution, height-diameter, order statistics, linear prediction, longitudinal analysis, optimization, forest planning, calibration, adjustment, localization.

PREFACE

This work was strongly based on the previous work of my supervisors, Matti Maltamo, Annika Kangas and Juha Lappi. Their ideas and earlier findings have had a strong influence on all stages of this work. The work of Annika Kangas and Matti Maltamo on diameter distributions and stand wise inventory, and the work of Juha Lappi on applications of mixed models and linear prediction in forestry, provided a good toybox for a young researcher to play with. My work has actually only consisted of taking different kinds of building blocks from this toybox, building new wobbly towers and trying to get them to stand upright. The supervisors have always been available for discussions and comments about the work being done. I am grateful for this unfailing support, which I have received thoughout the various stages of this work. I also wish to thank my pre-examiners Jeffrey H. Gove and Timothy G. Gregoire for their valuable comments on the work done.

The funding for this work was provided by the Academy of Finland (decision numbers 73392 and 200775). Annika Kangas, Jyrki Kangas and Kari T. Korhonen applied for the funding and managed the research projects. The work was carried out at the Finnish Forest Research Institute, Joensuu Research Centre, where Tuula Nuutinen, Jari Parviainen and Jyrki Kangas have provided me with good facilities for carrying out this research.

My position as an "observing associate member" of the MELA-group, led by Tuula Nuutinen, has given me an insight into the development of a forest planning system in practice and provided me with the contacts to help me with practical problems. The datasets used in this study have been collected by Timo Pukkala and FFRI. Pekka Leskinen, Juha Alho, and a group of researchers on forest assessment at the University of Joensuu, have given me constructive comments about original manuscripts. The course on Statistical Inference, held by Jukka Nyblom in 2003, was a source of inspiration for me. Without this course, Paper I would hardly have been written. Lisa Lena Opas-Hänninen revised the language of all parts of this thesis. Our telephone calls during the process were valuable private English lessons for me. Coffee breaks with Eero Muinonen, Ulla Mattila and Reetta Lempinen let me forget my research for a moment. The office employees of the research centre helped me always when I forgot my key at home or in my room. Leena Karvinen and Seija Partanen helped me with the makeup of this thesis. I am grateful to all these people for their contribution to this work.

However, my life is mostly outside the office. I give my warmest thanks to my wife, Hilkka, for her understanding and support during this work, and to my children, Juho and Iida, who use real building blocks for their towers, for reminding me about where the real life is. Their bright eyes and joyful voices remind me every day who and what I am living for.

Joensuu, August 2004

Lauri Mehtätalo

LIST OF ARTICLES

This thesis comprises the present summary and the following articles, which are referred to in the text by Roman numerals I-V.

I	Mehtätalo, L. Forthcoming. Localizing predicted diameter distribution with sample information. Revised manuscript (Forest Science).
II	Mehtätalo, L. 2004. An algorithm for ensuring compatibility between estimated percentiles of diameter distribution and measured stand variables. Forest Science 50(1): 20-32.
III	Mehtätalo, L. 2004. A longitudinal height-diameter model for Norway spruce in Finland. Canadian Journal of Forest Research 34(1): 131-140.
IV	Mehtätalo, L. Forthcoming. Height-diameter models for Scots pine and birch species in Finland. Submitted manuscript (Silva Fennica).
V	Mehtätalo, L. and Kangas, A. Forthcoming. An approach to optimizing field data collection in an inventory by compartments. To appear in Canadian Journal of Forest Research.

Paper V was planned jointly with Kangas. I built the system, calculated the results and was mainly responsible for writing the article.

CONTENTS

1 Introduction

The aim of forest planning is to find a management strategy for a forest area that maximizes the utility for the forest owner (Pukkala 1994). The traditional primary unit of forest planning in Finland is a forest stand. The forest plan includes management suggestions for every stand of the forest area under consideration. In order to collect the data for the plan, an inventory by compartments is carried out (Poso 1983).

A Finnish inventory by compartments is based on few horizontal point sample plots (HPS-plots, i.e. relascope sample plots, angle count sample plots, Bitterlich plots). The plots are established subjectively by the person carrying out the inventory at locations that seem representative for the stand. Using measurements and visual assessments of the plots, the most important characteristics, including basal area, basal area median diameter (*DGM*), height of a *DGM*-tree, stand age and site fertility class are assessed from each stratum of the growing stock in the stand (Paananen et al. 2000). The simulator of a forest planning system utilizes the stand wise characteristics to produce a stand description, including stock density, diameter distribution and height-diameter (*H-D*) curve of the stand. In the simulator of the MELA-system (Redsven et al. 2004), which is the core of most forest planning systems in Finland, they are predicted using the models of Mykkänen (1986), Veltheim (1987), Kilkki et al. (1989), Siipilehto (1999) and Kangas and Maltamo (2000b). The obtained stand description is used to generate a set of representative trees for each stand for the prediction of growth (Hynynen et al. 2002) and cutting removal in alternative management schedules.

The accuracy of the input of the Finnish stand simulation system based on partially visually assessed stand characteristics is reported to be rather low (Poso 1983, Mähönen 1984, Laasasenaho and Päivinen 1986, Pussinen 1992, Pigg 1994, Kangas et al. 2002, Haara and Korhonen 2004, Kangas et al. 2004), and prediction errors of growth models decrease it further (Kangas 1997, 1998a, 1998b, 1999). The data users are, however, satisfied with the accuracy, but would not like to see it decline (Uuttera et al. 2002). The aim of Finland's National Forest Programme 2010 (1999) is to increase the cover of forest plans from 50% to 75% by 2010. This requires decreasing the costs of planning per unit area (Heikinheimo 1999, Paananen 2002, Saramäki et al. 2003), which means that the current level of accuracy should be retained at a lower cost than previously.

One possibility to respond to these needs is to develop methods and models that provide as accurate assessments as the current inventory system but at a lower cost. Many studies have investigated possibilities to carry out the inventory from the air using aerial photographs (e.g. Pitkänen 2001, Anttila 2002a, Anttila and Lehikoinen 2002, Korpela 2004), satellite imagery (e.g. Hyvönen 2002, Saksa et al. 2003) or laser scanning approaches (e.g. Holmgren et al. 2003, Maltamo et al. 2004a). Even if these approaches are promising, they are not yet real alternatives to the inventory by compartments (Uuttera et al. 2002) and their development will take time. In other studies, the information of the previous inventory has been updated by utilizing growth models and information about treatments from forestry databases (Hyvönen and Korhonen 2003) or aerial photographs (Anttila 2002b) in order to lengthen the interval of two inventories. These approaches can be regarded as practical variations of the approach of Ståhl (1994), where it was suggested that if the expected utility of an inventory was greater than its costs then a new inventory of a forest stand should be carried out.

Another possibility is to search for possible new variables to be measured in the field that would provide more information at a lower cost than the currently measured stand variables. In many studies carried out in Finland, using accurate stem number in the calculations has been found to be useful (Siipilehto 1999, Kangas and Maltamo 2000b). However, several studies (e.g. Kangas et al. 2004) have reported large errors in measurements of stem number. The calibration algorithm of Kangas and Maltamo (2000a) made it possible to utilize different stand variables in the prediction of diameter distribution and they found that of several potential new variables the unweighted median, maximum and minimum diameters were the most promising alternatives for new measurements. In addition, new measurement equipment is being developed in order to make it possible for a single person to measure a large sample of diameters and heights rapidly in the field (Laasasenaho et al. 2002, Koivuniemi 2003).

Eid (2000) reported that losses caused by mistiming of harvests are most serious in young stands and in stands that are close to their economically optimal rotation age, while in middle-aged and over-mature stands the losses are smaller. Thus, it was suggested that the inventory data should be most accurate in stands where expected losses are greatest. In addition, Kangas and Maltamo (2002) proposed that different variables should be measured in different kinds of forests. Furthermore, Holmström et al. (2003) reported that in large stands, intensive field sampling should be carried out while in smaller stands a less intensive inventory might be satisfactory. All these studies indicate that money could be saved by varying the forest inventory strategy from stand to stand.

This study aims at reducing the costs and improving the accuracy of the traditional inventory by compartments by looking for new measurements, making the use of the collected data more efficient and allocating the measurement resources to measurements and stands where the utility is greatest. This requires a calculation system that is able to utilize various input information. The current system for production of stand description requires a fixed set of variables from each stand. Furthermore, it assumes that the input information is measured without errors, even if it is well known that the errors of stand measurements are large. Thus, in this study a new system for predicting stand description was developed.

In predicting forest characteristics, the use of regression models has become very popular. Some easily measurable stand variables, the most common of which are basal area, mean diameter, stand age and site type, are used to predict other stand characteristics that cannot be measured, or at least are difficult to measure, in the field. Examples of these are diameter distribution (e.g. Rennolls et al. 1985), tree height (e.g. Fang and Bailey 1998), stand growth (e.g. Woollons 1997), log volume reduction (e.g. Mehtätalo 2002), damages (e.g. Jalkanen and Mattila 2000) and the production of berries (e.g. Ihalainen et al. 2003). However, the most natural means to estimate any stand characteristic is to measure it. Thus, if measurement of the target variable is possible, it should be preferred over the measurements of covariates of a regression model (see e.g. Lappi 1997). This study applies this principle to diameter distributions and *H-D* curves. The random parameter approach and linear prediction theory provide effective tools for carrying it out (Lappi and Bailey 1988, Lappi 1991).

The aims of this study were

- To develop tools for producing a stand description. These tools include methods for predicting diameter distribution and models for the *H-D* relationship. The tools should be able to utilize different kinds of information measured at different levels of accuracy. Special attention is paid to the effective use of sample tree information. (I – IV.)
- To utilize the tools developed in order to construct a new system for producing stand description in forest planning in Finland (V).
- To utilize the system in the optimization of data collection in an inventory by compartments (V).

2 COMPONENTS OF STAND DESCRIPTION – A LITERATURE REVIEW

Stand description includes the information of the stand measurements in such a form that it can be used in stand simulation. An example of a very coarse description includes the name of the main tree species while a very detailed description may include a complete tree map of the stand with measured tree taper. However, stand description is always a simplification of the reality and the extent of simplification depends on the purpose of the stand description and the data available. The stand description of this study includes the stock density, diameter distribution and *H-D* curve.

The stock density is described either by the number of stems or by the basal area. In the Finnish system, the basal area is used because it can be measured with higher accuracy than the stem number and it is more strongly correlated with the total volume than the number of stems is. The measurement of the basal area of a sample plot is straightforward with an angle gauge and the basal area of the stand is calculated as the mean of the plot wise basal areas. The measurements of diameter distribution and the *H-D* relationship, on the other hand, are rather laborious to carry out in the field. Thus, these components are not measured in the field; instead, approaches for predicting them from stand and sample tree measurements are used. The next two subsections consider the approaches reported in the relevant literature.

2.1 Diameter distribution

2.1.1 Approaches

The diameter distribution is the basis of the stand description. Many approaches have been used to construct the diameter distribution of a stand. In the following, the approaches are divided into three main approaches: (i) those based on a sample of diameters, (ii) those based on prediction or recovery of the parameters of an assumed theoretical distribution model and (iii) those based on known diameter distributions of similar stands (imputation methods).

Approach (i). The most natural way to obtain the diameter distribution is to measure a diameter sample from the stand. If the sample is large enough, it can be used as such in the simulation (e.g. Pienaar and Harrison 1988, Nepal and Somers 1992, Tang et al. 1997). If the sample is small, it can be smoothed (e.g. Droessler and Burk 1989, Uuttera and Maltamo 1995) or a theoretical distribution function can be fitted to it (e.g. Bailey and Dell 1973, Zarnoch and Dell 1985, Van Deusen 1986, Shiver 1988, Lindsay et al. 1996, Zhou and McTague 1996, Scolforo et al 2003, Zhang et al. 2003). The theoretical distribution function can be fitted using the maximum likelihood method, the method of moments, methods based on linear regression or by utilizing properties of certain percentiles or stand variables.

Approach (ii). The measurement of a diameter sample is too time-consuming in many inventories. In these cases, the diameter distribution can be predicted with some easily measurable stand variables. Traditionally, these methods are divided into parameter prediction methods (PPM) and parameter recovery methods (PRM) (Hyink and Moser 1983). In the parameter prediction method, the parameters of the assumed distribution function are predicted with some measured stand variables using estimated regression models (e.g. Schreuder et al. 1979, Little 1982, Rennolls et al. 1985, Kilkki and Päivinen 1986, Kilkki et al. 1989, Maltamo 1997, Siipilehto 1999, Temesgen 2003, Robinson 2004). In the PRM the parameters are recovered from some stand variables using known relations between the stand variables and distribution parameters (e.g. Ek et al. 1975, Burk and Newberry 1984, Magnussen 1986, McTague and Bailey 1987, Kuru et al. 1992). The stand variables used in PRM may be, for example, percentiles or moments of the diameter distribution. The recovery is, however, possible only for as many parameters as there are measured stand variables that are linked with the diameter distribution. If the number of parameters is greater, partial recovery can be used, i.e. as many variables as possible are recovered and the other parameters are predicted (e.g. Kilkki and Päivinen 1986, Kangas and Maltamo 2000b, I, II, V).

Approach (iii) The third approach is to use known diameter distributions of similar stands as the predicted diameter distribution of the stand (e.g. Haara et al. 1997, Maltamo and Kangas 1998). These methods have also been called imputation methods (Ek et al. 1997, Temesgen 2003, Temesgen et al. 2003). The similar stands are selected from a neighborhood that is defined with a distance function. The imputation methods used in predicting the diameter distribution are the *k*-nearest-neighbor method (Altman 1992) and the most similar neighbor method (Moeur and Stage 1995). The approach of Nanos and Montero (2002), where an interpolated surface is used to carry information from geographical neighbors to the target stands, also belongs to this class of approaches.

The above approaches can also be combined. For example, the neighbors of the third approach may be smoothed diameter distributions instead of true distributions, thus combining approaches (i) and (iii) (Maltamo and Kangas 1998). Furthermore, sample information can be used to improve a predicted distribution; examples of this are the Bayesian approach of Green and Clutter (2000) where the prior information of neighboring stands (iii) is combined with sample information (i) and the approach of Paper I in the present thesis where sample information (i) is used to improve predicted diameter distribution (ii). The approach of Maltamo et al. (2003a) combining empirical distributions of large trees identified from a digital video imagery with predicted distribution of small trees is a combination of (i) and (ii).

2.1.2 Distribution families

No theoretical results have been presented regarding which distribution family should be used as a diameter distribution. Hence, many distribution families have been used. The most important factors affecting the goodness of fit of a distribution family is the number of free parameters and the flexibility of the distribution to represent all possible distributional forms.

The most commonly used distributional family is the Weibull distribution (Bailey and Dell 1973), either in a three parametric form with parameters for location, scale and shape, or in a two-parametric form, where the location parameter has been given a value of zero, i.e. the minimum diameter of the stand is assumed to be zero. An unrealistic property of the Weibull distribution is that it has no upper limit, i.e. the maximum diameter of a stand is infinity. This problem can be solved by truncating the distribution at a point that is regarded as the maximum possible diameter of the tree species. Another possibility is to use a reversed truncated Weibull distribution, so that the location parameter receives an interpretation of maximum diameter and the minimum diameter is zero (e.g. Kuru et al. 1992, Robinson 2004). Other distribution families used are, for example, Gram-Charlier (Cajanus 1914), normal (e.g. Nanang 1998), lognormal (Bliss and Reinker 1964), gamma (Nelson 1964), beta (e.g. Hafley and Schreuder 1977), the Johnson's distribution families (e.g. Zhou and McTague 1996), the Chaudhry-Ahmad family (Chaudhry and Ahmad 1993, Nanos and Montero 2001) and the exponential distribution (e.g. Cancino and Gadow 2002). Many of these families also need to be truncated in applications because of unlimited maximum and minimum diameters.

All the distribution families presented below are, however, too rigid in some stands. This happens, for example, in stands where the shape of the distribution is irregular or multimodal. Thus, approaches that combine several distributions have been presented. These approaches are the segmented distribution approach (Cao and Burkhart 1984), finite mixture approach (e.g. Zhang et al. 2001) and the percentile-based approach (e.g. Borders et al. 1987).

The segmented distribution is constructed from pieces of the selected distribution [Weibull in Cao and Burkhart (1984)], and the number of pieces and locations of cutting points are fixed before the fitting procedure. In the finite mixture approach, the density function is a weighted sum of several density functions of the selected distribution family. In the reported studies, the number of distributions has been two and the theoretical distribution used has been either the Weibull distribution (Zhang et al. 2001, Liu et al. 2002) or the bivariate normal distribution (Zucchini et al. 2001). The percentile-based approach expresses the distribution using a fixed number of percentiles, corresponding to predefined values of the distribution function. The continuous function is obtained by interpolation, either by linear (e.g. Borders et al. 1987, I, II) or by spline interpolation (Maltamo et al. 2000, Kangas and Maltamo 2000c). When linear interpolation is used, the obtained distribution consists of pieces of a uniform distribution (I). Thus, it can be regarded either as a segmented distribution approach or as a finite mixture approach using uniform distribution (See Section 3).

2.1.3 Compatibility of diameter distribution

Compatibility of stand description means that all stand variables calculated from the diameter distribution are equal to their measured variables. It is a desired property of the stand description and incompatibility may indicate that the information of the stand measurements is not utilized effectively. In PRM, the compatibility can be guaranteed with respect to as many stand variables as there are parameters in the distribution family used. However, it usually leads to a very complicated set of equations, whose solution does not necessarily exist in closed form.

Compatibility of basal area weighted and unweighted diameter distributions can be guaranteed through size-biased distribution theory (Gove and Patil 1998, Gove 2000, Gove 2003), where the relationships between the ordinary and basal area weighted forms of the diameter distribution are derived analytically. It makes it possible to derive the parameters of the ordinary distribution from the parameters of the basal area weighted distribution and vice versa. Thus, it provides tools for the recovery of the parameters of weighted distributions.

In PPM, the compatibility is very hard to ensure. Thus, many studies have presented algorithms that aim at a compatible stand description by adjusting the diameter distribution (Nepal and Somers 1992, Cao and Baldwin 1999, Kangas and Maltamo 2000a). In these algorithms, the adjustment is applied to the frequencies of the stand table and no new diameter classes are established in the adjustment. Furthermore, the stand variables calculated using the adjusted frequencies are required to equal the measurements exactly. These requirements may be too strict in practice, where measurements include errors. An alternative for these approaches is presented in Paper II.

2.2 Height-diameter models

2.2.1 Approaches

The *H-D* model is used to predict heights for trees with given diameters. The *H-D* relationship varies considerably between stands (Lappi 1997, Hökkä 1997, Jayaram and Lappi 2001, Eerikäinen 2003, Calama and Montero 2004, III, IV) and accuracy of the predicted *H-D* curve has a considerable effect on the accuracy of the stand volume estimate. As with diameter distributions, the approach used in height prediction depends on the data available and the possible approaches could be classified in a manner similar to the classification of the previous section, i.e. into (i) approaches based on a sample of heights, (ii) approaches predicting the parameters of the *H-D* curve without height measurements and (iii) approaches based on imputation. The studies using approaches belonging to the second (ii) category can further be divided into two classes: (ii-a) approaches which assume that a large forest area can be divided into stands, each of which has its own *H-D* curve, and (ii-b) approaches where a common *H-D* curve is assumed for larger areas, for example for states or regions.

Approach (i). The most natural approach in predicting the *H-D* curve of a stand is to fit an assumed curve to observed *H-D* data from the target stand. There are numerous studies concerning fitting different kinds of curves on observed *H-D* data (Curtis 1967, Omule and McDonald 1991, Arabazis and Burkhart 1992, Flewelling and de Jong 1992, Fang and Bailey 1998). The aim of these studies has been to find an appropriate functional form for the *H-D* curve and to show how the estimation should be done. A recent study of Zhang et al. (2004), which models the spatial variation within stand using geographically weighted regression, provides a new view to this approach.

Approach (ii-a). The variation of *H-D* curves between stands is usually taken into account by modeling the parameters of the *H-D* curve using stand specific variables as predictors (Veltheim 1987, Borders and Patterson 1990, Parresol 1992, Lynch and Murphy 1995, Fang and Bailey 1998, Knowe et al. 1998, Wang and Hann 1998, Hanus et al. 1999, Zeide and Vanderschaaf 2002). Furthermore, these models have been localized using measured heights, for example by re-scaling the model so that the measured mean height is obtained when the diameter equals the measured mean diameter, thus combining approach (ii-a) with approach (i). However, even though this approach of localization works rather well when the mean diameters and heights are accurate, it does not take into account the within-stand and between-stand variances of tree heights. In recent studies, in addition to the use of stand-specific predictors, the hierarchy of the data has been taken into account through a mixed model approach. It provides an effective and theoretically justified means for localizing the curves for a given stand using measured height(s) (Lappi 1997, Hökkä 1997, Jayaram and Lappi 2001, Calama and Montero 2004). The localization is possible even using just one measured height and the ratio of within-stand and between-stand variances determines how close the expected curve is to the measured height. The mixed model approach also provides natural approaches for taking into account the temporal development of the *H-D* curves (Lappi 1997, Eerikäinen 2003, III, IV).

Approach (ii-b). In models belonging to this approach, single values of the parameters of an assumed *H-D* model are estimated from a large dataset consisting of measurements from several locations of the target area (Huang et al. 1992, Zhang 1997, Peng 1999, Huang et al. 2000, Zhang et al. 2002, Colbert et al. 2002). When the model is used for prediction, the estimated parameter values are applied to the whole target area.

Approach (iii). With the exception of Maltamo et al. (2003b), imputation methods have not been used in the prediction of *H-D* curves, even though this could be done concurrently with the imputation of diameter distributions. The reason for this seems to be data related rather than methodological: the datasets used in the imputation of diameter distributions have not included tree heights.

2.2.2 Model forms

Many studies carried out on *H-D* curves have compared different functional forms for the *H-D* relationship. The number of functional forms tested exceeds 30 and no single form has been found to be superior. However, some forms have been among the best ones in many comparisons either in a linearized or nonlinear form. The three most frequently used functions are the allometric function, also called the power function (Greenhill 1881, Curtis 1968, Zeide and Vanderschaaf 2002, Eerikäinen 2003, Zhang et al 2004)

$$H=aD^b\ , \qquad (1)$$

Meyer's equation (Meyer 1940, Stout and Shumway 1982, Farr et al. 1989, Huang et al. 1992, Fulton 1999)

$$H = a\left(1-e^{-bD}\right) \qquad (2)$$

and the Korf curve, also called the Lundqvist or exponential function

$$H=ae^{-bD^{-c}}\ , \qquad (3)$$

where H is tree height, D is diameter and a, b and c are parameters.

Based on the biological growth pattern of a tree, Yuancai and Parresol (2001) recommended the use of functions with an inflection point, which means that the curve is S-shaped, and has an upper asymptote.

The allometric equation lacks both these properties. However, it has a strong mechanical basis: if the exponent b is given a value of 2/3, the stem is equally resistant to bending at different heights (Greenhill 1881, p. 66-73; see Zeide and Vanderschaaf 2002).

The Meyer equation has an upper asymptote of a, but it lacks an inflection point. It can be expanded to the Weibull type function by adding a positive power parameter to D, and to the Chapman-Richards type function by adding a power parameter to the whole expression in the parentheses. These expansions naturally also have an upper asymptote and the Chapman-Richards type of function also has an inflection point. Both of these expansions have been recommended and used as *H-D* curves, the former by Huang et al. (1992), Zhang (1997) and Ishii et al. (2000) and the latter by Huang et al. (1992), Zhang (1997) and Zhang et al. (2002).

The Korf curve has both an asymptote and an inflection point. The Korf curve has been used both in the form where c=1 (Curtis 1967, Zakrzewski and Bella 1988, Arabazis and Burkhart 1992, Calama and Montero 2004) and with other positive values of c (Huang et al. 1992, Lynch and Murphy 1995, Lappi 1997, Hökkä 1997, Jayaram and Lappi 2001, Colbert et al. 2002, III, IV). In addition to these functions, the Schnute function (Schnute 1981), which has both an inflection point and an upper asymptote, has been recommended in many studies (Huang et al. 1992, Zhang 1997, Yancai and Parresol 2001).

2.3 Other components

The diameter distribution and *H-D* curve of a stand are regarded as the most important components of stand description in forest management planning. Other important components are, for example, taper curves, spatial pattern and age distribution. It is well known that tree taper varies from stand to stand (Lappi 1986, Ojansuu 1993) and spatial patterns may be very different in different stands (Lin 2003). In the future, information about tree taper may be obtained from harvester measurements by using imputation methods and the spatial pattern may be determined by using high resolution remote sensing imagery (e.g. Uuttera et al. 1998). However, these approaches are not currently in use, and field measurements of these stand characteristics are too time consuming in inventories for forest management planning. Thus, the taper curves of Laasasenaho (1982) are assumed to apply to all stands and the spatial pattern within stands is assumed to be random.

The diameter distribution and *H-D* curve produce a static description of the stock structure, which is the basis of growth and yield prediction. The next important component of the stand description is a set of models that predicts the development of the stand (e.g. Hynynen et al. 2002). However, because the estimation of stand growth is out of the scope of this study, growth models will not be discussed any further.

2.4 Other approaches

An alternative to the separate prediction of diameter distribution and *H-D* curve is the use of a bivariate distribution as the joint distribution of heights and diameters. Ever since Schreuder and Hafley (1977) proposed it, Johnson's S_{BB} distribution has been widely utilized (Hafley and Buford 1985, Knoebel and Burkhart 1991, Siipilehto 1996, Tewari and Gadow 1999, Siipilehto 2000). Zucchini et al. (2001), however, found that a mixture of two bivariate normal distributions, which has more free parameters than the bivariate Johnson's S_{BB} distribution, fitted their Central European beech stand better. Furthermore, the trivariate Johnson's S_{BBB} distribution has been used in the estimation of the joint distribution of height, diameter and age (Schreuder et al. 1982). However, although these approaches are theoretically appealing, their utility is not self-evident when compared to the approach based on a univariate diameter distribution and the *H-D* curve (Knoebel and Burkhart 1991, Siipilehto 1996, 2000). Furthermore, a bivariate joint distribution of heights and diameters is obtained also by using the estimated diameter distribution and error distribution of the *H-D* curve.

Because diameter measurements are much easier to obtain than height measurements, the current practice is to predict tree height from its diameter. Laser scanning (e.g. Maltamo et al 2004b) and digital photogrammetry of trees from aerial photographs (Korpela 2004), which are promising alternatives to the inventory by compartments in the future, provide quite accurate height measurements, while diameter measurements are hard to obtain from the air. Thus, the development of these methods to realistic approaches in forest inventories may reverse the roles of height and diameter in the future (Maltamo et al. 2004b).

3 PERCENTILE-BASED DIAMETER DISTRIBUTION

This study uses the percentile-based approach in the prediction of diameter distributions. The percentile-based approach of diameter distributions was first presented by Borders et al. (1987) and has later been used in Borders and Patterson (1989), Maltamo et al. (2000), Kangas and Maltamo (2000b) and Eerikäinen and Maltamo (2003). The percentile-based diameter distribution was introduced because of its ability to reproduce multi-modal stand tables and the simplicity of the mathematics needed (Borders et al. 1987). More generally, it is much more flexible than the traditional parametric distributional families, e.g. Weibull and Johnson's S_B, because the implied assumptions about the form of the diameter distribution are weak (Maltamo et al. 2000, Kangas and Maltamo 2000b).

This study regards the percentile-based distribution as a piecewise defined uniform distribution. The interpretation of the residuals of the percentile models as the horizontal errors of the percentiles is emphasized, and methods for effective use of the error variances in calculations are developed (I, II, V). Furthermore, it is shown that measured sample order statistics can be regarded as measured percentiles, which can be plotted onto the figure of the c.d.f. and used in localizing the predicted percentiles for a given stand (I, V). Finally, exact analytical formulas for relationships between different stand variables are derived without transforming the distribution to a stand table. The formulas are utilized in adjusting the predicted percentiles in order to ensure compatibility of the stand description (II, V). The next five subsections summarize and complete the properties of the percentile-based approach derived in Papers I, II and V.

3.1 Distribution function and density

Let the diameter distribution of a stand be described with a strictly increasing vector of diameters, $\mathbf{d}=(d_1,d_2,\dots,d_k)'$, corresponding to predefined values, $\mathbf{p}=(p_1,p_2,\dots,p_k)'$, of the cumulative distribution function (c.d.f.) where $p_1=0$ and $p_k=1$. Assuming that the c.d.f. between consecutive percentiles is linear, the c.d.f. of tree diameter Y is (I):

$$F_{\mathbf{p}}\left(y|\mathbf{d}_m\right)=\begin{cases}0 & y<d_1\\ a_i+b_i y & d_i\le y<d_{i+1}, \text{for } i=1,\dots,k-1\\ 1 & y\ge d_k\end{cases}, \quad (4)$$

where

$$b_i=\frac{p_{i+1}-p_i}{d_{i+1}-d_i} \text{ and } a_i=p_i-b_i d_i .$$

(The capital letter Y is used for random variable and the lower case letter y for its realization; a notation that is commonly used in statistical literature.) The notation $F_{\mathbf{p}}(y|\mathbf{d}_m)$ emphasizes that vector $\mathbf{p}$ is a predefined constant vector that defines the distribution family used and vector $\mathbf{d}_m$ includes the parameters of the distribution in stand m. The density is obtained by differentiating (4):

$$f_{\mathbf{p}}\left(y|\mathbf{d}_m\right)=\begin{cases}0 & y<d_1\\ b_i & d_i\le y<d_{i+1}, i=1,\dots,k-1\\ 0 & y\ge d_k\end{cases}. \quad (5)$$

Table 1. Equations for the calculation of some stand characteristics from the percentile-based distribution of form (4).

	weighted	unweighted
unweighted mean	$\dfrac{\sum_{i=1}^{k-1} b_i\left(\ln d_{i+1} - \ln d_i\right)}{\sum_{i=1}^{k-1} b_i\left(\frac{1}{d_i} - \frac{1}{d_{i+1}}\right)}$	$\frac{1}{2}\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,2} - d_i^2\right)$
weighted mean	$\frac{1}{2}\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,2} - d_i^2\right)$	$\frac{3}{4}\dfrac{\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,4} - d_i^4\right)}{\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,3} - d_i^3\right)}$
unweighted median	$d \left\vert \dfrac{1}{\sum_{i=1}^{k-1} b_i\left(\frac{1}{d_i} - \frac{1}{d_{i+1}}\right)} \int_{d_1}^{d} \frac{f^G(t)}{t^2} dt = \frac{1}{2} \right.$	$d_{50\%}$
weighted median	$d_{50\%}$	$d \left\vert \dfrac{3}{\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,3} - d_i^3\right)} \int_{d_1}^{d} t^2 f^N(t) dt = \frac{1}{2} \right.$
quadratic mean	$\sqrt{\dfrac{1}{\sum_{i=1}^{k-1} b_i\left(\frac{1}{d_i} - \frac{1}{d_{i+1}}\right)}}$	$\sqrt{\dfrac{\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,3} - d_i^3\right)}{3}}$
stem number/basal area	$\sum_{i=1}^{k-1} \frac{4b_i}{\pi}\left(\frac{1}{d_i} - \frac{1}{d_{i+1}}\right)$	$\dfrac{12}{\sum_{i=1}^{k-1} \pi b_i\left(d_{i+1}^{\,3} - d_i^3\right)}$

Note. The equations for the basal area weighted diameter distribution are on the left and for the unweighted diameter distribution on the right. The vertical lines in the equations of median mean the value of d that satisfied the equation on the right hand side of the vertical line. The notation f^G means the density of the weighted diameter distribution and f^N the density of the unweighted distribution.

3.2 Moments of the distribution

The expectation or mean of tree diameter in a stand is calculated as the expectation of the percentile-based diameter distribution (Eqs 4 and 5). It follows straightforwardly from the definition of the expected value that (Casella and Berger 2002, p. 55)

$$
\begin{aligned}
E(Y) &= \int_{d_1}^{d_k} y f_{\mathbf{p}}\left(y \middle| \mathbf{d}_m\right) dy \\
&= \sum_{i=1}^{k-1} \int_{d_i}^{d_{i+1}} y b_i dy = \frac{1}{2}\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,2} - d_i^2\right)
\end{aligned}
\quad . \tag{6}
$$

For calculating variance, the second moment of the distribution, defined as $E(Y^2)$, is needed (Casella and Berger 2002, p. 59). It follows again from the definition of the expected value that

$$
\begin{aligned}
E(Y^2) &= \int_{d_1}^{d_k} y^2 f_{\mathbf{p}}\left(y \middle| \mathbf{d}_m\right) dy \\
&= \sum_{i=1}^{k-1} \int_{d_i}^{d_{i+1}} y^2 b_i dy = \frac{1}{3}\sum_{i=1}^{k-1} b_i\left(d_{i+1}^{\,3} - d_i^3\right)
\end{aligned}
\quad , \tag{7}
$$

and the variance is obtained using the well-known formula (Casella and Berger 2002, p. 60)

$$\mathrm{var}(Y) = E\left(Y^2\right) - \left[E(Y)\right]^2. \tag{8}$$

3.3 Relationships between percentiles and stand characteristics

Many computations with the theoretical distributions used in the description of forest structure are very complicated and lead to results that do not exist in closed form. A usual solution to this problem is to use a stand table approximation of the distribution (e.g. Nepal and Somers 1992). With the percentile-based diameter distribution (Eqs 4 and 5), many stand characteristics can be rather simply derived analytically (See the Appendix of Paper II). The most important stand characteristics are given in Table 1 for both basal-area weighted and unweighted diameter distributions.

3.4 Order statistics

The theory of order statistics is of great importance with the percentile-based diameter distribution because a measured sample order statistic is an unbiased estimator of a certain percentile of the underlying distribution (I).

Denote the r^{th} order statistic in a sample of size n by $Y_{r:n}$. If the sample is drawn from a distribution with a c.d.f. of the form (4), the exact distributions of order statistics can be derived rather easily.

The density of $Y_{r:n}$ is

$$f_{r:n}(y) = \begin{cases} \beta_0 b_i \left(a_i + b_i y\right)^{r-1} \left(1 - a_i - b_i y\right)^{n-r} & \text{if } d_i \le y < d_{i+1} \text{ for } i = 1,\dots,k-1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and the joint density of two order statistics $Y_{r1:n}$ and $Y_{r2:n}$ is

$$f_{r:n}(y_1,y_2) = \begin{cases} \beta_1 b_i b_j \left(a_i + b_i y_1\right)^{r_1-1} \left(a_j + b_j y_2 - a_i - b_i y_2\right)^{r_2-r_1-1} \left(1 - a_j - b_j y_2\right)^{n-r_2} & \text{if } d_i \le y_1 < d_{i+1};\ d_j \le y_2 < d_{j+1};\ y_1 \le y_2 \text{ for } i,j = 1,\dots,k-1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where

$$\beta_0 = \frac{n!}{(r-1)!(n-r)!} \quad \text{and} \quad \beta_1 = \frac{n!}{(n-r_2)!(r_2-r_1-1)!(r_1-1)!}$$

(Reiss 1989, I).

The expectation and variance of a single order statistic follow from the same definitions and general rule that were used in Equations 6, 7 and 8 (Casella and Berger 2002, p. 55-60, I). The covariance of two order statistics can be calculated using the corresponding definitions and rule (Casella and Berger 2002, p. 144, 170) for bivariate distribution, as shown in Paper I. Furthermore, the measured sample order statistic is an unbiased estimator of the $100p^{\text{th}}$ percentile of the diameter distribution of the stand, where

$$p = F_{\mathbf{p}}\left[E(Y_{r:n})|\mathbf{d}_m\right], \tag{11}$$

and $F_{\mathbf{p}}$ is the c.d.f. of the stand.

3.5 Considerations on the PPM with the percentile-based approach

Assume that the percentiles of the diameter distribution in stand m follow the model

$$\mathbf{d}_m = \mathbf{B}\mathbf{x}_m + \mathbf{e}_m, \tag{12}$$

where **B** includes the parameters, $\mathbf{x}_m$ the predictors and $\mathbf{e}_m$ the residual errors.

Assuming that the model is correct and **B** is known, the fixed part of the model, $\mathbf{Bx}_m$, gives the conditional expectations of the percentiles in the stand given $\mathbf{x}_m$,

$$\mathbf{Bx}_m = E(\mathbf{d}|\mathbf{x}_m), \tag{13}$$

and the vector $\mathbf{e}_m$ includes the horizontal deviations of the true percentiles of stand *m* from their conditional expectations or, in other words, stand effects (Figure 1 of I). Thus, the variance-covariance matrix of the residuals includes information about the between-stand variation of the percentiles, which is useful information in adjusting and localizing the percentiles (I, II, V).

4 DATASETS OF THIS STUDY

Two different datasets were used in the study. The first data, called mixed forest data, are a small data including 43 fixed rectangular plots from conifer mixtures of North Karelia. This dataset was used in the tests of the methods developed for diameter distribution prediction (I, II). The second dataset (INKA-data) is a larger dataset covering the whole of Finland. It includes remeasured plots from 757 stands. This dataset was used in the estimation of *H-D* curves (III, IV) and models for accuracy of stand structure prediction (V).

4.1 Mixed forests data (I, II)

The mixed forests data were originally collected for estimation of individual tree growth models (Pukkala et al. 1998). The plots were measured from mixed Scots pine-Norway spruce stands. Almost all stands are naturally regenerated and all stands represent the medium site fertility class [*Myrtillus* type in the classification system of Cajander (1926)]. The plot size varies between 600 and 3000 m^2, being smaller in dense stands than in open forests. The stands were selected to represent forests of different density, age, tree size, species composition, size difference between the two species and spatial distribution of trees in the stand. The stands are regarded as even-aged although there may be remarkable variation in the ages of the different tree species strata. However, the within-stand variation of a given tree species is small. The conifer mixtures of this kind are quite typical in Finland and they are considered even-aged forest stands in forest planning and management. Detailed size and growth measurements of all trees were made. However, this study used only the measured diameters (i.e. the empirical diameter distribution), stand age and site fertility. The minimum, maximum and mean values of the most important characteristics of the data are presented in Table 2 of Paper II.

4.2 INKA-data (III, IV, V)

The stands of the INKA-data (Gustafsen et al. 1988, Hynynen et al. 2002) are a subsample of the stands of the 7th National Forest Inventory in Southern Finland and of the 6th National Forest Inventory in Northern Finland. Only stands on mineral soils were included and sapling stands were excluded. The plots were established between the years 1976-1982 and they were remeasured twice with five-year intervals. The dominant tree species of the stands were Scots pine, Norway spruce or birch. Only healthy, single-storied stands with the proportion of major tree species being at least 50% of the total volume of the growing stock were included. The Scots pine dominated stands represent all fertility classes of mineral soils except the barest sites, which are unimportant for forest economy. The spruce and birch-dominated stands represent only stands with at least medium fertility.

A cluster of sample plots was established on each stand. The cluster included three fixed-radius circular plots, located 40 meters apart from each other (See Figure 1 of IV). The plot size varied according to the stand density so that the total number of sampled trees in a stand was at least 120 in Southern Finland and 100 in Northern Finland. The diameter at breast height was recorded from all trees of the plot. A smaller plot of more detailed measurements was established at the center of each plot, comprising 1/3 of the area of the sample plot. These measurements included, among others, sample tree heights.

The minimum, maximum and mean values of the most important characteristics of the data are given in Table 1 of Paper III, Table 1 of Paper IV and Table 1 of Paper V. Note that these tables are calculated from the sub-data used in these studies, including those stands of the original INKA-data that fulfill the requirements stated in these studies.

5 Methods

5.1 Mixed models (III, IV)

In the following a very short overview on mixed models is given through a forestry example. In many forestry problems, the data have a hierarchical structure that is due to the division of the forest area to several stands and plots. Furthermore, the plots may have been measured several times (e.g. Hökkä and Ojansuu 2004, I and II). Essentially, the stands, plots and measurement occasions of the model data are samples from populations of stands, plots and times and the model will be applied in new stands, plots and times that are not present in the modeling data. The mixed model provides a natural approach for these kinds of situations (see Davidian and Giltinan 1995, p. 63-124; McCulloch and Searle 2001, p. 1-27; Diggle et al. 2002, p. 169-189). Gregoire et al. (1995) summarizes well the merits of this approach in a situation where the structure of the data causes spatial and temporal autocorrelations between observations.

A linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}\,. \qquad (14)$$

The first term on the right hand side of (14) is called the fixed part and the last two terms the random part of the model. Matrix $\mathbf{X}$ is the design matrix of the fixed part, $\mathbf{Z}$ the design matrix of the random part, $\boldsymbol{\beta}$ the vector of fixed parameters, $\mathbf{b}$ the vector of random parameters and $\mathbf{e}$ the vector of residual errors (Lappi 1993, p. 133-153, McCulloch and Searle 2001, p. 156-163, Pinheiro and Bates 2000, p. 58-62). The parameters to be estimated are the fixed parameters, $\boldsymbol{\beta}$, and variance-covariance matrices of random parameters, var($\mathbf{b}$)=$\mathbf{D}$, and of residual errors, var($\mathbf{e}$)=$\mathbf{R}$. The parameters can be estimated, for example, with the restricted maximum likelihood method (REML) (see Pinheiro and Bates 2000, p. 75-79).

The fixed parameters of the model correspond to the parameters of an ordinary regression model and the fixed part gives the conditional expectation of the response variable $\mathbf{y}$ given the fixed predictors. The structure of the data determines the structure of $\mathbf{D}$ and $\mathbf{R}$. In our example, a block-diagonal structure of matrix $\mathbf{D}$ is assumed, which implies that trees of the same stand and plot have constant correlations, as have the trees from the same stand and different plots (see e.g. Pinheiro and Bates 2000, p. 60-62). Trees from different stands are assumed to be uncorrelated. Matrix $\mathbf{R}$ has usually a diagonal structure and if residuals have equal variance, it is a multiple of an identity matrix. However, spatial or temporal autocorrelations, for example, lead to a non-diagonal matrix $\mathbf{R}$. For more details on modeling the data structure and autocorrelations using matrices $\mathbf{D}$ and $\mathbf{R}$, see Pinheiro and Bates (2000, p. 201-270).

The variance-covariance matrix of the random part is

$$\mathbf{V}=\mathrm{var}(\mathbf{y})=\mathbf{ZDZ}'+\mathbf{R}\,. \qquad (15)$$

Equation (15) shows that the random parameters imply correlations between observations. Furthermore, it implies that the covariance between the observations and random parameters is $\mathrm{cov}[\mathbf{b},(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})']=\mathbf{DZ}'$ (McCulloch and Searle 2001, p. 255). This covariance can be utilized in the prediction of random effects using observations of the response variable.

5.2 Prediction of random variables (I, III, IV, V)

The error terms of statistical models are random variables. This study utilizes the theory of linear prediction to localize the models for a given stand by predicting these random variables with observations of the response from a sample. A short overview of the theory is given in this section.

Assume that we have a vector of random variables, $\mathbf{x}$, which can be divided into two parts

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{16}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are random vectors of length 1 or more. In the applications of this study, $\mathbf{x}_1$ is a vector of random effects (stand and time effects) and $\mathbf{x}_2$ is a vector of observed residuals (observed height or percentile minus its expected value). It is assumed that $E(\mathbf{x}_1)=\boldsymbol{\mu}_1$, $E(\mathbf{x}_2)=\boldsymbol{\mu}_2$, $var(\mathbf{x}_1)=\mathbf{V}_1$, $var(\mathbf{x}_2)=\mathbf{V}_2$, and $cov(\mathbf{x}_1,\mathbf{x}_2') = \mathbf{V}_{12}$. Using the notation of McCulloch and Searle (2000, p. 247), this can be written as

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_{12} \\ \mathbf{V}_{12}' & \mathbf{V}_2 \end{bmatrix} \right). \tag{17}$$

Assume that we have observed random vector $\mathbf{x}_2$ and we want to predict vector $\mathbf{x}_1$. The best predictor (BP) of $\mathbf{x}_1$ is the conditional expectation

$$BP(\mathbf{x}_1) = E(\mathbf{x}_1|\mathbf{x}_2) \tag{18}$$

(McCulloch and Searle 2001, p. 248). The best predictor can usually not be calculated because it requires the distribution of $\mathbf{x}_1|\mathbf{x}_2$. An estimator that requires only first and second moments is obtained by limiting the consideration to linear unbiased predictors. The Best Linear Unbiased Predictor (BLUP) of $\mathbf{x}_1$ is

$$BLUP(\mathbf{x}_1) = \hat{\mathbf{x}}_1 = \boldsymbol{\mu}_1 + \mathbf{V}_2\mathbf{V}_{12}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{19}$$

with the prediction variance of

$$var(\hat{\mathbf{x}}_1 - \mathbf{x}_1) = \mathbf{V}_1 - \mathbf{V}_{12}\mathbf{V}_2^{-1}\mathbf{V}_{12}' \tag{20}$$

(McCulloch and Searle 2001, p. 250). Thus, if the expectations and variance-covariance matrices of two random vectors are known and either of them is observed, the other can be predicted. The variance of the prediction error can be calculated using Equation (20). If $\mathbf{x}$ follows the multinormal distribution, BLUP is also BP.

The standard theory of mixed models utilizes BLUP to predict the realizations of the random effects in the modeling data (McCulloch and Searle 2001, p. 254-258). It is a special case of prediction, where the correlations between the observed random vector, $\mathbf{x}_2=\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$, and the vector of random effects, $\mathbf{x}_1=\mathbf{b}$, are generated by the structure of the data (Eq. 15), i.e. $\mathbf{V}_1=\mathbf{D}$ $\mathbf{V}_2=\mathbf{ZDZ'}+\mathbf{R}$ and $\mathbf{V}_{12}=\mathbf{DZ'}$. If measurements of the response variable from a new stand are available in applications, the realizations of the random parameters in that stand can be predicted (Lappi and Bailey 1998, Lappi 1991, III). Similarly, if we have several models that have correlated residual errors and the correlations are known, an observed response of any single model can be utilized in predicting the responses of the other models in that stand (Lappi 1991, I). Thus, the realized random effects and/or random errors of the models can be predicted in order to localize the models for that stand.

In practice, the variance-covariance matrices used are replaced with their estimates. A predictor obtained using these estimates is sometimes called EBLUP, Estimated Best Linear Unbiased Predictor (e.g. McCulloch and Searle 2001, p. 257).

5.3 Constrained optimization (II, V)

This study utilizes constrained optimization in adjusting the percentiles and stand variables to obtain a compatible stand description (II, V) and in searching for an optimal measurement strategy for a single stand (V). A general constrained optimization problem is defined as follows (Bazaraa and Shetty 1979, p. 2):

minimize

$$f(\mathbf{z}) \tag{21a}$$

subject to

$$g_i(\mathbf{z}) \le 0 \quad \text{for } i=1,\ldots,m \tag{21b}$$

$$h_j(\mathbf{z}) = 0 \quad \text{for } j=1,\ldots,l, \tag{21c}$$

where $\mathbf{z}=(z_1,z_2,\ldots,z_n)'$ is a vector of decision variables and f, $g_1,\ldots,g_m$ and $h_1,\ldots,h_l$ are functions of $\mathbf{z}$. The function f is called an objective function, functions $g_1,\ldots,g_m$ inequality constraints and functions $h_1,\ldots,h_l$ equality constraints. The solution to the above problem is a value of $\mathbf{z}$ that minimizes the function f and meanwhile satisfies the constraints. If functions f, $g_1,\ldots,g_m$ and $h_1,\ldots,h_l$ are linear in $\mathbf{z}$, the problem is a constrained linear optimization problem and if any of them is nonlinear in $\mathbf{z}$, the problem is nonlinear.

A considerable amount of research has been carried out to develop algorithms for solving optimization problems and modern software packages include functions for doing it. The adjustment problem (II, V), which has both a nonlinear objective function and nonlinear constraints, was solved using IMSL subroutine NCONF. It is based on successive quadratic programming (IMSL 1997). The problem used to search for an optimal measurement strategy for a forest stand has a nonlinear objective function and linear constraints. It was solved in the R-environment (R Development Core Team 2003) with the function ConstrOptim using the simplex method of Nelder and Mead (1964).

6 Summary of results

6.1 Localizing predicted diameter distribution with sample order statistics (I, V)

If one has a diameter sample of size n from a stand, the diameter of the r^{th} smallest tree in the sample is a measurement of the r^{th} order statistic of the sample. Measuring an order statistic in the field requires only the diameter measurement of one single tree and the knowledge whether the diameter of the other trees is larger or smaller than the measured diameter. Thus, the measurement could be carried out rather rapidly, even though practical studies on the measurements have not been carried out until now.

Paper I showed that a measured sample order statistic of an HPS-plot is a measurement of some percentile of the underlying basal area diameter distribution (see section 3.4). Correspondingly, if the sample is from a fixed plot, a sample order statistic is a measurement of some percentile of the ordinary (unweighted) diameter distribution. Furthermore, Paper I presented an algorithm for combining the information of expected percentiles of the diameter distribution (Eqs 12 and 13) with the information of sample order statistics. The algorithm is based on predicting the stand effects of a percentile-based diameter distribution model (Eq. 12) with BLUP. It utilizes the estimated variance-covariance matrix of stand effects and the sampling errors of the order statistics.

The algorithm was examined with the mixed forests data in Paper I and with the INKA-data in Paper V. The percentiles were predicted using the models of Kangas and Maltamo (2000b). An assumption behind the method is that the diameters of the sample plot are independently sampled from the diameter distribution of the stand. The experiment of Paper I was planned in such a manner that this assumption was valid. Thus, trees belonging to the simulated HPS-plots were randomly selected from among all trees of the stand and tree locations were not utilized. In reality, however, trees of a sample plot constitute the sample, which is independent and identically distributed only if the stand is spatially homogenous. In Paper V, the aim was to mimic a real inventory for forest management planning and the measured sample order statistics were obtained from true HPS-plots. In calculations, these plots were regarded as independent and identically distributed samples from the underlying diameter distribution, even though this assumption is violated in reality. Comparing the results of Papers I and V provides a view of how violating the assumption of independent and identically distributed sample affects the localized distributions.

Another difference between Papers I and V was that in Paper I the stand variables used in predicting the percentiles did not include sampling error, while in Paper V sampling error was included. These errors cause bias in the predictions of the percentiles and have a remarkable effect on their between-stand variance-covariance matrix. Under the assumption that the measurements follow a lognormal distribution, the bias-corrected predictions and a variance-covariance matrix of the stand effects that takes the measurement errors into account were derived in the Appendix of Paper V and used in localization.

A problem in Paper I was that because the measured median diameter was used as the predicted 50^{th} percentile and its variance was zero, the stand effects for the 50^{th} percentile were always zero. This caused peaks around the median diameter in the localized distributions. In Paper V, the use of a corrected variance-covariance matrix allowed also stand effects of the 50^{th} percentile to differ from zero, which solved the peak problem.

It was clearly seen that even one measured sample order statistic distinctly decreased the RMSE and absolute bias of volume and Reynolds' error index (Reynolds et al. 1988) (see Fig. 2 of I). In addition, increasing the number of measured sample order statistics improved the accuracy steadily. Furthermore, the method seemed to produce good predictions also in complex stands. The order statistics were clearly useful also in Paper V. This was seen in the decreasing RMSE of saw timber volume.

In Paper I, order statistics of independent samples improved the prediction of diameter distribution considerably, both when considered through RMSE of total volume and through error index. However, in Paper V, there was no longer a significant decrease in the RMSE of volume. This is partially because the diameters of a true HPS sample are not independent and identically distributed. Another reason is that the effect of measurement errors of stand variables overrides the effect of measured sample trees on the RMSE of volume. In Paper V, where stand variables included sampling error, the RMSE of Scots pine volume was, on average, 13.4, but in Paper I, where correct values were used, it was only 1.3.

6.2 Adjusting the predicted percentiles to obtain a compatible stand description (II, V)

Paper II proposed an algorithm for adjusting the diameter distribution in order to ensure its compatibility with stand variables. The primary aim of the study was to develop an algorithm that is able to utilize an inaccurate measurement of stem number in the prediction of diameter distribution. Instead of adjusting the stand table (Nepal and Somers 1992, Cao and Baldwin 1999, Kangas and Maltamo 2000a), the algorithm adjusts the predicted percentiles. This makes it possible to utilize the variances of the prediction errors in the adjustment and also to adjust minimum and maximum diameters, which is problematic in the adjustment of a stand table. In addition to prediction errors of percentiles, the algorithm takes into account the measurement errors of the stand variables. The obtained stand description is fully compatible, but the stand variables of the stand description may deviate from their measured values. The magnitude of the deviations depends on the variance of the measurement errors so that variables measured with low accuracy deviate more than variables measured with high accuracy.

The algorithm was tested in the mixed forests data using multinormal measurement errors of various magnitudes simulated to the true stand variables (basal area, *DGM* and stem number). When the measurement errors were low, the diameter distribution predictions with model 2 of Kangas and Maltamo (2000b) (which includes stem number as a predictor) were as accurate as those obtained by adjusting the predictions of model 1 of Kangas and Maltamo (2000b) (which does not include the stem number as a predictor). However, increasing the measurement errors made the accuracy of model 2 even poorer than that of model 1, while the adjustment algorithm improved the accuracy considerably. The adjustment algorithm was also superior when compared with an algorithm that adjusts stand table and does not utilize the measurement and prediction errors (Kangas and Maltamo 2000a).

Another test of the algorithm was carried out in Paper V, which showed that the use of stem number in an inventory by compartments does not affect the predictions of volume and saw timber volume significantly even if the measurement errors are taken into account. The reason is probably the same as in the localization, i.e. the errors of basal area and *DGM* dominate in practical forest inventory. Thus, an inaccurate measurement of stem number does not improve the accuracy of stand volume and saw timber volume. However, it may improve the accuracy of some other variables, for example, timber assortment proportions or stand growth, even though in the study of Kangas and Maltamo (2003), the accuracy of stand growth could not be improved considerably by the algorithm of Kangas and Maltamo (2000a).

The adjustment algorithm makes it also possible to use other stand variables than stem number in the adjustment by including them in the optimization as additional constraints. For example, arithmetic, basal area-weighted and quadratic mean diameter can be used. Equations for calculating them were presented in Table 1, but they are not yet implemented in the system.

6.3 H-D models from longitudinal data (III, IV, V)

Papers III and IV report models for the *H-D* relationship of the stand. The models were estimated by tree species and in applications the *H-D* curve of each tree species is predicted separately. The models apply the mixed model approach to longitudinal data, which is nowadays a standard approach in longitudinal studies (McCulloch and Searle 2001, p 187-219; Diggle et al. 2002, p 169-189).

In previous longitudinal studies of *H-D* curves, the development of the parameters has been linked to the stand age (e.g. Lappi 1997). However, Paper III showed that the development rate of an *H-D* curve of a shade-tolerant species stratum may differ substantially between stands of equal age. (See Figure 1 of Paper III). Two explanations for this were given: 1) shade tolerant trees may survive for a long time as an undergrowth in a closed forest and begin to grow rapidly as the amount of light increases and 2) the development rate depends on the site fertility, being slower on poor sites than on rich sites. Paper IV showed that the second explanation also holds with shade-intolerant tree species (Figure 1 of Paper IV). On the other hand, in stands with equal *DGM*, the development rate of *H-D* curves is fairly equal, with only a random stand-wise deviation that seems to be constant over time (Papers III and IV). Thus, linking the development of *H-D* curves with stand *DGM* led to rather simple linear mixed models, while linking the development with stand age would have required much more complicated models.

In order to provide a suitable model for different situations, five models with different sets of predictor variables were estimated for each tree species. The additional predictors used were stand location (x and y coordinates), altitude, cumulative temperature sum, site fertility, stand age and basal area. The mixed model approach makes it possible to localize the models for a new stand using any number of measured sample tree heights, as was demonstrated in Papers III and IV. Thus, if the height prediction obtained using only *DGM* of the stand is not accurate enough, the accuracy can be improved either by measuring more covariates and using another model or by measuring sample tree heights and localizing the model. However, as Fig. 4a of Paper IV demonstrates, despite the model used, the information of one sample tree often overrides the effect of additional predictors. More generally, the use of sample tree heights is recommended because it utilizes measured heights, instead of indeterminate correlations between stand variables and tree height, to improve height prediction.

The longitudinal approach of the model provides possibilities to carry *H-D* information from one point in time to another. For example, a localized future *H-D* curve can be predicted or old height sample trees can be used in localizing the *H-D* model for a given stand. Paper V showed that an old height measurement improves the accuracy of volume and saw timber volume predictions. Furthermore, Paper V showed that the number of height sample trees has a considerable effect on the accuracy of volume and saw timber volume predictions in a forest inventory by compartments and indicated that the number of height sample trees from one species stratum should be more than one.

6.4 A system for producing stand description in an inventory by compartments (V)

Paper V utilized the results of Papers I, II, III and IV to construct a system for the prediction of stand description in a forest inventory by compartments. An important feature of the system is that the stand description can be produced with a varying amount of input information. The minimum input requirement includes the estimates of basal area, *DGM*, stand age and site fertility class, all of which can be obtained from the measurement of one HPS-plot. Additional measurements that can be utilized are additional HPS-plots, any number of stem number measurements from fixed plots, any number of old or new height sample trees and any number of sample order statistics (i.e. quantile trees) from the HPS-plots. In addition, other stand variables that can be derived from the diameter distribution, e.g. arithmetic, basal-area weighted and quadratic mean diameters (See Table 1), can be used through the adjustment algorithm of Paper II. In addition to the measurements of the stand variables, the within-stand variance covariance matrix of basal area, *DGM* and the other stand variables used is needed. If only one sample plot is measured from a stand, a model-based estimate is required, but if more plots are measured, the estimate can be based on the observations of sample plots.

The possibility to use various input information makes the system useful in various kinds of inventories. Examples of these are the inventory of forest planning and the pre-harvest measurement of a stand marked for cutting. Furthermore, the measurement efforts can be directed according to the variable of interest. For example, when aiming at accurate estimation of the total amount of wood different measurements can be carried out than when aiming at accurate estimation of the saw log proportion.

6.5 Optimizing data collection in an inventory by compartments (V)

The system for producing a stand description was utilized to estimate models for the expected variance of total and saw timber volume given the values of the stand variables and numbers of stand measurements. The estimated models show that the most important factor affecting the accuracy of stand description is the number of HPS-plots, which provide information on the total basal area. The second most important factor is either the number of height sample trees or the number of quantile trees, depending on the stand properties and the aim of the inventory. If the aim is the accurate estimation of the total volume, the second most important factor (and the only one in addition to the number of HPS-plots) is the number of height sample trees. If the aim is the accurate estimation of the saw timber volume, the situation is different: in stands where the *DGM* is low, say less than 20 cm, the second and third most important factors are the numbers of quantile trees and height sample trees, respectively, but in stands with a larger *DGM*, their order is opposite (see Fig. 4 of Paper V). This is because in stands with a small *DGM*, the saw timber proportion depends strongly on the diameter distribution, whose prediction accuracy is improved considerably by quantile trees. On the other hand, in stands with a large *DGM*, almost all trees are saw timber trees and the proportion of saw timber depends on tree taper, the information on which is provided by the height sample trees. The measurement of stem number did not have a significant effect on the accuracy of total volume nor on the accuracy of saw timber volume.

Since the accuracy of stand description depends strongly on stand characteristics and the number of measurements, the measurement strategy of a stand should depend on stand characteristics and on the aim of the inventory. Paper V proposed an optimization approach to find the optimal strategy for each stand. Searching for the optimal strategy entailed solving a general linearly constrained optimization problem, where the numbers of different measurements were used as target variables. The optimization minimized the expected error of stand description subject to budget constraints. The expected error was a weighted average of the expected prediction errors of the total and saw timber volumes. For the definition of the constraints, time requirements of different measurements were needed; in Paper V ad hoc guesses were used. The solution of the optimization problem included concrete suggestions about the number of HPS-plot, height sample tree and quantile tree measurements, which makes the approach convenient for practical use.

7 Discussion

7.1 The system for producing a stand description

This study presented a new system for predicting stand description in an inventory by compartments. The system utilizes measurements of HPS-plots, stem number measurements, sample order statistics (i.e. quantile trees) and sample tree heights. In addition, old height measurements can be utilized, if available. The main differences between the proposed system and the systems that are currently used in Finland (e.g. Redsven et al. 2004) are the possibility to vary the set of measured variables and the utilization of measurement error variances of the stand variables. Thus, the accuracy of the produced stand description can be controlled through the number of different measurements. Furthermore, the measurement resources can be allocated to those measurements that provide most information about the variable of interest. The developed system was found promising in the prediction of stand description. However, it was not compared with the systems that are currently used in Finland and comparisons should be carried out in the future.

The system produces a static description of the stand. However, the *H-D* model of the system can also be used in predicting the *H-D* pattern of the stand in the future, and developing the system into a dynamic stand development model would require only additional longitudinal models of percentiles and stock density. Furthermore, all models of the system could be estimated simultaneously, which would provide the estimates of cross-model correlations. These would make it possible to cross-calibrate the models (Lappi 1991). For example, height measurement of spruce could be used in localizing the *H-D* model of pine in a spruce-pine mixture. This would make the use of measured forest data even more efficient.

The longitudinal *H-D* models (III and IV) provided a possibility to utilize old height measurements in the prediction of the current stand description and Paper V showed that old height measurements improve the prediction. As mentioned above, the principle used with height prediction could be generalized to other models of the system. For example, estimating longitudinal models for diameter percentiles would make it possible to utilize also old measured percentiles in predicting diameter distribution. If this approach proves to be useful, old measured information may become very valuable. Thus, in the near future, attention should be paid to saving all sample measurements in the database in such a form that they can be used later, to retaining old measurements in the database and to saving information about the origin of the data (measured or updated with models). For example, instead of saving only the mean height of the stand based on a height sample tree, the diameter and height of the height sample tree should be saved. Furthermore, all plot specific measurements of basal area and *DGM* should be saved in the database, since they are measured sample order statistics that may be very useful in the future.

In addition to the measurements of stand variables, the system developed in this study utilizes variances of their estimation errors. Paper V assumed that the forest data was collected with an objective sampling method that was based on measurements on randomly located sample plots. In the objective inventory, the measurement and prediction errors are quite controllable, because standard formulas of random sampling can be applied (e.g. Koivuniemi 2003, V). The current practice in Finland is, however, to use a subjective inventory method where plots are located subjectively and measurements are based on partly visual assessments. In the subjective inventory, the estimation of stand variables is regarded as more cost-efficient than in the objective inventory. However, processes generating errors in an inventory of this kind are very unclear and the errors depend strongly on the person carrying out the inventory (e.g. Haara 2003, Haara and Korhonen 2004, Kangas et al. 2004). If the measurement errors are utilized in calculations, as they are in the system of this study, the objective sampling method may be even more cost-efficient than the subjective sampling method. Thus, comparisons of subjective and objective sampling methods should be carried out in the future.

7.2 The use of sample information

The use of sample information is appealing because it provides possibilities to control the accuracy of predictions through the amount of sample measurements. However, using sample information alone would be a waste of information, because some stand variables (*DGM*, basal area, age, site fertility class) are measured anyway and they include information about the *H-D* curve and diameter distribution. This study used both information on the stand variables and the measured sample information in the prediction of the *H-D* curve and diameter distribution of a stand. The usefulness of the sample measurements seemed to be great when compared to other possibilities to improve the predictions. For example, Paper V showed that sample order statistics provide much more information on the diameter distribution than measurement of stem number. Furthermore, in the sample stand of Paper IV (Fig. 4a of Paper IV) the utility of additional covariates was marginal when compared to the utility of one height sample tree in the prediction of the *H-D* curve.

The measurement of a sample of diameters with the measurement equipment that is currently used is too laborious in an inventory for forest planning. However, quantile trees might be measured with little effort. This study found quantile trees to provide considerable information about the diameter distribution of the stand. This result is, in fact, a generalization of the result of Kangas and Maltamo (2002), where the most promising new measurements were minimum, maximum and arithmetic median diameter, i.e. the first, last and middle order statistics of a sample from the unweighted diameter distribution.

New measurement equipment (Laasasenaho et al. 2002) is being developed in order to make it possible for a single person to measure a large sample of diameters from a stand rapidly. The equipment is similar to an angle gauge, but the slot width can be adjusted and a laser telemeter is included. The diameter of a tree is measured by adjusting the slot width so that the tree exactly fills it and measuring the distance to the tree with a laser. However, the equipment has problems in measuring distances in branchy stands and in stands with dense undergrowth. The quantile tree approach could provide a solution to this problem: if the diameter cannot be measured, the rank of the tree on the plot could be assessed visually and the approach of Paper I used in the prediction of the diameter distribution of the stand.

The method for utilizing the measured sample order statistics is based on the observation that a measured sample order statistic is a measured percentile of the underlying diameter distribution. The algorithm requires the conditional expectations of the percentiles and the variance-covariance matrix of the prediction errors (i.e. stand effects). These are straightforwardly obtained when a percentile-based parameter prediction method is used and thus, this approach is recommended in the present study. However, it might be possible to localize also other distribution families with measured percentiles, as discussed in Paper I.

Measurements of quantile trees have not been carried out in practice, except for the measurements of the *DGM* tree from an HPS-plot, which is carried out by visually determining the median tree of the plot and measuring its diameter. The measurement of quantile trees in the field could be carried out as follows. Before counting any trees of the HPS-plot, one selects a tree that seems to belong to the plot and measures its diameter. When counting the trees with the angle gauge one determines visually for each tree on the plot if it is larger or smaller than the selected tree. However, in visual assessment measurement errors occur. Thus, the measurement accuracy, time consumption and usefulness of measured quantile trees with visually assessed ranks should be studied in the future.

A common practice in forestry is to predict various variables using regression models. This study utilized an approach where the models were localized using sample measurements of the response variable. The utilized approach was promising and other studies have also shown it to be useful (e.g. Lappi and Bailey 1988, Eerikäinen et al. 2002, Calama and Montero 2004). These observations raise the question of whether this principle should also be used with other models of the forest planning system. For example, should we begin to measure the growth of sample trees in order to predict the growth more accurately than it is currently predicted. This would require the use of such a model form and modeling approach that would make it possible to use sample measurements that are informative and can be carried out accurately and rapidly in the field.

7.3 Optimal allocation of stand measurements

As discussed previously, the studies of Eid (2000), Kangas and Maltamo (2002) and Holmström et al. (2003) indicated that money could be saved by varying the strategy of forest inventory from stand to stand. The savings consist of allocation of measurement time on the stand level and the area level. On the stand level, available measurement resources can be used for measurements that provide the largest amount of information about the specified target variable (e.g. total volume or saw timber volume). On the area level, more time can be spent on the measurement of those stands from which the most accurate information is needed, while in stands where the accuracy requirement is lower, the measurements can be limited to a minimum. For example, more time can be spent on stands that will be harvested in the near future than on sapling stands.

Paper V gave an operative tool to be used in finding the optimal measurement strategy for a single stand in practice. The optimization algorithm could be included in the field computer used in the inventory. The computer would carry out the optimization after the measurement of the first HPS-plot and suggest the combination of additional measurements. Furthermore, the suggestion would be updated after the measurement of each additional plot. The area level allocation of measurement resources to different stands was not optimized in Paper V in the sense that the resources were distributed between stands in a way that is most efficient with regard to some criterion. However, the maximum time requirement can vary between stands. For example, it can be defined to depend on stand characteristics.

The target variables in the optimization of data collection were the errors of total volume and saw timber volume. These variables were selected as examples in this study, because they depend on all components of the stand description, i.e. on the stand density, diameter distribution and *H-D* curve. If appropriate data are available, the approach of Paper V can also be used in the prediction of expected errors of some other target variables, e.g. pole volume or growth. In some cases, the interest may not lie in all three components. For example, the interest may lie in the accuracy of predicted stand structure, i.e. in the predictions of the diameter distribution and *H-D* curve, while the accuracy of stock density may be unimportant. In this case, the target variable could be the saw timber proportion instead of saw timber volume.

If the variable of interest is the diameter distribution alone, the target variable should be a variable that gives as much information as possible about the accuracy of diameter distribution. In different studies, different criterion variables have been used in measuring the accuracy of predicted diameter distribution. These are, for example, RMSE and bias of different variables derived from the diameter distribution (e.g. Maltamo et al. 1995, Temesgen et al. 2002, Robinson 2004). Furthermore, Reynolds' error index (Reynolds et al. 1988) and different goodness of fit test statistics and have been widely used (e.g. Liu et al. 2002, Zhang et al. 2003). All these could be used as target variables in the optimization. With basal area diameter distribution, the number of stems includes a considerable amount of information about the form of the distribution (e.g. Siipilehto 1999, Kangas and Maltamo 2000c). However, the measurement of stem number is very inaccurate (Kangas et al. 2004). Thus, Maltamo et al. (2003b) proposed that instead of trying to use the stem number in the prediction of diameter distribution, it could be used as a criterion variable measuring the accuracy of the predicted diameter distribution. Thus, if the aim of inventory is accurate basal area diameter distribution, a good target variable of the optimization could be the RMSE of stem number.

The aim of forest planning is to search for optimal management schedules for the stands of the forest area under consideration. Thus, in order to maximize the utility of a forest inventory with regard to forest planning, the effect of data collection on the management suggestions should be analyzed and the measurement strategy that minimizes the expected sum of costs and losses should be selected (Ståhl et al. 1994, Eid 2000 and Holmström et al. 2003). This cost-plus-loss approach suggests improving the accuracy of forest data whenever the costs of the improved accuracy are lower than the money saved in optimal harvest decisions. This study optimized the forest inventory of forest planning with respect to the accuracy of static stand description. This approach was selected because at this stage the study was limited to static prediction of stand description in the inventory by compartments. However, as discussed earlier, a natural extension of the system is a dynamic system with simultaneously estimated longitudinal models for the stock density, diameter distribution and *H-D* curve. With the extended system, management schedules could be produced and cost-plus-loss analyses of forest inventories would be possible.

8 REDUCING THE COSTS OF THE INVENTORY FOR FOREST PLANNING

This study aimed at responding to the need to decrease the costs of a forest inventory. Using the system proposed in this study, the costs might be decreased in the following ways.

1. By optimizing data collection on the stand level. A tool for optimization was proposed in Paper V. However, putting it into practice requires knowledge about the time requirements of different measurements. Furthermore, if quantile trees are used, their measurement accuracy and time requirement should be studied first.
2. By optimizing data collection on the area level. Because the proposed system makes it possible to vary the measurement combination from stand to stand, measurement resources could be allocated to stands where the accuracy requirement is highest.
3. By utilizing data of the previous inventory. This study utilized only height data, but estimation of longitudinal models for basal area and stock density could provide possibilities to utilize also old measurements of basal area and *DGM*. This requires that the measurements of sample plots should be carried out in such a way and saved in the database in such a form that they can be utilized later.
4. The proposed system might produce more accurate estimates than the currently used system even with the currently used measurement strategy, because the proposed height model is localized with a method that has a stronger theoretical basis than the one currently used. The accuracy of the current system and the proposed system should be compared in order to study which of them produces the most accurate estimates with the currently used measurement strategy.

If the aim is to improve the cost-effectiveness instead of reducing the costs of the forest inventory, one possibility, in addition to the four mentioned above, would be to improve the accuracy of forest data. If the expected savings in optimal harvest decisions are greater than the costs of the improved accuracy, the improvement is advantageous for the forest owner. Furthermore, an additional possibility to make the use of measured data more efficient would be to carry information between individual models of the model system. In particular, in a mixed stand, information could be carried from one tree species to another, as discussed previously.

References

Altman, N.S. 1992. Introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46(3): 175-164.

Anttila, P. 2002a. Nonparametric estimation of stand volume using spectral and spatial features of aerial photographs and old inventory data. Can. J. For. Res. 32(10): 1849-1857.

- 2002b. Updating stand level inventory data applying growth models and visual interpretation of aerial photographs. Silva Fenn. 36(2): 549-560.

- and Lehikoinen, M. 2002. Kuvioittaisten puustotunnusten estimointi ilmakuvilta puoliautomaattisella latvusten segmentoinnilla. Metsätieteen aikakauskirja 2002(3): 381-389.

Arabatzis, A. A. and Burkhart, H. E. 1992. An evaluation of sampling methods and model forms for estimating height-diameter relationships in loblolly pine plantations. For. Sci. 38(1): 192-198.

Bailey, R.L. and Dell, T.R. 1973. Quantifying diameter distributions with the Weibull function. For. Sci. 19(2): 97-104.

Bazaraa, M.S. and Shetty, C.M. 1979. Nonlinear programming, theory and algorithms. John Wiley & Sons, New York. 560 p.

Bliss, C. I. and Reinker, K. A. 1964. A lognormal approach to diameter distributions in even-aged stands. For. Sci. 10(3): 350-360.

Borders, B. E. and Patterson, W. D. 1990. Projecting stand tables: a comparison of the Weibull diameter distribution method, a percentile-based projection method, and a basal area growth projection method. For. Sci. 36(2): 413-424.

-, Souter, R.A., Bailey, R.L. and Ware, K.D. 1987. Percentile-based distributions characterize forest stand tables. For. Sci. 33(2): 570-576.

Burk, T.E. and Newberry, J.D. 1984. A simple algorithm for moment-based recovery of Weibull distribution parameters. For. Sci. 30(2): 329-332.

Cajander, A.K. 1926. The theory of forest types. Acta For. Fenn. 29(3): 1-108.

Cajanus, W. 1914. Über die Entwiclung gleichaltiger Waldbestände. Eine statsche Studie. Acta For. Fenn. 3. 142 p.

Calama, R. and Montero, G. 2004. Interregional nonlinear height-diameter model with random coefficients for stone pine in Spain. Can. J. For. Res. 34(1): 150-163.

Cancino, J. and von Gadow, K. 2002. Stem number guide curves for uneven-aged forests development and limitations. In von Gadow, K., Nagel, J. and Saborowski, J. (editors). Continuous cover forestry, assessment, analysis, scenarios. Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 163-174.

Cao, Q. V. and Baldwin, V. C. Jr. 1999. A new algorithm for stand table projection models. For. Sci. 45(4): 506-511.

- and Bukhart, H. E. 1984. A segmented distribution approach for modelling diameter frequency data. For. Sci. 30(1): 129-137.

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

Chaudhry, M.A. and Ahmad, M. 1993. On a probability function useful in size modeling. Can. J. For. Res. 23(8): 1679-1683.

Colbert, K.C., Larsen, D.R. and Lootens, J.R. 2002. Height-diameter equations for thirteen midwestern bottomland hardwood species. North. J. Appl. For. 19(4): 171-176.

Curtis, R. O. 1967. Height-diameter and height-diameter-age equations for second-growth douglas-fir. For. Sci. 13(4): 265-375.

Davidian, M. and Giltinan, D.M. 1995. Nonlinear models for repeated measurement data. Monographs on statistics and applied probability 62. Chapman & Hall/CRC, London. 359 p.

Diggle, P.J., Heagerty, P., Liang, K.-Y., Zeger, S.L. 2002. Analysis of longitudinal data. Oxford Statistical Science Series 25. Oxford University Press, Oxford. 379 p.

Droessler, T. D. and Burk, T. E. 1989. A test of nonparametric smoothing of diameter distributions. Scand. J. For. Res. 4(3): 407-415.

Eerikäinen, K. 2003. Predicting the height-diameter pattern of planted Pinus kesiya stands in Zambia and Zimbabwe. For. Ecol. and Manag. 175(1-3): 355-366.

- and Maltamo, M. 2003. A percentile based basal area diameter distribution model for predicting the stand development of Pinus kesiya plantations in Zambia and Zimbabwe. For. Ecol. and Manag. 172(1): 109-124.

-, Mabvurira, D., Nshubemuki, L. and Saramäki, J. 2002. A calibrateable site index model for Pinus kesiya plantations in southeastern Africa. Can. J. For. Res. 32(11): 1916-1928.

Eid, T. 2000. Use of uncertain inventory data in forestry scenario models and consequential incorrect harvest decisions. Silva Fenn. 34(2): 89-100.

Ek, A.R, Issos, J. N. and Bailey, R.L. 1975. Solving for Weibull diameter distribution parameters to obtain specified mean diameters. For. Sci. 21(3): 290-292.

-, Robinson, A.P., Radtke, P.J., Walters, D.K. 1997. Development and testing of regeneration imputation models for forests in Minnesota. For. Ecol. and Manage. 94(1-3): 129-140.

Fang, Z. and Bailey, R. L. 1998. Height-diameter models for tropical forests on Hainan Island in southern China. For. Ecol. and Manag. 110(1-3):315-327.

Farr, W.A., DeMars, D.J. and Dealy, J.E. 1989. Height and crown width related to diameter for open-grown western hemlock and Sitka spruce. Can. J. For. Res. 19(9): 1203-1207.

Finland's national forest programme 2010 1999. Ministry of Agriculture and Forestry Publications 2/1999. 40 p.

Flewelling, J.W. and de Jong, R. 1994. Considerations in simultaneous curve-fitting for repeated height-diameter measurements. Can. J. For. Res. 24(7): 1408-1414.

Fulton, M-R. 1999. Patterns in height-diameter relationships for selected tree species and sites in eastern Texas. Can. J. For. Res. 29(9): 1445-1448.

Gove, J. H. 2000. Some observations on fitting assumed diameter distributions to horizontal point sampling data. Can. J. For. Res. 30(4): 521-533

- 2003. A note on the relationship between the quadratic mean stand diameter and harmonic mean basal area under size-biased distribution theory. Can. J. For. Res. 33(8): 1587-1590.

- and Patil, G. P. 1998. Modelling basal area-size distribution of forest stands: a compatible approach. For. Sci. 44(2): 285-297.

Green, E.J. and Clutter, M. 2000. Using auxiliary information to estimate stand tables. Can. J. For. Res. 30(6): 865-872

Greenhill, G. 1881. Determination of the greatest height consistent with stability that a vertical pole or mast can be made, and of the greatest height to which a tree of given proportions can grow. Proceedings of Cambridge Philosophical Society. 4(2).

Gregoire, T.G., Schabenberger, O. and Barrett, J.P. 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. Can. J. For. Res. 25(1): 137-156.

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

Haara, A. 2003. Comparing simulation methods for modelling the errors of stand inventory data. Silva Fenn. 37(4): 477-491.

- and Korhonen, K.T. 2004. Kuvioittaisen arvioinnin luotettavuudesta. Manuscript.

-, Maltamo, M. and Tokola, T. 1997. The k-nearest-neighbour method for estimating basal-area diameter distribution. Scand. J. For. Res. 12: 200-208

Hafley, W.L. and Bufold, M.A. 1985. A bivariate model for growth and yield prediction. For. Sci. 31(1): 237-247.

- and Schreuder, H. T. 1977. Statistical distributions for fitting diameter and height data in even-aged stands. Can. J. For. Res. 7: 481-487.

Hanus, M.L., Marshall, D.D. and Hann, D.W. 1999. Height-diameter equations for six species in the coastal regions of the pacific northwest. Oregon State University, Forest Research laboratory. Research Contributions 25. 11 p.

Heikinheimo, M. (editor) 1999. Metsäsuunnittelun tietohuolto. Finnish Forest Research Institute, Res. pap. 741. 105 p.

Holmgren, J., Nilsson, M. and Olsson, H. 2003. Estimation of tree height and stem volume on plots using airborne laser scanning. For. Sci. 49(3): 419-428.

Holmström, H., Kallur, H. and Ståhl, G. 2003. Cost-Plus-loss analyses of forest inventory strategies based on *k*NN-assigned reference sample plot data. Silva Fenn. 37(3): 381-398.

Huang, S, Titus, S. J. and Wiens, D. P. 1992. Comparison of nonlinear height-diameter functions for major Alberta tree species. Can. J. For. Res. 22(9):1297-1304.

-, Price, D. and Titus, S.J. 2000. Development of ecoregion-based height-diameter models for white spruce in boreal forests. For. Ecol. and Manag. 129(1-3): 125-141.

Hyink, D. M. and Moser, J. W. Jr. 1983. A generalized framework for projecting forest yield and stand structure using diameter distributions. For. Sci. 29(1): 85-95.

Hynynen, J., Ojansuu, R., Hökkä, H., Siipilehto, J., Salminen, H., Haapala, P. 2002. Models for predicting stand development in MELA-system. Finnish Forest Research Institute, Res. pap. 835. 115 p.

Hyvönen, P. 2002. Kuvioittaisten puustotunnusten ja toimenpide-ehdotusten estimointi k-lähimmän naapurin menetelmällä Landsat TM -satelliittikuvan, vanhan inventointitiedon ja kuviotason tukiaineiston avulla. Metsätieteen aikakauskirja 2002(3): 363-379.

- and Korhonen, K.T. 2003. Metsävaratiedon jatkuva ajantasaistus yksityismetsissä. Metsätieteen aikakauskirja 2003(2): 83-96.

Hökkä, H. 1997. Height-diameter curves with random intercepts and slopes for trees growing on drained peatlands. For. Ecol. and Manag. 97(1): 63-72.

Hökkä, H. and Ojansuu, R. 2004. Height development of Scots pine on peatlands: describing change in site productivity with a site index model. Can. J. For. Res. 34(5): 1081-1092.

Ihalainen, M., Salo, K. and Pukkala, T. 2003. Empirical prediction models for Vaccinium myrtillus and V. vitis-idaea berry yields in North Karelia, Finland. Silva Fenn. 37(1): 95-108.

IMSL 1997. Fortran subroutines for mathematical applications. Math/Library. Vols 1 and 2. Visual Numerics. 1218 p. + Appendices.

Ishii, H., Reynolds, J.H., Ford, E.D. and Shaw, D.C. 2000. Height growth and vertical development of an old-growth Pseudotsuga-Tsuga forest in southwestern Washington State, U.S.A, Can.J.For.Res. 30(1): 17-24.

Jalkanen, A. and Mattila, U. 2000. Logistic regression models for wind and snow damage in northern Finland based on the National Forest Inventory data. For. Ecol. and Manag. 135(1-3): 315-330.

Jayaram, K. and Lappi, J. 2001. Estimation of height-diameter curves through multilevel models with special reference to even-aged teak stands. For. Ecol. and Manag.. 142(1-3): 155-162.

Kangas, A. S. 1997. On the prediction bias and variance in long-term growth projections. For. Ecol. and Manag. 96(3): 207-216.

- 1998a. Effect of errors-in-variables on coefficients of a growth model and on prediction of growth. For. Ecol. and Manag. 102(2-3): 203-212.

- 1998b. Uncertainty in growth and yield projections due to annual varition of diameter growth. For. Ecol. and Manag. 108(3): 223-230.

- 1999. Methods for Assessing uncertainty of growth and yield predictions. Can. J. For. Res. 29(9): 1357-1364.

- and Maltamo, M. 2000a. Calibrating predicted diameter distribution with additional information. For. Sci. 46(3): 390-396.

- and Maltamo, M. 2000b. Percentile-based basal area diameter distribution models for Scots pine, Norway spruce and birch species. Silva Fenn. 34(4): 371-380.

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

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

- and Maltamo, M. 2003. Calibrating predicted diameter distribution with additional information in growth and yield predictions. Can. J. For. Res. 33(3): 430-434.

-, Heikkinen, E. and Maltamo, M. 2002. Puustotunnusten maastoarvioinnin luotettavuus ja ajanmenekki. Metsätieteen aikakauskirja 2002(3): 425-440.

-, Heikkinen, E. and Maltamo, M. 2004. Accuracy of partially visually assessed stand characteristics – A case study of Finnish inventory by compartments. Can. J. For. Res. 34(4): 916-930.

Kilkki, P. and Päivinen, R. 1986. Weibull-function in the estimation of the basal area dbh-distribution. Silva Fenn. 20(2): 149-156.

-, Maltamo, M., Mykkänen, R. and Päivinen, R. 1989. Use of the Weibull function in estimating the basal area dbh-ditribution. Silva Fenn. 23(4): 311-318.

Knoebel, B.R. and Burkhart, H.E. 1991. A Bivariate distribution approach to modelling forest diameter distributions at two points in time. Biometrics 47(1): 241-253.

Knowe, S. A., Ahrens, G. R. and DeBell, D. S. 1998. Comparison of diameter-distribution-prediction, stand-table-projection and individual-tree-growth modelling approaches for young red alder plantations. For. Ecol. And Manag. 98(1): 49-60.

Koivuniemi, J. 2003. The accuracy of the compartmentwise forest inventory based on stands and located sample plots. University of Helsinki, Department of Forest Resource Management, Publications 36. 160 p.

Korpela, I. 2004. Individual tree measurements by means of digital aerial photogrammetry. Silva Fenn. Monographs 3. 93 p.

Kuru, G.A., Whyte, A.G.D. and Woollons, R.C. 1992. Utility of reverse Weibull and extreme value density functions to refine diameter distribution growth estimates. For. Ecol. and Manage. 48(1-2): 165-174.

Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Comm. Inst. For. Fenn. 108: 1-74.

- and Päivinen, R. 1986. On the checking of an inventory by compartments. Folia Forestalia 664. 19 p. (In Finnish with English summary)

-, Koivuniemi, J., Melkas, T. and Räty, M. 2002. Puuston mittaus etäisyyden- ja kulmanmittauslaitteella. Metsätieteteen aikakauskirja 2002(3): 493-497.

Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of Scots pine. Comm. Inst. For. Fenn. 134: 1-69.

- 1991. Calibration of height and volume equations with random parameters. For. Sci. 37(3): 781-801.

- 1993. Metsäbiometrian menetelmiä. Silva Carelica 24. University of Joensuu, 182 p.

- 1997. A longitudinal analysis of height/diameter curves. For. Sci. 43(4):555-570.

- and Bailey, R.L. 1988. A height prediction model with random stand and tree parameters: an alternative to traditional site index methods. For. Sci. 34(4): 907-927.

Lindsay, S. R., Wood, G. R. and Woollons, R. C.1996. Stand table modelling through the Weibull distribution and usage of skewness information. For. Ecol. And Manag. 81(1-3): 19-23.

Lin, C. 2003. Generating forest stands with spatio-temporal dependencies, Joensuun yliopiston yhteiskuntatieteellisiä julkaisuja 64. 123 p.

Little, S.N. 1982. Weibull diameter distributions for mixed stands of western conifers. Can. J. For. Res. 13: 85-88.

Liu, C, Zhang, L, Davis, C.J., Solomon, D.S. and Gove, J.H. 2002. A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. For. Sci. 48(4): 653-661.

Lynch, T. B. and Murphy, P. A. 1995. A compatible height prediction and projection system for individual trees in natural, even-aged shortleaf pine stands. For. Sci. 41(1):194-209.

Magnussen, S. 1986. Diameter distributions in Picea abies described by the Weibull model. Scand. J. For. Res. 1(4): 493-502.

Maltamo, M. 1997. Comparing basal area diameter distributions estimated by tree species and for the entire growing stock in a mixed stand. Silva Fenn. 31(1): 53-65.

- and Kangas, A. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28(8): 1107-1115.

-, Kangas, A., Uuttera, J., Torniainen, T. and Saramäki, J. 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogenous Scots pine stands. For. Ecol. And Manag. 133(3): 263-274.

-, Tokola, T. and Lehikoinen, M. 2003a. Estimating stand characteristics by combining single tree pattern recognition of digital video imagery and a theoretical diameter distribution model. For. Sci. 49(1): 98-109.

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

-, Eerikäinen, K., Pitkänen, J., Hyyppä, J. and Vehmas, M. 2004a. Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. Remote Sensing of Environment 90(3): 319-330.

-, Mustonen, K., Hyyppä, J., Pitkänen, J. and Yu, X. 2004b. The accuracy of estimating individual tree variables with airborne laser scanning in a boreal nature reserve. Can. J. For. Res. (In press)

McCulloch, C. E. and Searle, S. R. 2001. Generalized, linear, and mixed models. John Wiley & Sons. 325 p.

McTague, J. P. and Bailey, R. L. 1987. Compatible basal area and diameter distribution models for thinned loblolly pine plantations in Santa Catarina, Brazil. For. Sci. 33(1): 43-51.

Mehtätalo, L. 2002. Valtakunnalliset puukohtaiset tukkivähennysmallit männylle, kuuselle, koivuille ja haavalle. Metsätieteen aikakauskirja 2002(4): 575-592.

Meyer, H.A. 1940. A mathematical expression for height curves. Journal of forestry 38: 415-420.

Moeur, M., and Stage, A. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. 41(): 337-359.

Mykkänen, R. 1986. Weibull-funktion käyttö puuston läpimittajakauman estimoinnissa. of Joensuu, M.Sc. Thesis, 80 p.

Mähönen, M. 1984. Kuvioittaisen arvioinnin luotettavuus. University of Helsinki, M.Sc. thesis. 56 p.

Nanang, D.M. 1998. Suitability of the normal, log-normal and Weibull distributions for fitting diameter distributions of neem plantations in northern Ghana. For. Ecol. and Manag. 103(1): 1-7.

Nanos, N. and Montero, G. 2002. Spatial prediction of diameter distribution models. For. Ecol. and Manag.. 121(3): 147-158.

Nelder, J.A. and Mead, R. 1964. A simplex method for function minimization. Computer Journal 7: 308-313.

Nelson, T. C. 1964. Diameter distribution and growth of loblolly pine. For. Sci. 10: 105-115.

Nepal, S. K. and Somers, G. L. 1992. A generalized approach to stand table projection. For. Sci. 38(1): 120-133.

Ojansuu, R. 1993. Prediction of Scots pine increment using a multivariate variance component model. Männyn kasvun ennustaminen monimuuttuja- ja varianssikomponenttimallilla. Acta For. Fenn. 239. 72 p.

Omule, S.A.Y. and MacDonald R.N. 1991. Simultaneous curve-fitting for repeated height-diameter measurements. Can. J. For. Res. 21(9): 1418-1422.

Paananen, R. 2002. Uuden metsäsuunnittelujärjestelmän kehittämisen lähtökohtia ja tavoitteita. Metsätieteen aikakauskirja 2002(3): 532-536.

-, Valanne, K. and Ärölä, E. 2000. Solmu metsäsuunnittelun maastotyöopas. Metsätalouden kehittämiskeskus Tapio. Hakapaino Oy, Helsinki. 82 p.

Parresol, B. R. 1992. Baldcypress height-diameter equations and their prediction cionfidence intervals. Can, J. For. Res. 22(9):1429-1434.

Peng, C. 1999. Nonlinear height-diameter models for nine boreal forest tree species in ontario. Ontario Forest Research institute, Forest Research Report No. 155.

Pienaar, L.V. and Harrison, W. M. 1988. A stand table projection approach to yield prediction in unthinned even-aged stands. For. Sci. 34(3): 804-808.

Pigg, J. 1994. Keskiläpimitan ja puutavaralajijakauman sekä muiden puustotunnusten tarkkuus Metsähallituksen kuvioittaisessa arvioinnissa. University of Helsinki, M. Sc. thesis. 86 p.

Pinheiro, J.C. and Bates, D.M. 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York. 528 p.

Pitkänen, J. 2001. Individual tree detection in digital aerial images by combining locally adaptive binarization and local maxima methods. Can. J. For. Res. 31(5): 832-844.

Poso, S. 1983. Basic features of forest inventory by compartments. Silva Fenn. 17(4): 313-349. (In Finnish with English summary)

Pukkala, T. 1994. Metsäsuunnittelun perusteet. Gummerus, Jyväskylä. 242 p.

-, Miina, J., Kurttila, M. and Kolström, T. 1998. A spatial yield model for optimizing the thinning regime of mixed stands of Pinus sylvestris and Picea abies. Scand. J. For. Res. 13(1): 31-42.

Pussinen, A. 1992. Ilmakuvat ja Landsat TM – satelliittikuva välialueiden kuvioittaisessa arvioinnissa. University of Joensuu, M. Sc. thesis. 48 p.

R Development Core Team 2003. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Redsven, V., Anola-Pukkila, A., Haara, A,. Hirvelä, H., Härkönen, K., Kettunen, L., Kiiskinen, A., Kärkkäinen, L., Lempinen, R., Muinonen, E., Nuutinen, T., Salminen, O., Siitonen, M. 2004. MELA2002 reference manual (2nd edition). Finnish Forest Research Institute, Res. pap. 606 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.

Rennols, K., Geary, N. and Rollison, T.J.D. 1985. Characterizing diameter distribution by the use of Weibull distribution. Forestry 58(1): 57-66.

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

Robinson, A. 2004. Preserving correlation while modelling diameter distributions. Can. J. For. Res 34(1): 221-232.

Saksa, T., Uuttera, J., Kolström, T., Lehikoinen, M., Pekkarinen, A. and Sarvi, V. 2003. Clear-cut detection in boreal forest aided by remote sensing. Scand. J. For. Res. 18(6): 537-546.

Saramäki, J., Tikkanen, J. and Heino, E. (editors) 2003. Yksityismetsien suunnittelun uudet tuulet. Seminaari Ylivieskassa 26.11.2003. Finnish Forest Research Institute, Res. pap. 900. 64 p.

Schnute, J. 1981. A versatile growth model with statistically stable parameters. Can. J. Fish. and Aq. Sci. 38(9): 1128-1140.

Schreuder, H. T. and Hafley, W. L. 1977. A useful distribution for describing stand structure of tree heights and diameters. Biometrics 33: 471-478.

-, Hafley, W.L. and Bennett, F.A. 1979. Yield prediction for unthinned natural slash pine stands. For. Sci. 25(1) 25-30.

-, Bhattacharya, H.T. and McClure, J.P. 1982. Towards a unified distribution theory for stand variables using S_{BBB} distribution. Biometrics 38: 137-142.

Scolforo, J. R. S., Tabai, F.C.V., de Macedo, R.L.G., Acerbi. F.W.Jr. and de Assis, A.L. 2003. SB distribution's accuracy to represent the diameter distribution of Pinus taeda, through five fitting methods.For. Ecol. and Manag. 175(1-3): 489-496.

Shiver, B. D. 1988. Sample sizes and estimation methods for the Weibull distribution for unthinned slash pine plantation diameter distributions. For. Sci. 34(3): 809-814.

Siipilehto, J. 1996. Metsikön läpimitta- ja pituusjakauman kuvaaminen kaksiulotteisen todennäköisyysfunktion avulla. University of Helsinki, Department of forest resource management. 71 p.

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

- 2000. A comparison of two parameter prediction methods for stand structure in Finland. Silva Fenn. 34(4): 331–349.

Stout, B.B. and Shumway, D.L. 1982. Site quality estimation using height and diameter. For. Sci. 28(3): 639-645.

Ståhl, G, Carlsson, D. and Bondesson, L. 1994. A method to determine optimal stand data acquistion policies. For. Sci. 40(): 630-649.

Tang, S., Wang, Y., Zhang, L. and Meng, C-H. 1997. A distribution-independent approach to predicting stand diameter distribution. For. Sci. 43(4): 491-500.

Temesgen, H. 2003. Estimating stand tables from aerial attributes: a comparison of parametric prediction and most similar neighbour methods. Scand. J. For. Res. 18(3): 279-288.

-, LeMay, V.M., Froese, K.L. and Marshall, P.L. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. and Manag. 177(1-3): 277-285.

Tewari, V.P. and Gadov, K.v. 1999. Modelling the relationship between tree diameters and heights using S_{BB} distribution. For. Ecol and Manag. 119(1-3): 171-176.

Uuttera, J. and Maltamo, M. 1995. Impact of regeneration method on stand structure prior to first thinning. comparative study North Karelia, Finland vs. Republic of Karelia, Russian Federation. Silva Fenn. 29(4): 267-285.

-, Haara, A., Tokola, T. and Maltamo, M. 1998. Determination of the spatial distribution of trees from digital aerial photographs. For. Ecol. and Manag. 110(1-3): 275-282.

-, Hiltunen, J., Rissanen, P., Anttila, P. and Hyvönen, P. 2002. Uudet kuvioittaisen arvioinnin menetelmät - arvio soveltuvuudesta yksityismaiden metsäsuunnitteluun. Metsätieteen aikakauskirja 2002(3): 523-531.

Van Deusen, P. C. 1986. Fitting assumed distributions to horizontal point sample diameters. For. Sci 32(1): 146-148.

Veltheim, T. 1987. Pituusmallit männylle, kuuselle ja koivulle. M. Sc. thesis, University of Helsinki. 59 p. + appendices 29 p.

Wang, C-H and Hann, D.W. 1998. Height-diameter equations for sixteen tree species in the central western Willamette valley of Oregon. Oregon State University, College of Forestry, Research Paper 51.

Woollons R.C., Snowdon P., Mitchell N.D. 1997. Augmenting empirical stand projection equations with edaphic and climatic variables. For. Ecol. and Manag. 98(3): 267-275.

Yuancai, L. and Parresol, B.R. 2001. Remarks on height-diameter modeling. Research note. USD Forest service, southern research station. 5 pp.

Zakrzewski, W. T. and Bella, I. E. 1988. Two new height models for volume estimation of lodgepole pines. Can. J. For. Res. 18(2): 195-201.

Zarnoch, S.J. and Dell, T.R. 1985. An evaluation of percentile and maximum likelihood estimators of Weibull parameters. For. Sci. 31(1): 260-268.

Zeide, B. and Vanderschaaf, C. 2002. The effect of density on the height-diameter relationship. In: Outcalt, K.W. ed. 2002. Proceedings of the eleventh biennial southern silvicultural research conference. Gen. Tech. Rep. SRS-48. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 622 p.

Zhang, L. 1997. Cross-validation of non-linear growth functions for modelling tree height-diameter relationships. Ann. Bot. 79(3): 251-257.

-, Gove, J. H., Liu, C. and Leak, W.B. 2001. A finite mixture of two Weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. For.Sci. 48(4): 653-661.

-, Peng, C., Huang, S. and Zhou, X. 2002. Development and evaluation of ecoregion-based jack pine height-diameter models for Ontario. For. Chron. 78(4): 530-538.

-, Packard, K. C. and Liu, C. 2003.A comparison of estimation methods for fitting Weibull and Johnson's SB distributions to mixed spruce-fir stands in northeastern North America. Can. J. For. Res. 33(7): 1340-1347.

-, Bi, H., Cheng, P. and Davis, C.J. 2004. Modeling spatial variation in tree diameter-height relationships. For. Ecol and Manag. 189(1-3): 317-329.

Zhou, B. and McTague, J. P. 1996. Comparison and evaluation of five methods of estimation of the Johnson system parameters. Can. J. For. Res. 26(6): 928-935.

Zucchini, W., Schmidt, M. and von Gadow, K. 2001. A model for the diameter-height distribution in an uneven-aged beech forest and a method to assess the fit of such models. Silva Fenn. 35(2): 169-183.

Total of 170 references

I

LOCALIZING PREDICTED DIAMETER DISTRIBUTION WITH SAMPLE INFORMATION

Lauri Mehtätalo

Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, Fin-80101 Joensuu, Finland.

Abstract. This study presents a new method for predicting the diameter distribution of a stand. The method utilizes the percentile-based diameter distribution. The expected diameter percentiles are first predicted using stand measurements. Subsequently, the distribution is calibrated (localized) for the stand using sample order statistics, which consist of one or more diameters of sample trees and their ranks on the sample plot(s). These measurements can be carried out rather rapidly in the field, because the rank can be assessed visually. The sample order statistics can be interpreted as measured sample percentiles. The expectations, variances and covariances of the measured sample order statistics are derived using the theory of order statistics. Regression models are utilized to predict the conditional expectations of predefined percentiles, which are then combined with the measured percentiles using the best linear unbiased predictor (BLUP). The method was tested in a real dataset using simulated sample plots. The test showed that even with a small number of sample measurements, the Reynolds' error index and RMSE and bias of volume could be decreased remarkably. Furthermore, increasing the number of measurements improved the prediction steadily. The proposed method seems to be a promising tool in the prediction of diameter distributions of various forms and it seems to work also in complex stands.
Key words: percentile, order statistic, rank, BLUP.

Introduction

The most natural way of estimating the diameter distribution of a stand is to utilize a measured sample of diameters. However, the measurement of a diameter sample is time consuming and too cumbersome to be carried out by a single person. Thus, in many inventories, e.g. in an inventory by compartments for forest planning, there are no resources for measuring a diameter sample. This study proposes the measurement of sample order statistics as an alternative to the measurement of the complete sample, i.e. the measurement of one or more tree diameters and determination of the ranks of these diameters in the sample. The main point in measuring an order statistic in the field is that one needs to measure exactly the diameter of only one tree and for the other trees one only needs to know whether they are larger or smaller than the measured tree. In many cases, this can be assessed visually without walking to the base of the tree. Thus, order statistics of a horizontal point sample plot can be assessed rather easily by a single person.

There are three approaches used in predicting the diameter distribution of a stand. The first approach is based on a sample of diameters, which may either be smoothed somehow or used as such to represent the diameter distribution of the stand (e.g. Van Deusen 1986, Pienaar and Harrison 1988, Droessler and Burk 1989, Nepal and Somers 1992, Lindsay et al. 1996, Tang et al. 1997). In the second approach, parameters of some presupposed distribution family are predicted with some easily measurable stand characteristics (e.g. Hyink and Moser 1983, Rennolls et al. 1985, Kilkki and Päivinen 1986, Borders et al. 1987, Borders and Patterson 1990, Maltamo et al. 2000). In the first approach, the accuracy of prediction is good and can be improved by enlarging sample size, but measuring a sample of diameters is rather time consuming. In the second approach, on the other hand, the measurements can be obtained fairly rapidly but the accuracy is not good enough for many inventories, e.g. for an inventory for forest planning.

The third approach uses known relations between stand variables and diameter distribution to recover the distribution parameters (e.g. Burk and Newberry 1984). The recovery gives the only existing compatible solution of the assumed distribution family. However, the correct family is not known and thus, the obtained distribution is not the actual one even if the stand values are correct. The recovery is possible only for as many parameters as there are known stand variables related to the diameter distribution. Thus, with distribution families that have more free parameters than the number of measured stand variables, only partial recovery is possible. In Finland, partial recovery is commonly used in parameter prediction by setting the measured median to equal the median of the assumed distribution (e.g. Kilkki and Päivinen 1986, Maltamo et al. 2000); it is used also in this study.

In the first approach, the sample diameter distribution converges to the diameter distribution of the stand as sample size increases. Hence, if no information is lost when smoothing the sample diameter distribution, the predicted distribution approaches the actual distribution of the stand as sample size increases. With the second approach this does not hold: even if the values of the predictors were known exactly, the predicted parameters of the distribution would not be the actual parameters of the stand. Instead, assuming that the model is right, they are the conditional expectations of the parameters given the values of the predictors. In other words, they are the expected values of the parameters in a stand belonging to the population of stands with the given values of the predictors. In addition, the residual variances of the models of the parameters include information about how much the parameters of a single stand vary from their expected values.

The aim of this study was to develop a method that is able to predict the diameter distribution of a stand using measurements of stand variables and order statistics of diameter sample(s). The method combines the first and second approach presented above. It is shown that a measured order statistic is an unbiased estimator of some percentile of the underlying diameter distribution. Thus, the percentile-based diameter distribution (Borders et al. 1987) is a natural selection as the distribution family in this study. The expectations, variances and covariances of the sample order statistics are derived using the theory of order statistics. The conditional expectations of the percentiles are first obtained using the second approach. Subsequently, localized stand-level percentiles are predicted using the expected values of the percentiles, sample order statistics and information about their accuracy. The predicted expectations of the percentiles are combined with the sample information using the best linear unbiased predictor (BLUP). The usefulness of the method is demonstrated with a case study that is based on sample plots simulated using real data.

Method

Percentile-based diameter distribution

Let the diameter distribution of a stand be described with a vector of diameters $\mathbf{d}=(d_1,d_2,...,d_k)'$ corresponding to certain predefined values of the diameter distribution function F_Y, say vector $\mathbf{p}=(p_1,p_2,...,p_k)'$. For vector $\mathbf{p}$, $p_1=0$, $p_k=1$ holds and the elements $d_i,..., d_k$ and $p_i,..., p_k$ of vectors $\mathbf{d}$ and $\mathbf{p}$ are in increasing order. Thus, the diameter d_i is the $100^* p_i^{th}$ percentile of the diameter distribution F_Y. Assuming that the cumulative distribution function (c.d.f.) between consecutive percentiles is linear we get the c.d.f. of tree diameter Y:

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

where

$$b_i=\frac{p_{i+1}-p_i}{d_{i+1}-d_i} \text{ and } a_i=p_i-b_i d_i .$$

This formulation is the percentile-based diameter distribution utilized, for example, by Borders et al. (1987), Borders and Patterson (1990) and Mehtätalo (2004). Maltamo et al. (2000) and Kangas and Maltamo (2000a) used the same distribution except that the interpolation between consecutive percentiles was carried out with a spline function. In this study, however, linear interpolation is preferred in order to keep the computations simple. Thus, the resulting distribution can be interpreted as a piecewise defined uniform distribution (cf. Cao and Burkhart 1984) or as a finite mixture of k-1 uniform distributions (c.f. Liu et al. 2002). Because of the linear interpolation, many characteristics of the distribution, e.g. expectation, variance, median, and quadratic mean diameter, can be easily derived analytically (see Mehtätalo 2004). By derivation of (1) with respect to y one can see that the density of the diameter is a constant b_i at each interval $[d_i,d_{i+1})$. A distribution with a c.d.f. of the form (1) is later denoted with $Perc_{\mathbf{p}}(\mathbf{d})$.

Diameter distribution is commonly used to calculate number of stems, volume or some other stand characteristics between certain diameter limits. However, it can also be interpreted as a probability distribution giving the probability with which the diameter of a randomly selected tree is between two given diameters. In this case, a random sample of trees in a stand is regarded as a random sample of diameters from the diameter distribution of the stand.

Expectation, variance and covariance of order statistics

If one has measured the diameter of a tree and, in addition, knows its rank r in the sample of size n, the measured diameter is the r^{th} order statistic of the sample. This section shows that the measured sample order statistic is an estimate of certain percentile of the underlying population, i.e. it is a measured percentile of the stand. The next section demonstrates how this measured percentile can be used as additional information in predicting the diameter distribution of the stand. The idea is to calibrate (localize) the expected percentiles for a certain stand using the measured sample percentile(s). Because the measurements come from a sample or samples, they include some amount of sampling error. Hence, to be able to use the measured sample order statistics in calibration, we need to know their expectations and the variance-covariance matrix of their sampling error. Here they are derived using the exact results for distributions of order statistics.

Assume that $Y_1, ..., Y_n$ is an independent sample with a common distribution function F_Y. Denote the r^{th} order statistic of the sample by $Y_{r:n}$. The density of $Y_{r:n}$ is (Reiss 1989, p. 21, Casella and Berger 2002, p. 229)

$$f_{r:n}(y)=\beta_0(r,n)f_Y(y)\left[F_Y(y)\right]^{r-1}\left[1-F_Y(y)\right]^{n-r}, \quad (2)$$

where $\beta_0(r,n)=\frac{n!}{(r-1)!(n-r)!}$ and the functions $f_Y(y)$ and $F_Y(y)$ are the density and c.d.f of the random variable Y, respectively. It is easy to see that if Y is uniformly distributed between [0,1], $Y_{r:n}$ follows a beta distribution with the parameters r and n-r+1.

Next, the diameter distribution of a stand, $F_Y(y)$, is assumed to be known and of the form (1). Then a sample of diameters, $Y_1, ..., Y_n$, is a random sample from the distribution $Perc_{\mathbf{p}}(\mathbf{d})$. Writing (1) into (2) gives the density of the r^{th} order statistic at the i^{th} interval $[d_i, d_{i+1})$:

$$f_{r:n}{}^{i}(y)=\beta_0(r,n)b_i\left[a_i+b_iy\right]^{r-1}\left[1-a_i-b_iy\right]^{n-r}, \quad (3)$$

which is a piece of the beta distribution. The expectation of $Y_{r:n}$ is calculated as

$$E(Y_{r:n})=\beta_0(r,n)\int_{d_1}^{d_k} yf_{r:n}(y)dy\,.$$

Dividing the integral into subintervals according to the percentile intervals gives

$$E(Y_{r:n})=\beta_0(r,n)\sum_{i=1}^{k-1}\int_{d_i}^{d_{i+1}} yf_{r:n}{}^{i}(y)dy\,. \quad (4)$$

Utilizing the assumption that the distribution $F_Y(y)$ is the diameter distribution of the stand, its inverse (quantile function) $F_Y^{-1}(p)$, gives the $100*p^{th}$ diameter percentile of the stand (Reiss 1989, p. 14-15). Thus, the value of the c.d.f. corresponding to the expectation of the sample percentile is the value of p satisfying $E(Y_{r:n})=F_Y^{-1}(p)$, which gives the solution

$$p=F_Y\left[E(Y_{r:n})\right]. \quad (5)$$

Hence, the diameter of the r^{th} largest tree in a sample of size n, $Y_{r:n}$, is an unbiased estimator of the $100*p^{th}$ percentile of the stand, assuming that the c.d.f. of the stand is F_Y.

The calculation of the second moment of $Y_{r:n}$ corresponds to formula 4:

$$E(Y_{r:n}{}^2)=\beta_0(r,n)\sum_{i=1}^{k-1}\int_{d_i}^{d_{i+1}} y^2f_{r:n}{}^{i}(y)dy$$

giving the variance

$$\mathrm{var}(Y_{r:n})=E(Y_{r:n}{}^2)-\left[E(Y_{r:n})\right]^2. \quad (6)$$

The joint density of two order statistics $Y_{r1:n}$ and $Y_{r2:n}$ from a sample of size n from a population with the c.d.f. $F_Y(y)$ is (Reiss 1989, p. 30-31, Casella and Berger 2002, p. 230)

$$\begin{aligned}f_{r_1,r_2:n}(y_1,y_2)&=\beta_1(r_1,r_2,n)f_Y(y_1)f_Y(y_2)\times\\ &\left[F_Y(y_1)\right]^{r_1-1}\left[F_Y(y_2)-F_Y(y_1)\right]^{r_2-r_1-1}\left[1-F_Y(y_2)\right]^{n-r_2}\end{aligned} \quad (7)$$

if $y_1<y_2$ and 0 otherwise, where $r_1<r_2$ and

$$\beta_1(r_1,r_2,n)=\frac{n!}{(n-r_2)!(r_2-r_1-1)!(r_1-1)!}\,.$$

Writing (1) into (7) gives the joint density in the case of the percentile-based diameter distribution within each quadrangle $[d_i, d_{i+1})*[d_j, d_{j+1})$:

$$\begin{aligned}f_{r_1,r_2:n}{}^{i,j}(y_1,y_2)&=\beta_1(r_1,r_2,n)b_ib_j\times\\ &(a_i+b_iy_1)^{r_1-1}(a_j+b_jy_2-a_i-b_iy_1)^{r_2-r_1-1}(1-a_j-b_jy_2)^{n-r_2}\end{aligned} \quad (8)$$

for $y_1<y_2$ and 0 otherwise. To calculate the covariance between these order statistics we need to calculate the expectation of their product first. Integrating each quadrangle

$[d_i, d_{i+1})*[d_j,d_{j+1})$ separately for $i,j=1,\ldots,k-1$ gives

$$E\left(Y_{r_1:n}Y_{r_2:n}\right)=$$
$$\beta_1(r_1,r_2,n)\sum_{j=1}^{k-1}\sum_{i=1}^{k-1}\int_{d_i}^{d_{i+1}}\int_{d_j}^{d_{j+1}} y_1 y_2 f_{r_1:n,r_2:n}^{\;i,j}(y_1,y_2)dy_2dy_1 .$$

Using this result and the expectation of single order statistic, (4), the covariance of two order statistics is calculated as

$$\text{cov}\left(Y_{r_1:n},Y_{r_2:n}\right)=E\left(Y_{r_1:n}Y_{r_2:n}\right)-E\left(Y_{r_1:n}\right)E\left(Y_{r_2:n}\right). \qquad (9)$$

Since the densities (3) and (8) are n-1th and n-2th order polynomials, respectively, exact calculation of the integrals may lead to degeneration of computing accuracy with large values of n and need to be calculated numerically in applications.

Predicting localized percentiles

Assume that models predicting k diameter percentiles are available and the $100*p_i^{th}$ percentile in stand m follows the model $d_{im} = \mu_{im} + e_{im}$ for i=1,2,…,k. Writing all percentiles as a vector, the model for the whole set of percentiles in stand m is

$$\mathbf{d}_m = \boldsymbol{\mu}_m + \mathbf{e}_m , \qquad (10)$$

where $\mathbf{d}_m=(d_{1m}, d_{2m}, \ldots, d_{km})'$, $\boldsymbol{\mu}_m=(\mu_{1m},\mu_{2m}, \ldots, \mu_{km})'$ and $\mathbf{e}_m=(e_{1m}, e_{2m}, \ldots, e_{km})'$ with $E(\mathbf{e}_m)=0$. Assuming that the model is correct and the parameters are known, vector $\boldsymbol{\mu}_m$ includes the conditional expectations of the percentiles given the values of the stand variables, $E(\mathbf{d}|\mathbf{x}_m)$. These are later referred to as expected percentiles of the stand. Vector $\mathbf{e}_m$ includes the stand effects of stand m, i.e. the deviations of the stand-level percentiles from their conditional expectations. Estimation of the model (Equation 10) using a simultaneous regression technique (SUR) produces an estimate of the variance-covariance matrix of stand-level deviations, var($\mathbf{e}_m$), denoted by $\mathbf{D}_{k\times k}$. Writing the predefined p-values into vector $\mathbf{p}=(p_1,p_2, \ldots,p_k)'$, the diameter distribution based on expected percentiles of stand m can be written as $Perc_{\mathbf{p}}(\boldsymbol{\mu}_m)$ and the diameter distribution of stand m, correspondingly, as $Perc_{\mathbf{p}}(\boldsymbol{\mu}_m + \mathbf{e}_m)$. Since the following calculations apply to one stand, the stand index m is dropped hereafter.

The measured order statistics are used to predict the vector of stand effects, $\mathbf{e}_{k\times 1}$., using the standard linear prediction theory (see e.g. Lappi 1986, 1991, 1997). Assume that we have q measured order statistics from a stand in vector $\mathbf{d}^*=(Y_{r1:n1}, Y_{r2:n2}, \ldots, Y_{rq:nq})'$. The measurements follow the model

$$\mathbf{d}^* = \boldsymbol{\mu}^* + \mathbf{e}^* + \boldsymbol{\varepsilon} , \qquad (11)$$

where the measured diameters are in vector $\mathbf{d}^*_{q\times 1}$, their conditional expectations in vector $\boldsymbol{\mu}^*_{q\times 1}$, the stand effects in vector $\mathbf{e}^*_{q\times 1}$ and the sampling error in vector $\boldsymbol{\varepsilon}_{q\times 1}$. For the random part of the model E($\mathbf{e}^*$)=E($\boldsymbol{\varepsilon}$)=$\mathbf{0}$ and cov($\mathbf{e}^*$, $\boldsymbol{\varepsilon}'$)=$\mathbf{0}$ holds. At this stage we assume that the actual distribution of the stand, $Perc_{\mathbf{p}}(\boldsymbol{\mu} + \mathbf{e})$, is known (I return to this later). The p-values corresponding to the measurements are obtained using formulas (4) and (5) and written into vector $\mathbf{p}^*_{q\times 1}$. The conditional expectations of measurements are then calculated as $\boldsymbol{\mu}^*=F^{-1}(\mathbf{p}^*)$, where F^{-1} is the inverse of the distribution based on expected percentiles, $Perc_{\mathbf{p}}(\boldsymbol{\mu})$.

For predicting the realized stand effects of model 10, the sampling variance-covariance matrix of order statistics, var($\boldsymbol{\varepsilon}$), denoted by $\mathbf{R}_{q\times q}$, the variance-covariance matrix of the stand effects of model 11, var($\mathbf{e}^*$), denoted by $\mathbf{D}^*_{q\times q}$ and covariance matrix of stand effects of models 10 and 11, cov($\mathbf{e}$,$\mathbf{e}^{*\prime}$), denoted by $\mathbf{C}_{k\times q}$ are needed. The variances of order statistics are calculated using formula (6) and written on the diagonal of $\mathbf{R}$. If one has measured several order statistics from the same sample plot, their covariances are calculated using formula (9) and written in the corresponding cells of matrix $\mathbf{R}$. The covariances of order statistics from different plots are naturally zero. Matrices $\mathbf{D}^*$ and $\mathbf{C}$ are obtained from matrix $\mathbf{D}$ by interpolating it for the values of $\mathbf{p}^*$.

In localization, the unobserved random vector $\mathbf{e}$ is predicted using the observed random vector $\mathbf{e}^*+\boldsymbol{\varepsilon}=\mathbf{d}^*-\boldsymbol{\mu}^*$. For the prediction, we

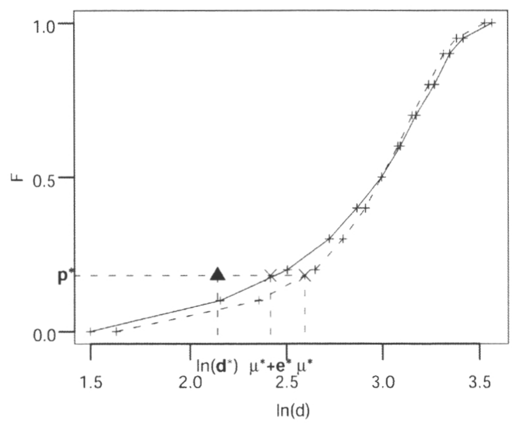


Figure 1. The distributions based on the expected percentiles, **μ** (dashed line) and on the localized percentiles, **μ+e** (solid line) obtained using the observation $Y_{3:13}$=8.5 cm (▲).

need to derive the variances and covariances of these random vectors using the known matrices **D**, **D***, **C** and **R**. The variance of stand effects of model 10, var(**e**), is straightforwardly **D**. Since cov(**e***, **ε**)=**0** and cov(**e**, **ε**)=**0**, var(**e***+**ε**)=**D***+**R** and cov[**e**,(**e***+**ε**)']=cov(**e**,**e***')=**C**. Using the notation of McCulloch and Searle (2001, p. 247), these results can be expressed as

$$\begin{bmatrix} \mathbf{e} \\ \mathbf{e}^*+\boldsymbol{\varepsilon} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D}^*+\mathbf{R} \end{bmatrix} \right).$$

The Best Linear Unbiased Predictor (BLUP) of **e** is calculated as

$$\hat{\mathbf{e}} = \mathbf{C}(\mathbf{D}^*+\mathbf{R})^{-1}(\mathbf{d}^*-\boldsymbol{\mu}^*) \qquad (12)$$

with the prediction variance of

$$\mathrm{var}(\hat{\mathbf{e}}-\mathbf{e}) = \mathbf{D}-\mathbf{C}(\mathbf{D}^*+\mathbf{R})^{-1}\mathbf{C}' \qquad (13)$$

(McCulloch and Searle 2001, p. 250). The stand level diameter percentiles are obtained by adding the predicted stand effects to the expected percentiles (see model 10).

The stand-level diameter distribution would already be needed in the calculation of **p*** and **R**. Since it is the result of the prediction and is not known when **p*** and **R** are calculated, the solution is searched iteratively. At the first iteration step, the expected diameter distribution of the stand, $Perc_p(\boldsymbol{\mu})$, is used as the stand-level distribution to approximate **p*** and **R**. Subsequently, these approximations are used to predict the stand-level diameter distribution, $Perc_p(\boldsymbol{\mu}+\hat{\mathbf{e}})$, with BLUP. Furthermore, this prediction is used to calculate new approximations of **p*** and **R** and the prediction of stand effects is carried out again. Repeating this until the predicted stand-level percentiles converge gives the final predicted stand-level diameter distribution.

In some cases, we may obtain two observations of the same percentile. This happens when there are two measurements of the same percentile from different plots, i.e. the number of tally trees is the same on two sample plots from the same stand and diameters with the same rank are measured from both plots. This means that the same row (and column) is included in matrix **D*** twice. It is therefore singular and the calibration cannot be carried out. The non-singularity of matrix **D*** can be guaranteed by treating the several measurements of the same order statistic as one measurement in the calculations. In this case, the mean of measured diameters is used as the observed percentile and the elements of matrix **R** are calculated using general rules for the variances and covariances of sums.

Numerical example

This section presents a numerical example of the calibration algorithm. The model of Kangas and Maltamo (2000a) was used to predict the conditional expectations of k=11 percentiles of basal-area diameter distribution. The predictors of the model are basal area median diameter (*DGM*), age, basal area and soil type. The models predict the 0^{th} 10^{th}, 20^{th}, 30^{th}, 40^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th}, 95^{th} and 100^{th} percentiles, i.e.

$$\mathbf{p}=(0\ 0.1\ 0.2\ 0.3\ 0.4\ 0.6\ 0.7\ 0.8\ 0.9\ 0.95\ 1.0)'.$$

In addition, the known *DGM* is used as the 50^{th} percentile. The models have been estimated using seemingly unrelated regression and the estimated variance-covariance matrix **D** was obtained from the original SUR-fit (Table 1). The model of percentiles is estimated on a logarithmic scale, i.e. the model is of the form (cf. Equation 10)

$$\ln(\mathbf{d}) = \boldsymbol{\mu} + \mathbf{e}\,. \tag{14}$$

In order to give a numerical example of the proposed algorithm, the diameter distribution of a Norway spruce stand was predicted, where *DGM* is 20 cm, basal area is 22 m^2/ha , stand age is 64 years and site fertility class is mesic. The predicted logarithmic percentiles using these values (i.e. conditional expectations given the values of stand variables) were

$$\boldsymbol{\mu} = \begin{pmatrix} 1.63 \\ 2.35 \\ 2.65 \\ 2.79 \\ 2.91 \\ 3.08 \\ 3.15 \\ 3.24 \\ 3.31 \\ 3.38 \\ 3.53 \end{pmatrix}.$$

In addition to the known stand variables, the third smallest tree of a horizontal point sample of 13 trees has been observed to be 8.5 cm in diameter, i.e. $Y_{3:13}$ has been observed to be ln(8.5)=2.140. The expectation of the sample order statistic $Y_{3:13}$ of the distribution $Perc_p(\boldsymbol{\mu})$ is (Equation 4) $E(Y_{3:13})$= 2.595. Equation 5 gives the value p^*= 0.182, which means that the measured order statistic is a measurement of the 18.2th percentile of the diameter distribution. The observed percentile has been plotted onto Figure 1 at the location (2.140,0.182). Note that we have k=11 percentiles of predefined percentage values and q=1 measured percentiles. Thus, $\mathbf{p}^*$, $\mathbf{e}^*$, ε, $\mathbf{R}$ and $\mathbf{D}^*$ are scalars and $\mathbf{C}$ is a vector with a length of 11.

The next step is to predict the stand effects of the percentiles, **e** (Equation 14). In order to be able to do it, we need the variances of sampling errors (**R**), the variances of the stand effects (**D***) and the covariances between the stand effects of model 14 and the stand effect of the 18.2th percentile (**C**). The variance of sampling error is (Equation 6) $\mathbf{R}$=var(ε)= 0.0577 and the variance of stand effect is obtained by interpolation of matrix **D** (Table 1) for the value of 0.182.

Linear interpolation gives $\mathbf{D}^*$=var(e^*)= 0.0746+(0.0293-0.0746)/0.10*(0.182-0.1)=0.0376. The covariances (**C**) are also obtained by linear interpolation of matrix **D** (Table 1) as

$$\mathbf{C} = \begin{pmatrix} 0.0501+(0.0223-0.0501)/0.10*(0.182-0.1) \\ \vdots \\ -0.00903+(-0.00774+0.00903)/0.10*(0.182-0.1) \end{pmatrix} = \begin{pmatrix} 0.0274 \\ \vdots \\ -0.00798 \end{pmatrix}.$$

The stand effects of model 14 are predicted as (Equation 12)

$$\hat{\mathbf{e}} = \begin{pmatrix} 0.0274 \\ \vdots \\ -0.00798 \end{pmatrix} (0.0376 + 0.0577)^{-1} \begin{pmatrix} 0.0274 & \cdots & -0.00798 \end{pmatrix} (2.140 - 2.595) = \begin{pmatrix} -0.131 \\ \vdots \\ 0.0381 \end{pmatrix}.$$

The localized percentiles are obtained by adding the stand effects to their conditional expectations (see Figure 1).

The next step in the calculations would be to calculate the values of **p*** and **R** again by assuming that the sample has been drawn from the localized distribution and to iterate this until convergence. However, in this example the iteration is not carried out.

Table 1. The within-stand variance covariance matrix (**D**) of the percentile models of Kangas and Maltamo (2000a) for Norway Spruce (Kangas, A., personal communication).

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
d_1	0.162										
d_2	0.0501	0.0746									
d_3	0.0223	0.0349	0.0293					(symm)			
d_4	0.0107	0.0156	0.015	0.0142							
d_5	0.00689	0.00877	0.00935	0.00933	0.0098						
d_6	0.00021	-0.00269	-0.00265	-0.00153	-0.00093	0.00319					
d_7	-0.00274	-0.0051	-0.00395	-0.00245	-0.00133	0.00301	0.00592				
d_8	-0.00548	-0.00729	-0.00592	-0.00357	-0.00249	0.00296	0.00584	0.00868			
d_9	-0.00655	-0.00814	-0.00656	-0.00447	-0.00328	0.00311	0.00579	0.00835	0.011		
d_{10}	-0.00672	-0.00818	-0.00699	-0.00479	-0.00351	0.00305	0.00597	0.00837	0.011	0.0135	
d_{11}	-0.00982	-0.00903	-0.00774	-0.00541	-0.00366	0.00281	0.00625	0.00863	0.0109	0.0138	0.025

Testing with real data

Arrangement of the test

The calibration was tested in a small dataset consisting of 43 fixed rectangular sample plots from mixed Scots pine-Norway spruce stands. For the test, Norway spruce trees were selected, because the number of Norway spruce sample trees per stand is much greater (42-222) in the data than the number of Scots pine trees (13-168). Furthermore, the Norway spruce data is more challenging than the Scots pine data, because it includes various forms of the diameter distribution, including symmetric, skewed, extremely wide, mound-shaped, bimodal and multi-modal forms.

The dataset has been originally collected by Pukkala et al. (1994) for productivity studies and it has been further used by Kangas and Maltamo (2000b) and Mehtätalo (2004) in testing the performance of diameter distribution prediction algorithms. The size of the sample plots in the data varied from 600 to 3000 m^2, the number of tally trees from 42 to 222, the basal area of Norway spruces from 1.54 to 24.07 m^2/ha and *DGM* from 5.5 to 33.9 cm. Because of the fairly small number of trees in each plot, it was assumed that a slightly smoothed distribution describes the distribution of the stand better than the actual measured one (Droessler and Burk 1989, Maltamo and Kangas 1998). Hence, the actual distribution was smoothed with a Gaussian kernel (Härdle 1990, p.15-20) using bandwidth determined by the function

$$h_m = \frac{w_m}{21}\frac{6.5}{\sqrt{N_m}},$$

where w_m is the width of the actual distribution and N_m is the number of sample trees in stand m (see Mehtätalo 2004). The smoothed tree stock of the original sample plot is subsequently referred to as a stand.

To illustrate the effect of calibration on the accuracy of the predicted diameter distribution of the stand, a varying number of horizontal point samples were simulated in each stand and two sample trees were randomly selected from each plot as sample measurements of sample order statistics. The number of sample plots in a stand was varied systematically from 1 to 6, thus resulting in the number of sample trees from any one stand being 2, 4, 6, 8, 10 or 12. The trees belonging to a sample plot were selected by generating a uniformly (0,1) distributed random number for each tree of the stand and selecting those trees for which the random number was less than the sampling probability of the tree. The localized predictions of diameter distributions were calculated by predicting the stand effects of model 14 using these measurements.

The models of Kangas and Maltamo (2000a) were used in the test. In this test the exact values of the stand characteristics were used in predicting the expected percentiles to guarantee an equally accurate predictions regardless of the number of sample plots. To ensure safe interpolation of matrix **D**, linear interpolation was used; and to guarantee logical behavior of the interpolated variances and covariances, the interpolation of variances was carried out before interpolating the covariances (see the numerical example presented above). Since the models were originally estimated on the logarithmic scale, the calibration was carried out on the same scale. Thus, the logarithmic percentiles and diameter measurements were used instead of the arithmetic ones in all calculations. When transforming the logarithmic percentiles onto the arithmetic scale, the bias caused by the logarithm transformation was corrected by adding half of the prediction variance to the predicted logarithmic percentiles before applying the exponential function (see Lappi 1991). The correction terms were obtained for the expected percentiles from the diagonal of **D** and for the localized percentiles from the diagonal of $\mathrm{var}(\hat{\mathbf{e}} - \mathbf{e})$ (Equation 13).

In some stands, the calibration algorithm failed for two reasons. Firstly, the iteration did not converge before the maximum number of iterations was reached and secondly, there was some iteration step after which the distribution was not monotone and thus the iteration could not be continued. These situations were circumvented in the test by repeating the sampling until the calibration did not fail. However, to get a reliable figure about the performance of the method in practice, the proportion of those cases where the first attempt failed for one of the first two reasons was calculated.

To compare the predicted distributions with the actual one, bias and relative root mean square error of volume (in per cents) were calculated. The volume was calculated in 1 cm classes, using the volume functions of Laasasenaho (1982) with tree *DBH* as the predictor. In addition, an error index proposed by Reynolds et al. (1988) was used in the comparisons to measure the goodness of fit of the distributions. The error index was calculated in 2 cm classes using the basal area as weight. Thus, the error index was the sum of the absolute differences between the actual and predicted basal areas of the diameter classes. The calculation was repeated 100 times to decrease the effect of sampling error in the results.

The calculations were carried out in R, an environment for statistical computing (www.r-project.org, Venables and Ripley 2002). In addition, the numerical integrations needed in the calculation of expectations, variances and covariances of order statistics were carried out with IMSL-subroutines(IMSL 1997), which were linked into R.

Test results

The effect of calibration on the accuracy of the predicted diameter distribution was considerable. Even two sample trees per stand decreased the RMSE of volume from 2.71to 2.52 and the absolute bias of volume from 0.84 to 0.66. The mean of the error index also decreased clearly, from 4.67 to 4.37. Furthermore, increasing the number of sample trees improved the accuracy of the localized distributions steadily (Figure 2) with the result that 12 sample trees reduced the RMSE of volume by 29 per cent and the absolute bias by 62 per cent when compared to the distributions based on expected percentiles. The prediction results were relatively accurate because the exact stand variables were used in the prediction of expected percentiles and all calculations of volume were based only on tree diameter, not on tree height.

To demonstrate the effect of iteration, Figure 2 presents the results for localized distributions after the first iteration step and converged iteration. The iteration had a considerable effect on the accuracy of volume prediction. With two sample trees, the accuracy of volume prediction after the first iteration step was even lower than that of the distributions based on expected percentiles, but with a larger number of sample trees, the first iteration step also improved the accuracy. However, regardless of the number of sample trees, the advantage of iteration for the accuracy of volume predictions was remarkable. The effect of iteration on the error index, on the other hand, was somewhat confusing. With two sample trees, the values of the error indices were almost the same after the first iteration step and converged iteration. With a large number of sample trees, on the other hand, the iteration increased the value of the error index when compared to the value after the first iteration step. I will return to this.

Visual examination of the localized distributions showed that they are realistic (Figure 3). In most cases the localized distributions were more consistent with the form of the actual distribution than the distributions based on expected percentiles were. Furthermore, the calibration algorithm could produce skewed and multimodal distributions in cases where the distribution based on conditional expectations was rather symmetric and unimodal (Figure 3, stands a and b). In addition, calibration usually moved the predicted minimum and maximum diameters in the right direction. The prediction errors of the localized percentiles were smaller than the errors of the expected percentiles (Figure 4).

In some cases the short distance between consecutive percentiles caused quite high peaks in some diameter classes (e.g. stand a in Figure 3). However, the frequencies of the diameter classes beside the peaks were usually low, which compensates for the error caused by peaks. Hence, peaks do not increase the RMSE and absolute bias of stand volume remarkably but they may have an effect on the error index. Visual examination of the localized distributions showed that the peaks were in many cases higher after converged iteration than after the first iteration step. In addition, increasing the number of sample trees increased the occurrence of high peaks. This may be an explanation for the result that if there were a large number of sample trees the error indices after the first iteration step were lower than those after converged iterations. The peaks are in many cases in percentile intervals just below or above the 50th percentile (=*DGM*) and may be a consequence of using the measured *DGM* as the 50th percentile. A peak arises when the measured *DGM* is far from the midpoint between the localized 40th and 60th percentile.

For convergence it was required that the sum of the absolute differences of the percentiles at subsequent iterations was less than 10^{-6}. The

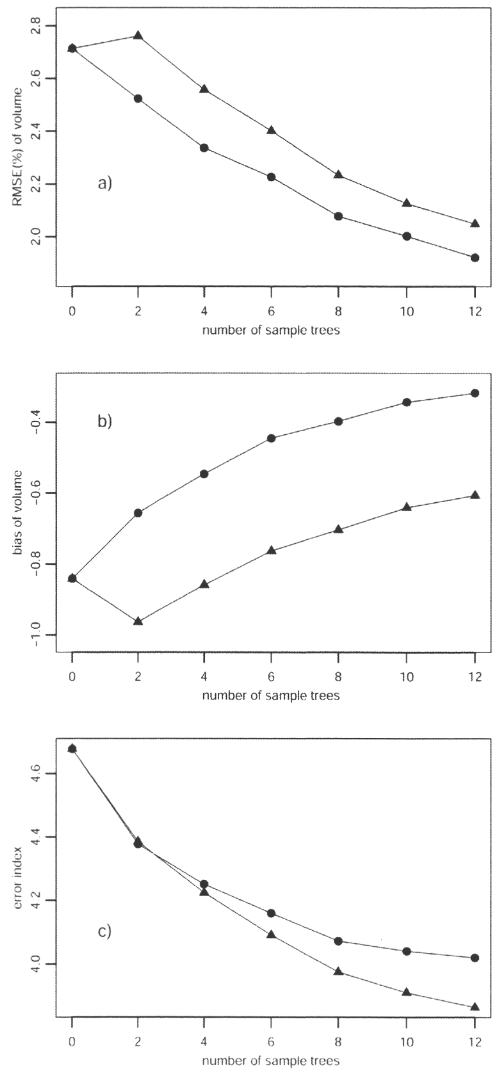


Figure 2. The effect of increasing the number of sample trees on the relative RMSE (a) and bias (b) of volume and the basal area weighted error index (c). The results are presented both using the predictions after the first iteration step (▲) and after converged iterations (●).

maximum number of iterations was 50. A sample with which the calibration was successful could easily be found for all stands. In the following, the proportions of failed calibrations are presented for the first sample attempted. The number of sample trees had a clear effect

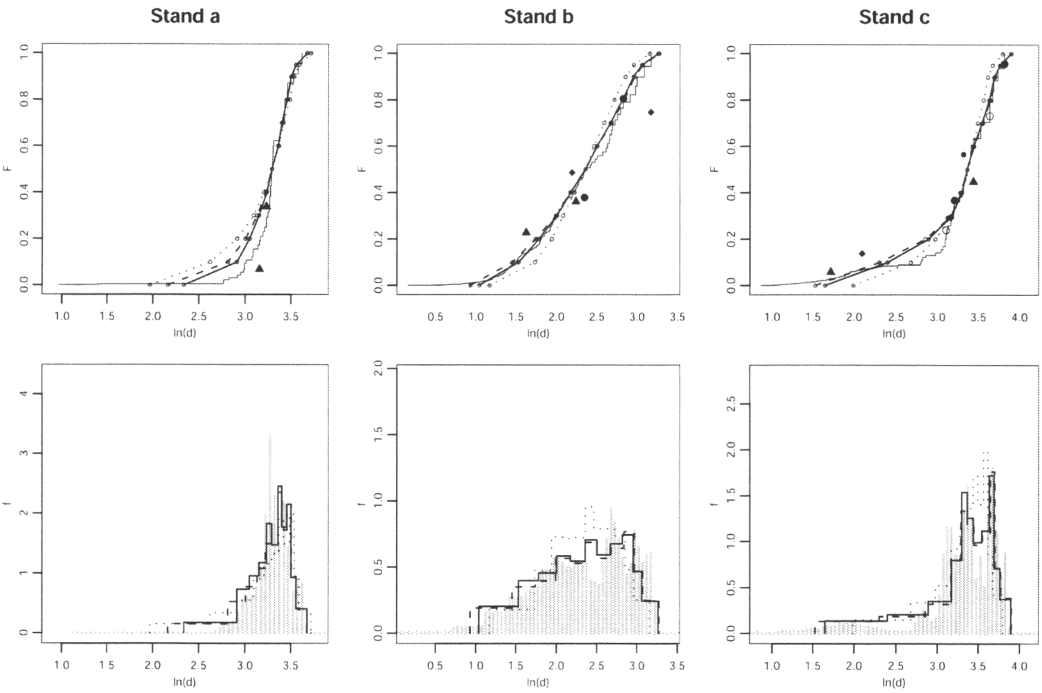


Figure 3. Examples of the predicted and localized distributions with 2, 6 and 10 measured sample trees from 1, 3 and 5 sample plots, respectively. The upper graphs are the cumulative distributions and the lower ones the densities. The dotted lines are the distributions based on expected percentiles, dashed lines the localized distributions after the first iteration step and solid lines the localized distributions after converged iteration. The actual distribution of the stand is the stepped line in each of the upper graphs and the histogram in the lower ones. In the upper graphs, the marks are the measurements (the same symbols are used for measurements from the same sample plot).

on the convergence: with two sample trees per stand the iteration converged in 98% of the stands and the average number of iterations in these stands was 8.28, while with 12 sample trees the proportion of converged iterations was only 88% with an average number of iterations of 12.0. Of the unsuccessfully localized distributions, the iteration did not converge in 7%, and the percentiles were not monotone in some iteration step in the other 93%.

The above calculations were carried out with the Norway spruces of the test data. Because Norway spruce is a shade tolerant tree species, its diameter distribution may have various forms. The same calculations were carried out also with the Scots pines of the test stands. Scots pine is a shade intolerant tree species, which usually has a unimodal diameter distribution. Thus, the diameter distribution of Scots pine is much easier to predict than that of Norway spruce. With Scots pines, RMSE and absolute bias of volume were lower than in the Norway spruce data when using the expected percentiles in the prediction. In addition, the effect of calibration on the RMSE and absolute bias were clearly stronger. The error index was, however, greater in the Scots pine data than in the Norway spruce data, but the effect of calibration was, again, stronger. For example, in the Scots pine data, two sample trees decreased the RMSE of volume from 1.31 to 1.10, absolute bias from 1.03 to 0.63 and mean of the error index from 5.76 to 5.41. Thus, the calibration seems to be clearly useful in both Scots pine and Norway spruce data.

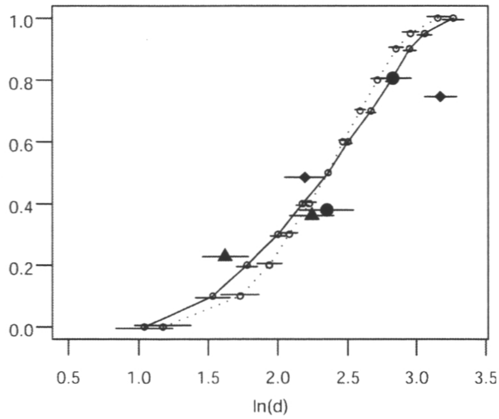


Figure 4. The effect of calibration on the accuracy of the predicted percentiles in stand b of Figure 3. The dotted line is the distribution based on expected percentiles of the stand, the solid line is the localized distribution, and the marks are the sample measurements. The lengths of the horizontal lines at the measurements and predicted percentiles show the standard deviation of the sampling and prediction error, respectively. The standard deviations are obtained from the diagonal elements of the matrices $\mathbf{D}$, $\mathrm{var}(\hat{\mathbf{e}} - \mathbf{e})$ (eqn 11) and $\mathbf{R}$ (eqn 6) for the expected, localized and measured percentiles, respectively.

Discussion

This study presented an algorithm for combining the information from sample order statistics and the predictions of expected diameter percentiles. The test results showed that using this method the accuracy of predicted diameter distribution could be improved considerably with a small number of sample tree measurements. Furthermore, by increasing the number of sample trees, the predicted distribution seemed to approach the actual distribution of the stand.

Because of its flexibility, the percentile-based diameter distribution has been found to be a good alternative in stands with complex diameter distributions (Maltamo et al. 2000). This study showed that using measured order statistics further utilizes the flexibility of the percentile-based method. Thus, the proposed method seems to be a promising alternative for the prediction of diameter distributions in complex stands.

A clear advantage of the method is its flexibility in producing predictions of different accuracy. In other words, using this method, a realistic prediction of the diameter distribution can be achieved with quite a small number of measurements (i.e. even with no sample order statistics measured) and the accuracy of the predicted distribution can be improved steadily by increasing the time used in measuring sample trees. Hence, any realistic predefined accuracy requirement can be obtained with a sufficient number of diameter measurements and, on the other hand, any allotted measurement time can be used effectively for improving the accuracy of prediction. Therefore the method can be useful in various inventories with different accuracy requirements: for example, for tactical forest planning one or two sample trees may produce a sufficiently accurate prediction, while for the pre-harvest measurement of a stand some more measurements may be needed.

The aim of this study was to explain the methodology, demonstrate its use and show how the proposed method works with accurate measurements of order statistics from i.i.d. samples. No comparisons with other methods were carried out, since no methods that utilize the same input information are available. However, the models used in the prediction of expected percentiles have been thoroughly tested (Kangas and Maltamo 2000b). They were found superior when compared to the parameter prediction methods based on the Weibull distribution. Furthermore, the percentile models of Scots pine were found to produce approximately as accurate predictions as the *k*-nearest neighbor method of Maltamo and Kangas (1998). This study showed that predictions based on expected percentiles can further be remarkably improved by the use of measured order statistics.

When deriving the expectations, variances and covariances of order statistics, the sample was assumed to be independent and identically distributed (i.i.d.), i.e. the stand was assumed to be spatially homogeneous. The method was tested with simulated i.i.d. samples, where selecting sample trees was not based on the location of the tree in the stand. This approach was selected to test the method in a situation where the assumptions behind the method are valid. In practice, however, stand characteristics are spatially correlated within a stand. Thus, the test results of this study are an example of localization in a somewhat ideal situation. In a practical situation, where samples are not i.i.d, the advantage achieved by calibration may be smaller than in the test situation of this study. In addition, it may be profitable to distribute the measurements to several plots in a non-homogenous stand to get a spatially more representative sample. A more reliable view of the advantage of the method would be achieved by testing the method in a dataset with real measured sample plots.

Figure 3 showed that the localized distributions may be of various forms, including bimodal and skewed forms. However, for example in stand c of Figure 3, all combinations of sample order statistics did not produce a bimodal distribution. Thus, attention should be paid to the selection of measured order statistics. For example, in a stand where distribution seems to be bimodal, the measurements of maximum diameter of the lower peak and minimum diameter of the upper peak would probably provide more information about the form of the distribution than two randomly selected sample order statistics. Furthermore, measurement of the smallest possible sawtimber tree as a sample order statistic would probably lead to more accurate estimation of sawtimber volume than measurement of a randomly selected tree. Thus, in the future, the usefulness of different strategies in the selection of sample order statistics should be studied.

Since the percentile models used predict the diameter distribution weighted by basal area, the sample plots of this study were horizontal point samples. However, the approach presented can be used as such with an unweighted diameter distribution. In this case, fixed-radius sample plots should be used instead of horizontal point sample plots

Since the measured order statistics are regarded as measured percentiles, the approach of this study requires the percentile-based diameter distribution. However, the approach could be generalized also to other distribution families, when the parameter prediction method is used. This requires derivation of the expectations, variances and covariances of order statistics using the distribution family used. Furthermore, the covariances between the measured percentiles and the parameters of the distribution family should be known. Estimates of them would be obtained by fitting simultaneously models for the parameters of the distribution family and percentiles of the distribution. The required covariances of the measured percentiles and distribution parameters would be obtained by interpolation of the estimated within-stand variance covariance matrix for the p-values of the measurements.

No measurement error was assumed for the sample plot measurements. In practice, however, measurements always include error. When measuring sample order statistics, the error may be either a measurement error of diameters or an error in the determination of the rank of the tree on the plot. The measurement error of diameter is quite easy to take into account by adding the variance of measurement error on the diagonal of matrix **R** (see Lappi 1991). The error in the determination of rank, on the other hand, is somewhat more complicated. It can be regarded as a measurement error of diameter, the magnitude of which is related to the difference between the diameters of subsequent trees in the sample plot. Hence, the measurement error of rank should increase the diagonal elements of **R** proportionally to the absolute difference between subsequent diameters on the sample plot. However, most errors in determining the rank presumably happen with trees that have diameters close to each other. In this case the effect of measurement error is quite small and should not cause serious problems.

In practice, the predictors of percentiles include measurement or sampling error, which may be large. Therefore, in addition to the error caused by the between stand variation, the predictions of expected percentiles also include error variation and bias caused by the measurement error of the predictors. Taking this error into account would require derivation of prediction error variances and covariances of the expected percentiles as well as some kind of bias correction on the expected percentiles. Since the aim of this study was merely to introduce the concept of combining the two types of studies, the work required for taking such errors into account was not tackled.

A problem with the proposed method was that the localized distribution sometimes had peaks around *DGM*. A simple solution to the peak problem would be to use linear interpolation between the localized 40th and 60th percentiles instead of using *DGM* as the localized 50th percentile. Another, more elegant solution would be to use the measurement error variance of the *DGM* as the within-stand variance of the 50th percentile and derive its effect on matrix **D**. Thus, taking the effect of measurement errors of stand variables into account would probably solve the peak problem.

If a practical situation had been simulated, the expected percentiles should have been predicted using stand variables calculated from the sample plots. However, it would have affected the accuracy of the expected percentiles since the number of sample plots varied in different simulations. To make the test show only the effect of calibration on the accuracy of the predicted distribution the exact stand characteristics were used instead of the means of the sample plots.

The procedure used with the unsuccessfully localized distributions was to pick a new sample from the stand and try again, in the test situation of this study. In applications, a procedure for the situations where the calibration fails is needed. A simple approach would be to drop one or more sample trees from among the measured sample trees and to try to find a solution by using the rest of the sample trees. However, this would waste expensive field information. Because the main reason for failure was the non-monotony of the percentiles, the proportion of failures could be decreased remarkably by ensuring the monotony through formulation of the percentile models.

The calculation of exact expectations, variances and covariances of order statistics with numerical integration was rather slow, especially when the number of measured sample order statistics was high. For example, the calculation of the 4300 sample plots with two sample trees took 1.5 hours with a modern personal computer and increasing the number of sample trees to 12 increased the time requirement to almost 4 hours. In these calculations, IMSL subroutines DCONG and DQDAG (IMSL 1997) were used. With a view to speeding up the computations, the calibration was also carried out using asymptotical expectations and variances of sample order statistics (Reiss 1989, p. 109-110), integrals of which can be calculated analytically. With this method the accuracy of the localized distributions was clearly better than that of the distributions based on expected percentiles, but nevertheless clearly worse than using exact results. This could be expected, since the asymptotical results are quite far from the exact ones with such a sample size that is obtained using relascope factor 1. The time requirement could probably be decreased by adjusting the parameters of the numerical integration algorithm. Furthermore, more effective algorithms might be available and the development of faster computers will eliminate this problem in the future.

However, even though the results of this study represent a somewhat ideal situation and there are many things that could be taken into account in order to improve the method, it is a promising method that brings together the good properties of the parameter prediction method and of the use of sample information.

Acknowledgements

This work was part of the project "Statistical modeling for forest management planning", which was carried out at the Finnish Forest Research Institute and funded by the Academy of Finland (decision number 73392). I would like to thank Dr Juha Lappi for the clarifying discussions we had during this work. In addition, I am grateful to Professor Annika Kangas, Professor Matti Maltamo, Dr Pekka Leskinen, the two anonymous referees and the associate editor for their valuable comments on the manuscript. Finally, I wish to thank Dr Lisa Lena Opas-Hänninen for revising my English.

Literature cited

Borders, B.E., Souter, R.A., Bailey, R.L. and Ware, K.D. 1987. Percentile-Based Distributions Characterize Forest Stand Tables. For. Sci. 33(2): 570-576.

Borders, B.E. and Patterson, W.D. 1990. Projecting Stand Tables: A Comparison of the Weibull Diameter Distribution Method, a Percentile-Based Projection Method, and a Basal Area Growth Projection Method. For. Sci. 36(2): 413-424.

Burk, T.E. and Newberry, J.D. 1984. A Simple Algorithm for Moment-Based Recovery of Weibull Distribution Parameters. For. Sci. 30(2): 329-332.

Cao, Q.V. and Burkhart, H.E. 1984. A Segmented Distribution Approach for Modeling Diameter Frequency Data. For. Sci. 30(1): 129-137.

Casella, G. and Berger, R.L. 2002. Statistical Inference, Second Edition. Duxbury Advanced Series, Pacific Grove. 660 p.

Droessler, T.D. and Burk, T.E. 1989. A Test of Nonparametric Smoothing of Diameter Distributions. Scand. J. For. Res. 4: 407-415.

Hyink, D.M. and Moser, J.W. Jr. 1983. A Generalized Framework for Projecting Forest Yield and Stand Structure Using Diameter Distributions. For. Sci. 29(1): 85-95.

Härdle, W. 1990. Smoothing Techniques with Implementation in S. Springer Series in Statistics. Springer-Verlag, New York. 256 p.

IMSL 1997. Fortran Subroutines for Mathematical Applications. Math/Library. Volumes 1&2. Visual Numerics. 1218 p. + Appendices.

Kangas, A. and Maltamo, M. 2000a. Percentile-Based Basal Area Diameter Distribution Models for Scots Pine, Norway Spruce and Birch Species. Silva Fenn. 34(4): 371-380.

Kangas, A. and Maltamo, M. 2000b. Performance of Percentile Based Diameter Distribution Prediction and Weibull Method in Independent Data Sets. Silva Fenn. 34(4): 381-398.

Kilkki, P. and Päivinen, R. 1986. Weibull-function in the estimation of the basal area dbh-distribution. Silva Fenn. 20(2): 149-156.

Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Comm. Inst. For. Fenn. 108. Finnish Forest Research Institute, Helsinki, Finland. 74 p.

Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of Scots pine. Comm. Inst. For. Fenn. 134: 1-69.

Lappi, J. 1991. Calibration of Height and Volume Equations with Random Parameters. For. Sci. 37(3): 781-801.

Lappi, J. 1997. A Longitudinal Analysis of Height/Diameter Curves. For. Sci. 43(4): 555-570.

Lindsay, S.R., Wood, G.R. and Woollons, R.C. 1996. Stand table modelling through the Weibull distribution and usage of skewness information. For. Ecol. and Manag. 81: 19-23.

Liu, C., Zhang, L., Davis, C.J., Solomon, D.S. and Gove, J.H. 2002. A Finite Mixture Model for Characterizing the Diameter Distributions of Mixed-Species Forest Stands. For. Sci. 48(4): 653-661.

Maltamo, M. and Kangas, A. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28: 1107-1115.

Maltamo, M., Kangas, A., Uuttera, J., Torniainen, T. and Saramäki, J. 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogenous Scots pine stands. For. Ecol. and Manag. 133: 263-274.

McCulloch, C. E. and Searle, S. R. 2001. Generalized, Linear and Mixed Models. John Wiley & Sons. 325 p.

Mehtätalo, L. 2004. An algorithm for ensuring compatibility between estimated percentiles of diameter distribution and measured stand variables. For. Sci. 50(1): 20-32.

Nepal, S.K. and Somers, G.L. 1992. A Generalized Approach to Stand Table Projection. For. Sci. 38(1): 120-133.

Pienaar, L.V. and Harrison, W.M. 1988. A Stand Table Projection Approach to Yield Prediction in Unthinned Even-Aged Stands. For. Sci. 34(3): 804-808.

Pukkala, T. Vetteenranta, J, Kolström, T. and Miina, J. 1994. Productivity of mixed stands of Pinus sylvestris and Picea abies. Scandinavian Journal of Forest Research 9: 143-153.

Reiss, R.-D. 1989. Approximate Distributions of Order Statistics With Applications to Nonparametric Statistics. Springer Series in Statistics. Springer-Verlag, New York. 355 p.

Rennols, K., Geary, N. and Rollison, T.J.D. 1985. Characterizing diameter distribution by the use of Weibull distribution. Forestry 58(1): 57-66.

Reynolds, M.R. Jr., Burk, T.E., Huang, W-C. 1988. Goodness-of-fit Tests and Model Selection Procedures for Diameter Distribution Models. For. Sci. 34(2): 373-399.

Tang, S., Wang, Y., Zhang, L. and Meng, C-H. 1997. A Distribution-Independent Approach to Predicting Stand Diameter Distribution. For. Sci. 43(4): 491-500.

Van Deusen, P.C. 1986. Fitting Assumed Distributions to Horizontal Point Sample Diameters. For. Sci. 32(1): 146-148.

Venables, W.N. and Ripley, B D. 2002. Modern Applied Statistics with S. Fourth edition. Springer-Verlag, New York. 495 p.

II

An Algorithm for Ensuring Compatibility Between Estimated Percentiles of Diameter Distribution and Measured Stand Variables

Lauri Mehtätalo

ABSTRACT. It is difficult to formulate a diameter distribution model that is compatible with many stand variables. In previous studies, compatibility of diameter distribution has been ensured with the aid of calibration (adjustment) based on making small changes to the predicted frequencies of diameter classes. In these methods, the minimum and maximum diameters cannot be changed, and the measurement error of the stand variables is not taken into account. In this study, two calibration methods based on minimizing deviations from predicted percentiles were developed. Because minimum and maximum diameters were among the predicted percentiles, there were no problems in changing them in the calibration. In the first method, the measurement error of the stand variables was not taken into account. In the second method, in addition to deviations from predicted percentiles, deviations from the measured stand variables were allowed, and the measurement error was taken into account by weighting each term of the objective function inversely by its error variance. The methods were tested in a dataset of Finnish mixed coniferous forests. Both methods were found to be better than the reference method used because the minimum and maximum diameters could be changed. Even if the measurement error was large, the second method was still advantageous, while the other methods were of no use. For. Sci. 50(1):20–32.

Key Words: Diameter distribution, compatibility, calibration, nonlinear optimization, measurement error, percentile.

Diameter distribution is an essential stand characteristic when assessing, for example, volume, basal area, or stem number. It enables the calculation of the volumes of trees between certain diameter limits, thus providing a tool for assessing, for example, the volume and monetary value of logs obtained in a harvest. The diameter distribution can also be projected into the future, enabling the prediction of future incomes and the effect of different harvest schedules on these. The results can then be compared with each other in a forest management planning process. The accurate description of the stand structure is very important when simulating stand development for forest management planning, and in long-term simulations the small trees are also of interest.

Different theoretical distributions have been used for describing the diameter distribution of a stand. These include, for example, the lognormal (Bliss and Reinker 1964), Weibull (Bailey and Dell 1973), beta (Loetsch et al. 1973, p. 48–61) and Jonson's SB (Hafley and Schreuder 1977) distributions. In recent decades, the Weibull distribution has been the most commonly used because of its flexibility, the fairly straightforward interpretation of its parameters, and the closed

Lauri Mehtätalo, Researcher, Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, FIN-80101 Joensuu, Finland—Phone: +358-10-2113051; Fax: +358-10-2113113; E-mail: lauri.mehtatalo@metla.fi.

Acknowledgments: This work was part of the project "Statistical modeling for forest management planning," which was carried out at the Finnish Forest Research Institute and funded by the Academy of Finland (decision number 73392). I would like to thank Juha Lappi, Annika Kangas, Matti Maltamo, and the three anonymous referees for their valuable comments on the earlier drafts of the manuscript. Professor Kangas also provided the computer program for the reference method. Finally, I wish to thank Lisa Lena Opas-Hänninen for revising my English.

Manuscript received May 5, 2002, accepted May 16, 2003.

form of its cumulative distribution function (Bailey and Dell 1973). Many methods for estimating the parameters of the Weibull distribution have been presented. The maximum likelihood method or the method of moments can be used if a sample dbh-distribution from the stand is available (Van Deusen 1986, Lindsay et al. 1996). If a sample is not available, the parameters can be predicted with the parameter prediction method (PPM) or estimated using the parameter recovery method (PRM). In PPM, the parameters are estimated by developing regression models with stand characteristics as independent variables, and parameters of a target stand are predicted using measured stand variables. In PRM, the parameters of the distribution are recovered from some known or predicted stand characteristics. Recovery can be based on the moments of the distribution (Burk and Newberry 1984), known characteristics of certain percentiles (Bailey and Dell 1973, Shiver 1988) or some measured stand variables (Ek et al. 1975, Hyink and Moser 1983). When predicting the future distribution of a stand, the development of the parameters, stand variables, moments or percentiles is modeled as a function of time.

It has been demonstrated that theoretical distributions work well in even-aged unthinned forests; but in more heterogeneous forests, distributions which are purely parametric have been found to be too rigid (e.g., Cao and Burkhart 1984, Maltamo and Kangas 1998) and more adaptable methods are needed. In a percentile-based method (Borders et al. 1987), the diameter distribution of a stand is described with percentiles of the distribution. The percentiles can be predicted using known stand characteristics, and the development of the percentiles with respect to stand age can also be modeled (Borders and Patterson 1990). The continuous distribution function between the percentiles can be interpolated either with linear interpolation (Borders et al. 1987) or with an adequate spline function (Maltamo et al. 2000).

If a measured sample of tree diameters is available, the sample distribution can be used as such in calculating present volume. If the sample is small, the sample distribution can be smoothed, for example with a kernel method, to obtain a distribution that better describes the underlying population distribution (Droessler and Burk 1989, Maltamo and Uuttera 1998). Fitting a theoretical distribution to the sample distribution can have a detrimental effect: if the population distribution does not follow the theoretical distribution used, some of the information on the form of the distribution is lost (Borders and Patterson 1990, Nepal and Somers 1992). A good estimate of the future distribution may be obtained by projecting the present sample diameters and stand variables into the future with individual tree growth models (Pienaar and Harrison 1988, Nepal and Somers 1992, Tang et al. 1997) and adjusting the projected diameter distribution to obtain a stand description compatible with the projected stand variables.

The incompatibility of the distribution may be a problem in those cases where the distribution is projected into the future with models (Nepal and Somers 1992, Cao and Baldwin 1999), or if all the measured stand variables have not been used in the prediction of the diameter distribution, or if the diameter distribution model used does not guarantee compatibility with all the stand variables used as predictors (Kangas and Maltamo 2000a). In such cases, calibration (adjustment) of the diameter distribution can be useful. In the calibration, the predicted frequencies of the diameter classes are adjusted slightly to obtain a distribution compatible with all the measured or projected stand variables. This approach has two drawbacks. The first is that new diameter classes are not constructed, and problems arise if the minimum and maximum diameters of the uncalibrated distribution are far from the true diameters. The other drawback is that the stand characteristics used are assumed to be accurate, i.e., no measurement or prediction error is assumed.

In Finland, forest management planning is based on stand-wise inventories carried out approximately every tenth year, and stand development is predicted with individual tree growth models using representative trees picked from the predicted diameter distribution of the stand (Kilkki et al. 1989). The costs of the inventories are kept low to make forest planning inexpensive. Therefore, rather than measuring a sample diameter distribution to predict the stand diameter distribution, the assessed stand characteristics are used. The basal-area weighted form of the diameter distribution (basal area diameter distribution) (see Gove and Patil 1998) is used to ensure the accurate description of the large, valuable trees. In the inventories, an angle-gauge is used to determine the basal area and basal area median diameter of a few, subjectively chosen sample plots, and the stand variables are calculated using the sample plot measurements. Due to within-stand variance and measurement error, the total error in the measurement may be as high as 20% of the mean values of the variables (Poso 1983, Laasasenaho and Päivinen 1986, p. 9).

Depending on site features, damage, the forest owner's activity, etc., Finnish forest stands may differ markedly from each other. One stand may have dense natural undergrowth, while in another there may be no undergrowth at all; a stand may be a mixed forest or a one-species forest; it may comprise many age cohorts or be even-aged; it may be naturally regenerated, planted, or sowed, etc. In this situation, it might be advantageous to change the set of assessed variables according to the type of forest stand. However, it would require a large dataset and much work to model diameter distributions with all measurement combinations for all types of stands in order to determine which variables work in any given type of stand. It would also be very hard to formulate models that produce compatible distributions with all measurement combinations. A simpler approach is to predict the distribution with commonly measured stand variables and to take other variables into account through calibration.

The aim of this study was to develop a calibration method that could be used in Finnish forest management planning. This meant that (1) the method could not prohibit the formulation of new diameter classes and (2) the measurement error of the stand characteristics had to be taken into account. The calibration was based on the percentiles of the diameter distribution instead of the frequencies of the diameter classes.

At the end of the study, the effect of calibration in the predicted diameter distributions is demonstrated with a case study.

The Method

A diameter distribution compatible with the measured basal area median diameter and basal area is obtained by scaling the basal area diameter distribution with the measured basal area of the calculation unit, i.e., by multiplying the two. If additional stand characteristics have been measured, e.g., stem number or mean diameter, the resulting distribution may not be consistent with all the measurements. This study proposes a method for calibrating the predicted percentiles. The derivations are presented for the basal area diameter distribution, but they can be easily reformulated for a stem number diameter distribution.

Calibration of the Percentiles of a Basal-Area Diameter Distribution

Let us assume that the diameter distribution is described using percentiles of the basal area diameter distribution and the cumulative distribution is linear between one percentile and the next (Figure 1). Denote the ordered set of M percentiles by $\{d_1, d_2, \ldots, d_m\}$ and the corresponding percentage values by $\{p_1, p_2, \ldots, p_m\}$, where $p_1 = 0$ and $p_m = 100$. (The usual notation, e.g., $d_{50\%}$, is used to show exactly which percentile is under consideration.) Suppose that we have predictions of the percentiles and denote them by $\{\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_M\}$. In addition, we may have measurements of some stand variables. When presenting the calibration methods, it is assumed that the basal area G, basal area median diameter d_{Gm} and stem number N have been measured. Denote these measurements by $\hat{G}$, $\hat{d}_{Gm}$ and $\hat{N}$, because they are from a sample and are therefore estimates of the true values. Let us assume that the measurements are unbiased, i.e., $E(G) = G$, $E\left(\hat{d}_{Gm}\right) = d_{Gm}$ and $E\left(\hat{N}\right) = N$.

We want to modify the predicted percentiles to make the predicted distribution consistent with the measured stand variables. Here it is also assumed that the stand variables are known exactly, i.e., they are assumed to be measured without error. The problem is, in fact, to find a calibrated distribution that is as close to the predicted distribution as possible and concurrently satisfies the calibrating equations. This leads us to formulate the calibration as an optimization problem, where the distance between the predicted and the calibrated distribution is minimized, and the measured stand characteristics are included in the optimization problem as constraints.

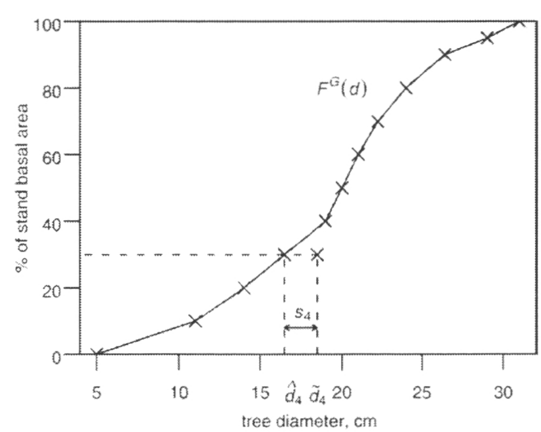


Figure 1. An example of a cumulative predicted percentile-based distribution and the predicted and calibrated fourth percentile.

First, for each predicted diameter, we define a deviational variable s_i, which implies the deviation of the calibrated ith percentile from the predicted percentile. The calibrated ith percentile is $\tilde{d}_i = \hat{d}_i + s_i$.

The optimization problem is
minimize

$$z\left(s_1, s_2, \ldots, s_M\right) \tag{1}$$

subject to

$$\hat{d}_i + s_i + \varepsilon \le \hat{d}_{i+1} + s_{i+1}, \quad i = 1, 2, \ldots, M-1, \tag{2}$$

$$\hat{d}_1 + s_1 + \varepsilon \ge 0, \tag{3}$$

$$\frac{4\hat{G}}{100\pi} \times \sum_{i=1}^{M-1}\left[\frac{p_{i+1} - p_i}{\left(\hat{d}_{i+1} + s_{i+1}\right) - \left(\hat{d}_i + s_i\right)}\left(\frac{1}{\left(\hat{d}_i + s_i\right)} - \frac{1}{\left(\hat{d}_{i+1} + s_{i+1}\right)}\right)\right] = \hat{N}, \tag{4}$$

$$s_{50\%} = 0, \tag{5}$$

where $s_{50\%}$ is the deviational variable corresponding to $d_{50\%}$. The objective function (1) measures the distance between the calibrated and predicted diameter distribution, the constraints (2) and (3) are monotony constraints and constraints (4) and (5) are calibration constraints determining the relations between the measured stand variables. A detailed discussion on the optimization problem follows.

To obtain a monotone distribution, the calibrated percentile $\tilde{d}_i$ must be less than the calibrated percentile $\tilde{d}_{i+1}$ for $i = 1, \ldots, M-1$. Thus we get 11 monotony constraints, one for each percentile interval. In optimization, because we have to use the operation less than or equal to instead of the operation less than, we need to add a small constant implying the smallest difference allowed between two consecutive percentiles (e.g., 0.1 cm), denoted by ε in constraints (2) and (3). Note that it is not required that the predicted set of percentiles be monotone for calibration to be possible. Since calibrated percentiles are always monotone, the procedure can, as a byproduct, make a nonmonotone set of percentiles monotone.

In this situation we need two calibration constraints: the first for the relation between the stem number and the basal area and the second for the basal area median diameter. The first calibration constraint [Equation (4)] is formulated as $\tilde{N} = \hat{N}$, where $\tilde{N}$ is the stem number of the calibrated distribution. (The notation ^ is used on the measured or predicted variables and the notation ~ on the calibrated ones). The calibrated stem number is calculated using Equations (A7) and (A8) in the Appendix, substituting the calibrated values $\tilde{d}_i = \hat{d}_i + s_i$ and $\tilde{d}_{i+1} = \hat{d}_{i+1} + s_{i+1}$ for d_i and d_{i+1}, respectively. The second calibration constraint [Equation (5)] is formulated as $\tilde{d}_{50\%} = \hat{d}_{Gm}$, where $\hat{d}_{Gm}$ is the measured median of the distribution. Letting the predicted median equal the measured median we get Equation (5). If we have other measurements connected with the distribution, they can be formulated as additional calibration constraints in the problem. Some examples of these are presented later.

We require that in the objective function (1), (a) negative and positive deviations of the same absolute value cause equal increase in the value of the objective function, and (b) the objective function takes the prediction error of the percentiles into account. Estimates of the prediction errors of the percentiles (and later also the covariances between consecutive percentiles) are supposed to be known from the modeling stage of the percentiles. As a first trial for the objective function we examine the function

$$z = \sum_{i=1}^{M} \frac{s_i^2}{\sigma_i^2}, \tag{6}$$

where $\sigma_i^2 = \mathrm{var}\left(\hat{d}_i - d_i\right)$ is the prediction error variance of d_i. In this formulation, the sign of the deviational variable does not matter, and the increase caused in the objective function by the change of a percentile is inversely proportional to its prediction error. However, this function will move the extreme percentiles rather than the intermediate ones. This is because the deviations are assumed to be independent. However, the error terms of the percentile models are strongly correlated (for an example, see Table 1a), because each percentile model explains partly the location of the distribution and partly the shape of the distribution. Thus seemingly good weights in the objective function may cause bad results since the objectives of a multiple objective optimization problem are strongly correlated (Steuer 1986, p. 198). The correlation can be made less strong by dividing the distribution into the location and the shape of the distribution. This is done by reformulating the deviational variables as differences between consecutive percentiles and letting the median diameter $d_{50\%}$ explain the location of the distribution. The absolute correlations between the prediction errors of the differences are markedly lower than those of the predicted percentiles (Table 1b). The objective function becomes

$$z = \sum_{i=1}^{M-1} \frac{\left(s_{i+1} - s_i\right)^2}{\sigma_{i,i+1}^2}, \tag{7}$$

where

$$\sigma_{i,i+1}^2 = \mathrm{var}\left[\left(\hat{d}_{i+1} - d_{i+1}\right) - \left(\hat{d}_i - d_i\right)\right].$$

The Measurement Error of Stand Variables

In the previous section, the calibration did not allow the calibrated values of stand characteristics to deviate from their measured values. However, as discussed before, in practice the measurement (or prediction) error may be significant. If the distribution is calibrated strictly to the erroneous measurements, the calibration may not succeed, or the calibrated

Table 1. The estimated standard deviations (diagonal) and correlations (off diagonal) of the prediction error of the percentiles (a) and the difference between consecutive percentiles (b) in the Scots pine models of Kangas and Maltamo (2000b) without stem number. The basal area median diameter ($d_{50\%}$) is assumed to be measured without error.

(a)	$d_{0\%}$	$d_{10\%}$	$d_{20\%}$	$d_{30\%}$	$d_{40\%}$	$d_{50\%}$	$d_{60\%}$	$d_{70\%}$	$d_{80\%}$	$d_{90\%}$	$d_{95\%}$	$d_{100\%}$
$d_{0\%}$	4.06	0.61	0.48	0.35	0.28	0.00	−0.14	−0.18	−0.20	−0.20	−0.25	−0.21
$d_{10\%}$		2.95	0.83	0.69	0.59	0.00	−0.18	−0.27	−0.28	−0.27	−0.29	−0.21
$d_{20\%}$			2.12	0.84	0.70	0.00	−0.15	−0.29	−0.30	−0.29	−0.28	−0.19
$d_{30\%}$				1.49	0.81	0.00	−0.13	−0.28	−0.30	−0.31	−0.29	−0.22
$d_{40\%}$					1.15	0.00	−0.12	−0.21	−0.23	−0.23	−0.21	−0.14
$d_{50\%}$						0.00	0.00	0.00	0.00	0.00	0.00	0.00
$d_{60\%}$							0.83	0.66	0.53	0.41	0.42	0.29
$d_{70\%}$								1.21	0.75	0.68	0.64	0.48
$d_{80\%}$									1.79	0.83	0.77	0.60
$d_{90\%}$										2.60	0.90	0.72
$d_{95\%}$											2.97	0.77
$d_{100\%}$												4.39

(b)	$d_{10\%}$-$d_{0\%}$	$d_{20\%}$-$d_{10\%}$	$d_{30\%}$-$d_{20\%}$	$d_{40\%}$-$d_{30\%}$	$d_{50\%}$-$d_{40\%}$	$d_{60\%}$-$d_{50\%}$	$d_{70\%}$-$d_{60\%}$	$d_{80\%}$-$d_{70\%}$	$d_{90\%}$-$d_{80\%}$	$d_{95\%}$-$d_{90\%}$	$d_{100\%}$-$d_{95\%}$
$d_{10\%}$-$d_{0\%}$	3.26	−0.07	−0.04	−0.07	−0.18	0.01	−0.04	0.02	0.02	0.10	0.06
$d_{20\%}$-$d_{10\%}$		1.69	0.15	0.06	0.15	0.13	0.03	0.05	0.05	0.13	0.03
$d_{30\%}$-$d_{20\%}$			1.18	0.12	0.23	0.11	0.12	0.08	0.03	0.07	−0.04
$d_{40\%}$-$d_{30\%}$				0.87	0.07	0.06	0.20	0.11	0.14	0.05	0.07
$d_{50\%}$-$d_{40\%}$					1.15	0.12	0.17	0.13	0.13	0.02	−0.01
$d_{60\%}$-$d_{50\%}$						0.83	−0.04	0.13	0.08	0.14	0.01
$d_{70\%}$-$d_{60\%}$							0.90	0.04	0.32	0.00	0.08
$d_{80\%}$-$d_{70\%}$								1.19	0.09	0.03	0.11
$d_{90\%}$-$d_{80\%}$									1.49	−0.03	0.17
$d_{95\%}$-$d_{90\%}$										1.32	−0.02
$d_{100\%}$-$d_{95\%}$											2.82

distribution may have abnormally high peaks in some diameter classes. Therefore it is better to calibrate the distribution only approximately to the measured values. In this section the optimization problem is modified to take the measurement error into account.

In addition to the deviational variables of the percentiles, deviational variables for the measured stand characteristics are introduced. The calibrated basal area and stem number are defined as $\tilde{G} = \hat{G} + s_G$ and $\tilde{N} = \hat{N} + s_N$, where the s-variables are the corresponding deviational variables.

The optimization problem is now

minimize

$$z = \sum_{i=1}^{M-1} \frac{(s_{i+1} - s_i)^2}{\sigma_{i,i+1}^2} + \frac{s_{50\%}^2}{\sigma_{50\%}^2} + \frac{s_G^2}{\sigma_G^2} + \frac{s_N^2}{\sigma_N^2} \quad (8)$$

subject to

$$\hat{d}_i + s_i + \varepsilon \le \hat{d}_{i+1} + s_{i+1}, i = 1, 2, \ldots, M-1, \quad (9)$$

$$\hat{d}_1 + s_1 + \varepsilon \ge 0, \quad (10)$$

$$\frac{4(\hat{G} + s_G)}{100\pi} \times \sum_{i=1}^{M-1} \left[\frac{p_{i+1} - p_i}{(\hat{d}_{i+1} + s_{i+1}) - (\hat{d}_i + s_i)} \left(\frac{1}{(\hat{d}_i + s_i)} - \frac{1}{(\hat{d}_{i+1} + s_{i+1})} \right) \right] = \hat{N} + s_N \quad (11)$$

$$\hat{G} + s_G \ge 0, \quad (12)$$

$$\hat{N} + s_N \ge 0, \quad (13)$$

where

$$\sigma_{50\%}^2 = \text{var}(\hat{d}_{Gm} - d_{Gm}),$$

$$\sigma_G^2 = \text{var}(\hat{G} - G)$$

and

$$\sigma_N^2 = \text{var}(\hat{N} - N)$$

are the measurement error variances of the stand characteristics. The changes to the previous formulation, i.e., Equations (1)–(5), are adding three terms to the objective function (8), substituting the calibrated basal area and stem number for the measured ones in the constraint (11), dropping the basal area median diameter constraint (5) from the problem formulation (no longer needed since we no longer assume that the calibrated median is equal to the measured median) and adding the nonnegativity constraints of the calibrated stand variables (12)–(13).

Additional Stand Variables in the Calibration

In addition to basal area, stem number, and basal area median diameter, we may also have other measured stand variables. This section presents some examples of how additional stand variables are included in the optimization problem.

In most cases, the new stand variables can be included in the optimization problem as additional calibration constraints. However, a nonnegativity constraint for the new calibrated stand variable should be added [cf. Equations (12) and (13)]. The calibration constraints are formulated by calculating analytically the corresponding variable from the diameter distribution and equating it with the calibrated stand variable, i.e., the measured stand variable plus the corresponding deviational variable. Some analytically calculated stand variables are presented in the Appendix. The measurement error of the new variables is taken into account by adding the corresponding term into the objective function, as was done in objective function (8) [cf. objective function (7)]. In the following, there are three examples of new constraints.

The constraint for the mean diameter of the stem number diameter distribution is formulated by equating the expected value of the distribution (A10) with the corresponding measured mean diameter as in

$$\frac{\sum_{i=1}^{M-1} \tilde{a}_i \left[\ln(\tilde{d}_{i+1}) - \ln(\tilde{d}_i) \right]}{\sum_{i=1}^{M-1} \tilde{a}_i \left[\frac{1}{\tilde{d}_i} - \frac{1}{\tilde{d}_{i+1}} \right]} = \bar{d}_N + s_{\bar{d}_N} \quad (14)$$

and the constraint for the mean diameter of the basal-area diameter distribution is formulated by equating the expected value of the basal area diameter distribution (A9) with the corresponding measured mean diameter as in

$$\frac{1}{2} \sum_{i=1}^{M-1} \tilde{a}_i \left[\tilde{d}_{i+1}^2 - \tilde{d}_i^2 \right] = \bar{d}_G + s_{\bar{d}_G}. \quad (15)$$

The constraint for the quadratic mean diameter is obtained similarly by equating the calculated quadratic mean diameter (A12) with its calibrated value as in

$$\sqrt{\frac{1}{\sum_{i=1}^{M-1} \left[\tilde{a}_i \left(\frac{1}{\tilde{d}_i} - \frac{1}{\tilde{d}_{i+1}} \right) \right]}} = \hat{d}_q + s_{d_q}. \quad (16)$$

In Equations (14)–(16), $\tilde{a}_i$ is the slope of the calibrated diameter distribution between the calibrated percentiles $\tilde{d}_i$ and $\tilde{d}_{i+1}$ [see Equation (A5)]:

$$\tilde{a}_i = \frac{1}{100} \frac{p_{i+1} - p_i}{\tilde{d}_{i+1} - \tilde{d}_i}, \; i = 1, 2, \ldots, M-1. \quad (17)$$

The minimum and/or maximum diameter can be used in calibration, simply by using the measured value as the prediction of the minimum and/or maximum diameter. After this, the "predicted" minimum and/or maximum diameter equals the measured one, and the only change to be made to the optimization is to correct the variances $\sigma_{1,2}^2$ and $\sigma_{M-1,M}^2$ in the objective function (7). The variance $\sigma_{1,2}^2$ consists of three terms:

$$\sigma_{1,2}^2 = \mathrm{var}(\hat{d}_1 - d_1) + \mathrm{var}(\hat{d}_2 - d_2) \\ -2\,\mathrm{cov}(\hat{d}_1 - d_1, \hat{d}_2 - d_2). \tag{18}$$

Here it is assumed that the percentile d_1 has been measured and the measurement error is uncorrelated with the prediction error of the percentile d_2. Thus, the first term in Equation (18) is replaced by the measurement error variance of the minimum diameter, the second term remains as before, and the third term is zero. The variance becomes

$$\sigma_{1,2}^2 = \mathrm{var}(\hat{d}_1 - d_1) + \mathrm{var}(\hat{d}_2 - d_2), \tag{19}$$

where $\mathrm{var}(\hat{d}_1 - d_1)$ is the measurement error variance of the minimum diameter. The variance $\sigma_{M-1,M}^2$ can be calculated similarly.

Because consecutive percentiles are correlated, the measured minimum and maximum also include some information about the other percentiles. In Equation (19) this information was not used. Some kind of rescaling of the other percentiles before the calibration might be useful. One could use, for example, a crude linear scaling such as the one used by Kangas and Maltamo (2000c), but a better alternative would be a scaling that takes advantage of the covariance structure of the percentiles.

A Case Study

The effect of calibration on the diameter distributions is demonstrated by the following case study. The predicted percentiles are calibrated with stem number in two ways: first ignoring the measurement error and then taking it into account. The methods are compared with a published calibration method (Kangas and Maltamo 2000a) and also with a method where all the stand variables are already used as explanatory variables in the percentile models (Kangas and Maltamo 2000b).

Data

The empirical data of this study are from mixed pine-spruce (*Pinus sylvestris, Picea abies*) stands (Pukkala et al. 1998). The dataset consists of sample trees measured from 43 rectangular plots of 600 to 3000 m^2 situated on mineral soils. The dataset was separated into two sets based on the tree species, and these parts were calculated separately. Some characteristics of the data are presented in Table 2.

Due to the varying number of measured trees and different plot sizes in the different stands, the measured diameter distributions contain different amounts of sampling error. In this study, it was assumed that a Gaussian kernel-smoothed (Härdle 1990, p.15–20) distribution described the population better than the actual measured distribution. The value of the smoothing parameter (i.e., the standard deviation of the normal distribution used) was determined as a function of the number of measured trees and the width of the distribution, using information from recent studies (Maltamo and Uuttera 1998, Uusitalo 1997, p. 29) and visual examination of smoothed distributions. The selected function was

$$h_s = \frac{w_s}{21}\frac{6.5}{\sqrt{n_s}}, \tag{20}$$

where w_s is the observed width of the measured distribution in cm and n_s is the number of measured trees in stand s. The value of the smoothing parameter varied from 1.06 to 1.89 in the Scots pine subdata and from 0.98 to 2.18 in the Norway spruce subdata.

Predicting and Calibrating the Distribution

The percentile models of Kangas and Maltamo (2000b) were used to predict the diameter distribution. The models predict M = 12 percentiles of the basal area diameter distribution, i.e., the percentiles 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 100. The measured basal area median diameter is used as the 50% point. For each tree species, two model sets with different predictors are presented. The predictors of model set 1 are the stand basal area, basal area median diameter, stand age, and a dummy variable for mineral soil. In model set 2, the stem number is included as an additional predictor. In this study, linear interpolation was used to produce a continuous distribution function between the predicted percentiles.

Table 2. Some characteristics of the data. G_S is the basal area of the tree species and G_T the total basal area of the stand (m^2/ha), N is the stem number (1/ha), V is the total volume (m^3/ha), d_{Gm} is the basal area median diameter (cm), T the stand age (yr) and n the number of measured trees in the stand.

	Scots pine			Norway spruce		
	Min	Mean	Max	Min	Mean	Max
G_S	2.97	14.9	32.3	1.54	10.9	24.1
G_T	4.82	27.3	46.2	4.82	27.3	46.2
N	92.9	391	1600	247	954	2170
V	11.0	100	224	6	79.0	208.3
d_{Gm}	6.47	25.9	35.2	5.53	17.8	33.9
T	19	68.5	102	19	68.5	102
n	13	47.1	168	42	109	222

Table 3. The error scenarios. Error is simulated from the multi-normal distribution using correlation structure as follows: $\mathrm{corr}(e_{dgM}, e_G) = -0.29$, $\mathrm{corr}(e_{dgM}, e_N) = 0.01$ and $\mathrm{corr}(e_N, e_G) = 0.56$.

Scenario	Standard deviation of the error term (% of the true value)		
	σ_6	σ_G	σ_N
I	0.5	0.5	0.5
II	4	5	6
III	8	10	12
IV	12	15	18
V	16	20	24

According to Poso (1983), the within-stand standard deviation of basal area and basal area median diameter may be 32% and 23% of the true value, respectively, and thus the sampling errors using, for example, five sample plots are 14.3% and 10.3%, respectively. In addition, the measurements contain a random measurement error. Because exact information on the error variance was not available, the error was taken into account by using five error scenarios, each with different error variances (Table 3). The standard deviation of the measurement error was assumed to be proportional to the true value of the stand variable, and the errors were assumed to be correlated, because in an inventory, all stand characteristics are measured in the same sample plots. The correlations were obtained from unpublished data. The whole dataset was calculated once with each error scenario.

For each stand the distribution was predicted with both model set 1 and 2. The distribution obtained with model set 1 (m1) was calibrated using three methods. In the first method, called percentile calibration 1 (pc1), the percentiles are calibrated ignoring the measurement error (see Table 4). In the second method (pc2) measurement error is taken into account. The third method was the method of Kangas and Maltamo (2000a), which was used as a reference method. In this method, the frequency table is calibrated to obtain a distribution satisfying the calibration constraints (frequency table calibration, ftc) using the distance function based on the square root (see Kangas and Maltamo 2000a). The calibrations pc1 and ftc are later called "strict calibrations" because they do not take the measurement error into account. The calibrated distributions were compared with the distribution predicted with model set 2 (m2) in order to compare calibrated distributions with a distribution predicted with regression models that use the same information as used in the calibrations. If the calibration did not succeed, or the percentile set obtained with model set 2 was not monotone, the prediction obtained with model set 1 was used as the prediction of the distribution. Because of the rather small number of plots, all calculations were repeated 50 times to diminish the effect of sampling error.

The calibrations of the percentiles were carried out using IMSL subroutine DNCONG (IMSL 1997, p. 1003–1007), which solves a nonlinear optimization problem with nonlinear constraints. The program for the calibration method of Kangas and Maltamo (2000a) and the covariance matrices for the percentile models were obtained from Professor Kangas.

Comparison of the Methods

The goodness of fit of the predicted distributions was measured with the error index proposed by Reynolds et al. (1988). The error index of a diameter distribution is calculated as

$$e = \sum_{i=1}^{K} w_i \left| f_i - \hat{f}_i \right|, \tag{21}$$

where f_i and $\hat{f}_i$ are the true and predicted stem numbers of class i, and w_i is the weight of class i. In this study, tree basal area was used as weight, and hence the error index is the sum of the absolute deviations of the class-wise basal areas. Averages of the error indices in the dataset were calculated in order to compare the calibration methods.

The relative root mean square error of volume (in percent) was calculated as

$$RV = 100 \sqrt{\frac{\sum_{i=1}^{n} \left(V_i - \hat{V}_i\right)}{n}} \Big/ \overline{V}, \tag{22}$$

where V_i and $\hat{V}_i$ are the true and estimated volumes of stand i, and n is the total number of stands. Tree volumes were calculated with the dbh-based volume functions of Laasasenaho (1982, p. 41).

Results

When the measurements included only a slight error, there was no difference between the method that takes the measurement error into account (pc2) and the method that does not (pc1) (Figure 2). However, when the measurement error was increased, the strict calibration method pc1 even in-

Table 4. The methods compared in the case study.

Method	Initial predicted distribution	Calibration method	Changes in the predicted minimum and maximum allowed	Measurement error of stand variables taken into account
pc1	model set 1	Equations (7), (2)–(5)	Yes	No
pc2	model set 1	Equations (8)–(13)	Yes	Yes
ftc	model set 1	Kangas & Maltamo 2001a	No	No
m1	model set 1	—	—	—
m2	model set 2	—	—	—

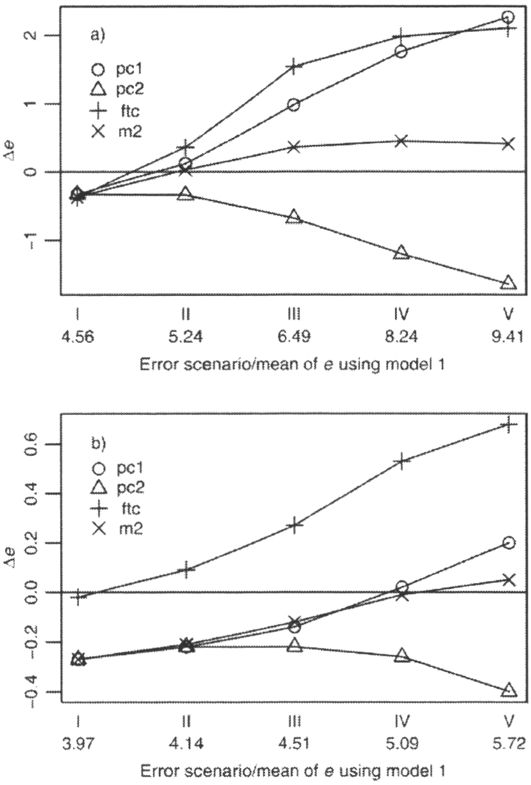


Figure 2. The average of the error index in the Scots pine (a) and Norway spruce (b) subdata sets. The value in the x-axis legend is the mean value of the predicted distributions (method m1) and the lines represent the difference between the other methods and method m1 (see Table 4).

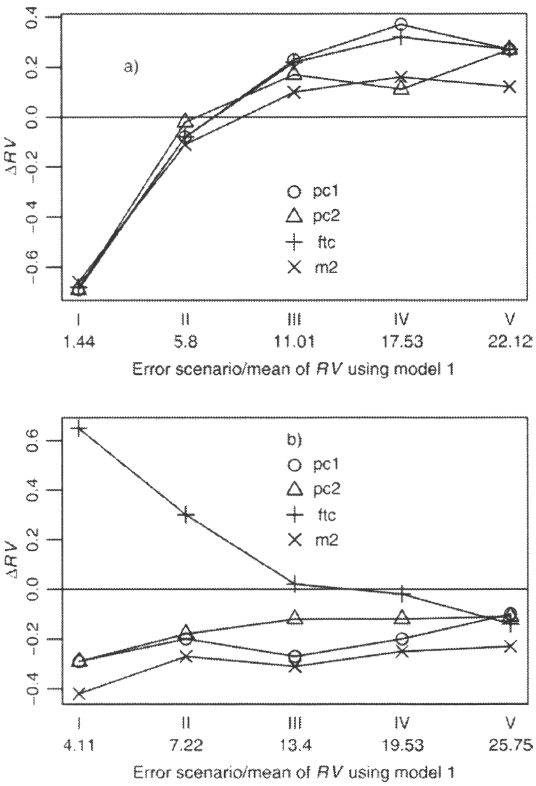


Figure 3. The difference of the relative RMSE of volume between method m1 and the other methods in the Scots pine (a) and Norway spruce (b) data sets.

creased the value of the error index, while the calibration method pc2 clearly decreased it. With respect to the RMSE of volume, no clear differences between the calibration methods can be seen (Figure 3). When the measurement error was small, the RMSE of volume decreased in all calibrations, but when the measurement error increased, the RMSE of volume was even slightly increased by the calibrations.

The error index of the pc2 method was also lower than that of the m2 method, where the stem number was a predictor in the percentile models. The method m2, however, seems to have produced slightly smaller RMSEs of volume than any of the calibration methods.

The percentile calibrations worked better than the reference method (ftc), especially with respect to the Norway spruce data (Figures 1 and 2). The explanation is that the percentile calibrations can produce new diameter classes while the reference method cannot. The trees in the Norway spruce data are much smaller, and the stem number is greater than in the Scots pine data (Table 2). If the true minimum of the distribution is small, its overestimation causes a large underestimation in the stem number of the stand and vice versa. Because the ftc method does not produce new diameter classes, the frequencies of the smallest or largest classes must become very high to satisfy the stem number constraint and this causes peaks in the calibrated distribution (see Figure 4c).

To summarize, if there is very little measurement error in the stand variables, and the distribution models predict the minimum and maximum diameters well, all calibration methods are equally good. If the predicted minimum and maximum differ from the true values, the percentile calibrations are better than the frequency table calibration. Furthermore, when the measurement error of the stand variables increases, the goodness of fit of the distributions is clearly better with the pc2 method than with the strict calibrations (pc1 or ftc) or even with the m2 method, because the pc2 calibration takes the reliability of the measurement into account. Hence, when simulating forest development, it may be advisable to use the assessed stem number as a calibration variable rather than as a predictor of the diameter distribution model. When the present volume is estimated, calibrating with an erroneous stem number seems to be of very little use.

Some examples of diameter distributions obtained with different methods are presented in Figure 4 and Table 5. In subfigure a, the prediction m1 is not very good, but the stand characteristics calculated from the predicted distribution are close to their measured values. Thus, there are no remarkable conflicts between measured stand characteristics and the

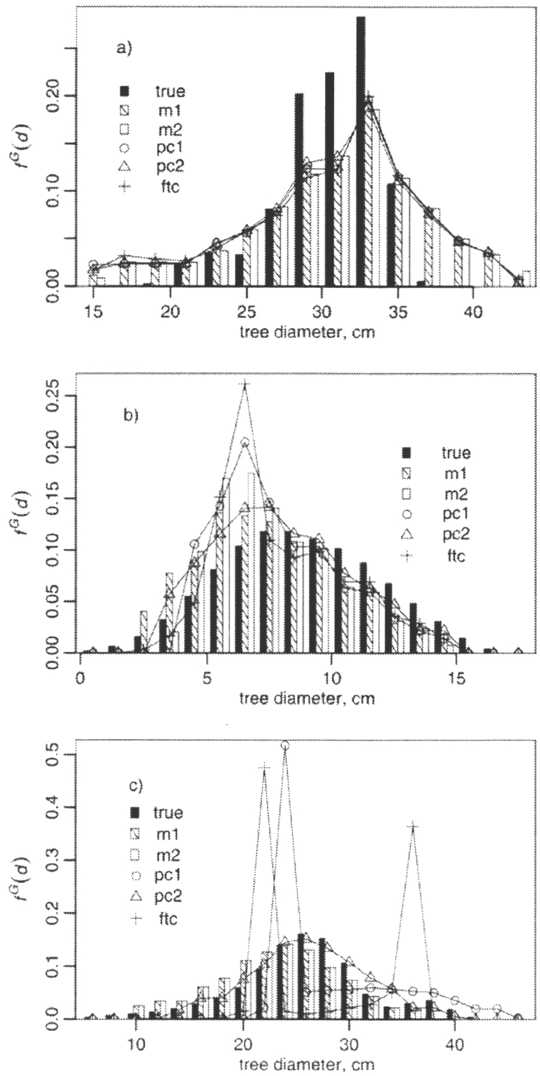


Figure 4. Examples of diameter distribution predictions in three Scots pine stands.

distribution to be calibrated (m1). Hence, the calibrated distributions are quite similar to the predicted m1 distribution. Subfigure b shows quite a typical situation, where the measurements and the initial predicted distribution m1 are mutually conflicting, which causes peaks in the strictly calibrated pc1 and ftc distributions. When differences between the measured and calibrated stand variables are allowed, the calibrated distribution is between the true and initial predicted m1 distributions, and no excessive changes in the form of the distribution have been necessary. In subfigure c, the measurements are strongly conflicting, because the measured stem number and basal area median diameter are clearly underestimates, and the basal area is clearly an overestimate (Table 5). The strict calibrations have moved many trees to the diameter classes just below the basal area median diameter in order to decrease the stem number without changing the basal area median diameter. Model 2 has not been able to produce a nonmonotone percentile set, and thus the prediction of model 1 has been used. The pc2 method has moved all the percentiles to the right and corrected the conflicting measurements; hence the distribution is quite close to the true distribution.

If the measurements are so strongly conflicting that the set of calibration equations is inconsistent, the calibration fails. The unfeasibility of the calibration problem is a serious problem when the measurements contain errors. In the Scots pine data with error scenario V (Table 4), 15.1% of the problems were unfeasible with the ftc method and 3.0% with the pc1 method. The difference between these methods arises from the fact that the predicted minimum and maximum diameters can be changed in pc1. With the pc2 method, the calibration problem was never unfeasible—in practice it is a universal result. It is always possible to find a solution to the calibration problem because each calibration constraint has a deviational variable with no bounds on the right-hand side [and on the left-hand side only the lower bound determined by constraints (12) and (13)].

The calibration taking the measurement error into account changes the values of mutually conflicting stand variables. Hence, with the calibration method pc2, the resulting RMSE of basal area was slightly lower than that of the measurement (Figure 5). The RMSE of the basal area median diameter was reduced by as much as 4% in error scenario V. All calibrations produced an RMSE of stem number considerably smaller than that obtained with model set 1, but only the pc2 method produced an RMSE of stem number approximately equal to that of the measured stem number. With the strict calibrations pc1 and ftc, if the calibration succeeds, the calibrated stem number is exactly equal to the measured stem number. However, when the calibration fails, the initial predicted distribution has to be used and therefore, with the strict calibrations the RMSE of stem number is quite far from the RMSE of the measured stem number.

Because the RMSE of the stand variables was reduced with the pc2 method, the calibrated stand variables can be used as new estimates of the stand variables in the other models of the stand simulator. The distributions of the calibrated basal area, basal area median diameter, and stem number were symmetric and close to a normal distribution.

Table 5. True, measured and calibrated stand variables of the distributions in Figure 4.

	Stand variable								
	Stand (a)			Stand (b)			Stand (c)		
	d_{Gm}	G	N	d_{Gm}	G	N	d_{Gm}	G	N
True	31.7	6.55	92.9	8.43	2.98	786	26.0	17.6	449
Measured	31.8	6.76	107	7.17	2.99	811	23.5	20.3	370
Calibrated (pc2)	31.6	6.79	107	7.67	2.89	843	26.0	19.3	416

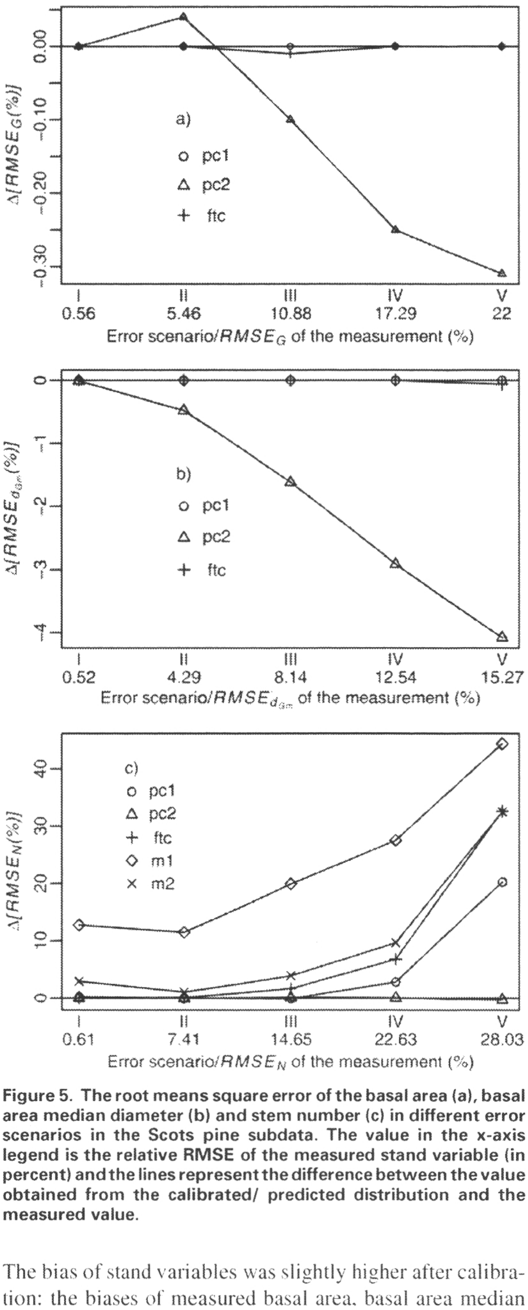


Figure 5. The root means square error of the basal area (a), basal area median diameter (b) and stem number (c) in different error scenarios in the Scots pine subdata. The value in the x-axis legend is the relative RMSE of the measured stand variable (in percent) and the lines represent the difference between the value obtained from the calibrated/ predicted distribution and the measured value.

The bias of stand variables was slightly higher after calibration: the biases of measured basal area, basal area median diameter and stem number in the Scots pine data using error scenario V were 0.09, –0.18, and 2.5, and after calibration they were –0.22, 0.60, and 11.21, respectively. However, the change in the biases is so small that it hardly causes problems in applications.

Discussion

The aim of this study was to develop a calibration method able to produce new diameter classes in the distribution while taking the measurement error of the stand variables into account. When percentiles of the distribution are calibrated instead of the frequencies of a stand table, the minimum and maximum diameters are changed as easily as other percentiles. The reliability of the predicted percentiles can be taken into account in the formulation of the objective function by weighting each term inversely with the corresponding error variance.

When the continuous diameter distribution is interpolated using linear interpolation, many stand variables can be easily derived analytically from the predicted distribution. In calibration they can be formulated as calibration constraints by equating them to their measured values. When deviational variables are used also for the stand variables, the calibrated distributions need not fulfill the measurements exactly. The measurement error of the stand variables can be easily taken into account in the objective function by weighting the corresponding deviational variables with the inverse of the corresponding measurement error variance. Calibration results were presented using the measured basal area, basal area median diameter and stem number as calibration variables, but other stand variables can also be easily included in the calibration, as shown in the section "Additional Stand Variables in the Calibration."

One problem in the formulation of the calibration problem was that the errors of the percentiles and stand variables are correlated. The correlation between the prediction errors of the percentiles was made less strong by reformulating the objective function to minimize the sum of percentile interval widths. This reformulation decreased the correlation coefficients of the terms in the objective function from between 0.61 and 0.90 to between –0.07 and 0.15. The resulting correlation was regarded as so weak that it was not taken into account. The measurement error correlation of stand variables was not taken into account either. In the earlier stages of the study, the Mahalanobis distance $\mathbf{s}'\mathbf{V}^{-1}\mathbf{s}$ was used as a distance function, where $\mathbf{s}$ is a vector of the deviational variables and $\mathbf{V}$ their variance-covariance matrix. This formulation takes the prediction error correlation into account, but it was found to be problematic because it was very vulnerable to the estimation error of the matrix $\mathbf{V}$. In addition to the variances and covariances of the prediction errors of the percentiles, this matrix requires the covariances between the measurement errors of the stand variables and the prediction errors of the percentiles. The effect of the measurement errors in the predictors of the percentiles on the error variances and covariances should also have been estimated.

In this study, two methods for calibrating predicted percentiles of diameter distribution were developed. The first method assumed that the calibration variables were measured without error, and it is thus comparable to previous calibration methods. In the comparisons, this method was found to be a little better than previous methods, and the improvement arose from allowing the predicted minimum and maximum of the distribution to be changed. This method was a good calibration method when there were no measurement errors in the stand characteristics; but when there were errors in the measurements, it worsened the goodness of fit of the distribution, as did the reference method.

In the second method, the error variance of the measurements was taken into account, and the results showed that this was necessary. The effect of calibration on the RMSE of volume was negligible, but the effect on the value of the error index measuring the goodness of fit can be clearly decreased by calibration. This was also seen in the figures of the calibrated distributions. The strictly calibrated distributions produced peaks in the distributions (Figure 4), while the method taking the measurement error into account produced smooth distributions holding the form of the predicted distribution, even if the stand measurements were very conflicting. This method did not have the problem of unfeasibility either, which was a problem in the other methods. With respect to the error index, the calibrated distribution was also better than the distribution predicted with models using the calibration variables as predictors and it seems to be useful to take the erroneous measurements into account through calibration rather than use them as predictors of the diameter distribution model. In addition, it seems to be useful, as well as safe, to use the calibrated stand variables as corrected estimates of the stand variables in the other models of the simulation system.

A problem with this second method is that it can be used effectively only when the initial diameter distribution is described with percentiles. Naturally, percentiles can be calculated from any predicted diameter distribution, but the prediction error of percentiles calculated, for example, from a Weibull distribution is at least very difficult, if not impossible, to compute. In addition to prediction errors of the percentiles, the measurement error variances of the stand variables are needed. These variances should be estimated from some data and the susceptibility of the method to the estimation error in these variances should also be studied. In this study, only basal area, basal area median diameter and stem number were used as calibration variables. The use of other stand variables should be tested. The use of calibrated stand variables in other models of the simulation system also needs to be tested before applying the idea in practice. However, the calibration method pc2 seems to be a promising tool in predicting the diameter distribution for stand simulations.

Literature Cited

BAILEY, R.L., AND T.R. DELL. 1973. Quantifying diameter distributions with the Weibull Function. For. Sci. 19(2):97–104.

BLISS, C.I., AND K.A. REINKER. 1964. A lognormal approach to diameter distributions in even-aged stands. For. Sci. 10(3):350–360.

BORDERS, B.E., R.A. SOUTER, R.L. BAILEY, AND K.D. WARE. 1987. Percentile-based distributions characterize forest stand tables. For. Sci. 33:570–576.

BORDERS, B.E., AND W.D. PATTERSON. 1990. Projecting stand tables: A comparison of the Weibull diameter distribution method, a percentile-based projection method, and a basal area growth projection method. For. Sci. 36(2):413–424.

BURK, T.E., AND J.D. NEWBERRY. 1984. A simple algorithm for moment-based recovery of Weibull distribution parameters. For. Sci. 30(2):329–332.

CAO, Q.V., AND V.C. BALDWIN, JR. 1999. A new algorithm for stand table projection models. For. Sci. 45(4):506–511.

CAO, Q.V., AND H.E. BURKHART. 1984. A segmented distribution approach for modelling diameter frequency data. For. Sci. 30(1):129–137.

DROESSLER, T.D., AND T.E. BURK. 1989. A test of nonparametric smoothing of diameter distributions. Scand. J. For. Res. 4:407–415.

EK, A.R, J.N. ISSOS, AND R.L. BAILEY. 1975. Solving for Weibull diameter distribution parameters to obtain specified mean diameters. For. Sci. 21(3):290–292.

GOVE, J.H., AND G.P. PATIL. 1998. Modelling basal area-size distribution of forest stands: A compatible approach. For. Sci. 44(2):285–297.

HAFLEY, W.L., AND H.T. SCHREUDER. 1977. Statistical distributions for fitting diameter and height data in even-aged stands. Can. J. For. Res. 7:481–487.

HARDLE, W. 1990. Smoothing techniques with implementation in S. Springer Series in Statistics. Springer-Verlag, New York. 256 p.

HYINK, D.M., AND J.W. MOSER, JR. 1983. A generalized framework for projecting forest yield and stand structure using diameter distributions. For. Sci. 29(1):85–95.

IMSL. 1997. Fortran subroutines for mathematical applications. Math/Library. Vols. 1 and 2. Visual Numerics. 1218 p. + Appendices.

KANGAS, A., AND M. MALTAMO. 2000a. Calibrating predicted diameter distribution with additional information. For. Sci. 46(3):390–396.

KANGAS, A., AND M. MALTAMO. 2000b. Percentile-based basal area diameter distribution models for Scots pine, Norway spruce and birch species. Silv. Fenn. 34(4):371–380.

KANGAS, A., AND M. MALTAMO. 2000c. Performance of percentile based diameter distribution prediction and Weibull method in independent data sets. Silv. Fenn. 34(4):381–398.

KILKKI, P., M. MALTAMO, R. MYKKANEN, AND R. PAIVINEN. 1989. Use of the Weibull function in estimating the basal area dbh-distribution. Tiivistelmä: Weibul-funktion käyttö pohjapinta-alan läpimittajakauman estimoinnissa. Silv. Fenn. 23(4):311–318.

LAASASENAHO, J. 1982. Taper curve and volume functions for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae 108:1–74. Finnish For. Res. Inst., Helsinki, Finland. 74 p.

LAASASENAHO, J., AND R. PAIVINEN. 1986. Kuvioittaisen arvioinnin tarkistamisesta. English summary: On the checking of inventory by compartments. Folia Forestalia 664. Finnish For. Res. Inst., Helsinki, Finland. 16 p.

LINDSAY, S.R., G.R. WOOD, AND R.C. WOOLLONS. 1996. Stand table modelling through the Weibull distribution and usage of skewness information. For. Ecol. Manage. 81:19–23.

LOETSCH, F., F. ZÖHLER, AND K.E. HALLER. 1973. Forest inventory 2. BLV Verlagsgeschellschaft. München, Germany. 469 p.

MALTAMO, M., AND A. KANGAS. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28: 1107–1115.

MALTAMO, M., AND J. UUTTERA. 1998. Angle-count sampling in the description of forest structure. For. Landsc. Res. 1:447–471.

MALTAMO, M., A. KANGAS, J. UUTTERA, T. TORNIAINEN, AND J. SARAMÄKI. 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogenous Scots pine stands. For. Ecol. Manage. 133:263–274.

NEPAL, S.K., AND G.L. SOMERS. 1992. A generalized approach to stand table projection. For. Sci. 38(1):120–133.

PIENAAR, L.V., AND W.M. HARRISON. 1988. A stand table projection approach to yield prediction in unthinned even-aged stands. For. Sci. 34(3):804–808.

POSO, S. 1983. Kuvioittaisen arviointimenetelmän perusteita. English summary: Basic features of forest inventory by compartments. Silv. Fenn. 17(4):313–349.

PUKKALA, T., J. MIINA, M. KURTTILA, AND T. KOLSTRÖM. 1998. A spatial yield model for optimizing the thinning regime of mixed stands of *Pinus sylvestris* and *Picea abies*. Scand. J. For. Res. 13:31–42.

REYNOLDS, M.R., JR., T.E. BURK, AND W-C. HUANG. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. For. Sci. 34(2):373–399.

SHIVER, B.D. 1988. Sample sizes and estimation methods for the Weibull distribution for unthinned slash pine plantation diameter distributions. For. Sci. 34(3):809–814.

STEUER, R.E. 1986. Multiple criteria optimization. Wiley series in probability and mathematical statistics-applied. Wiley, New York. 546 p.

TANG, S., Y. WANG, L. ZHANG, AND C-H. MENG. 1997. A distribution-independent approach to predicting stand diameter distribution. For. Sci. 43(4):491–500.

UUSITALO, J. 1997. Pre-harvest measurement of pine stands for sawing production planning. Acta For. Fenn. 259. 56 p.

VAN DEUSEN, P.C. 1986. Fitting assumed distributions to horizontal point sample diameters. For. Sci. 32(1): 146-148.

APPENDIX

The Stem Number of a Basal-Area Diameter Distribution

Let G be the basal area and N the stem number of a stand and $f^G(x)$ the density of the basal-area diameter distribution as a function of the tree diameter x. The density function adds up to unity. Denote the basal area of a tree, whose diameter is x, by $g(x)$. The density function of the stem number is thus (see Gove and Patil 1998)

$$f^N(x) = \frac{1}{N} G f^G(x) / g(x). \tag{A1}$$

Assuming that the basal area G of the stand is known, the stem number between diameters d_1 and d_2 can be calculated as a definite integral of the density function of the stem number:

$$N_{d1,d2} = N \int_{d_1}^{d_2} f^N(x)dx = G \int_{d_1}^{d_2} f^G(x)/g(x)\,dx, \tag{A2}$$

and the number of stems in the stand can be calculated as

$$N = G \int_{d_{min}}^{d_{max}} f^G(x)/g(x)\,dx. \tag{A3}$$

Some Results of a Percentile-Based Diameter Distribution

Assume that the basal area diameter distribution of a stand is described with M percentiles $d_1, d_2, \ldots, d_M$ of certain percentage values $p_1, p_2, \ldots, p_M$ of the cumulative basal area diameter distribution, where p_i is a fixed percentage value implying how large a proportion of the total basal area consists of trees, the diameter of which is less than d_i. Assume further that the cumulative distribution function is linear consecutive percentiles d_i and d_{i+1}

$$F_i^G(x) = a_i x + b_i, \tag{A4}$$

where the slope a_i is calculated as

$$a_i = \frac{1}{100} \frac{p_{i+1} - p_i}{d_{i+1} - d_i}. \tag{A5}$$

The density function is the first derivative of the cumulative distribution

$$f_i^G(x) = F_i^{G\prime}(x) = a_i. \tag{A6}$$

Assuming that the cross-section of a tree is circular in shape and using equations (A2) and (A6), the stem number between diameters d_i and d_{i+1} is calculated as

$$\begin{aligned} N_i &= G \int_{d_i}^{d_{i+1}} \frac{f_i^G(x)}{g(x)} dx \\ &= G \int_{d_i}^{d_{i+1}} \frac{a_i}{\pi \left(\frac{x}{2}\right)^2} dx = \frac{4Ga_i}{\pi} \left(\frac{1}{d_i} - \frac{1}{d_{i+1}} \right). \end{aligned} \tag{A7}$$

The stem number of the distribution is the sum of the stem numbers of all these intervals

$$N = \sum_{i=1}^{M-1} N_i. \tag{A8}$$

The expected value of a basal-area diameter distribution is calculated as

$$\bar{d}_G = \int_{d_{min}}^{d_{max}} x f^G(x) dx$$

$$= \sum_{i=1}^{M-1} \int_{d_i}^{d_{i+1}} x a_i dx = \frac{1}{2} \sum_{i=1}^{M-1} a_i \left(d_{i+1}^{\ 2} - d_i^{\ 2} \right). \tag{A9}$$

The expected value of the stem number diameter distribution is calculated as

$$\bar{d}_N = \int_{d_{min}}^{d_{max}} x f^N(x) = \frac{1}{\sum_{i=1}^{M-1} \int_{d_i}^{d_{i+1}} \frac{a_i}{x^2} dx} \sum_{i=1}^{M-1} \int_{d_i}^{d_{i+1}} \frac{a_i}{x} dx$$

$$= \frac{\sum_{i=1}^{M-1} a_i \left(\ln d_{i+1} - \ln d_i \right)}{\sum_{i=1}^{M-1} a_i \left(\frac{1}{d_i} - \frac{1}{d_{i+1}} \right)}. \tag{A10}$$

The quadratic mean diameter of the stand is the diameter of the tree representing the mean basal area. If the basal area and number of stems of a stand are known, the mean of the tree-wise basal areas is calculated as *G/N* and the diameter of the tree representing this basal area (the quadratic mean diameter) as

$$d_q = \sqrt{\frac{4G}{\pi N}}. \tag{A11}$$

Substituting *G* with the known basal area and *N* with the calculated stem number (Equation A8) we get, after cancellations:

$$d_q = \sqrt{\frac{1}{\sum_{i=1}^{M-1} a_i \left(\frac{1}{d_i} - \frac{1}{d_{i+1}} \right)}}. \tag{A12}$$

III

A longitudinal height–diameter model for Norway spruce in Finland

Lauri Mehtätalo

Abstract: A height–diameter (*H–D*) model for Norway spruce (*Picea abies* (L.) Karst.) was estimated from longitudinal data. The Korf growth curve was used as the *H–D* curve. Firstly, *H–D* curves for each stand at each measurement time were fitted, and the trends in the parameters of the *H–D* curve were modeled. Secondly, the trends were included in the *H–D* model to estimate the whole model at once. To take the hierarchy of the data into account, a mixed-model approach was used. This makes it possible to calibrate the model for a new stand at a given point in time using sample tree height(s). The heights may be from different points in time and need not be from the point in time being predicted. The trends in the parameters of the *H–D* curve were not estimated as a function of stand age but as a function of the median diameter of basal area weighted diameter distribution (d^{Gm}). This approach was chosen because the stand ages may differ substantially among stands with similar current growth patterns. This is true especially with shade-tolerant tree species, which can regenerate and survive for several years beneath the dominant canopy layer and start rapid growth later. The growth patterns in stands with a given d^{Gm}, on the other hand, seem not to vary much. This finding indicates that the growth pattern of a stand does not depend on stand age but on mean tree size in the stand.

Résumé : Une équation de prédiction de la hauteur par le diamètre (*H–D*) pour l'épinette de Norvège (*Picea abies* (L.) Karst.) a été calibrée à partir de données longitudinales. La courbe *H–D* est représentée par la courbe de croissance de Korf. Des courbes *H–D* ont d'abord été ajustées pour chaque peuplement et chaque prise de mesures, et les variations dans les paramètres de la courbe *H–D* ont été modélisées. L'équation *H–D* a ensuite été réajustée après y avoir incorporé les modèles de variation de ses paramètres. La procédure d'estimation par modèle mixte a été utilisée pour tenir compte de la hiérarchie des données. Cette procédure permet de calibrer le modèle pour un nouveau peuplement à un moment donné dans le temps en utilisant un échantillon des hauteurs d'arbre. Les hauteurs peuvent provenir de mesures prises à différents moments dans le temps et n'ont pas besoin d'être prises au même moment que la prédiction. La variation des paramètres de l'équation *H–D* n'a pas été modélisée en fonction de l'âge du peuplement, mais plutôt en fonction du diamètre médian de la distribution des diamètres pondérée par la surface terrière (d^{Gm}). En effet, des peuplements d'âge très différent peuvent avoir des patrons de croissance similaires. C'est le cas particulièrement chez les essences tolérantes à l'ombre qui peuvent se régénérer et survivre pendant plusieurs années sous couvert et entamer une croissance rapide par la suite. Par contre, les patrons de croissance dans les peuplements ayant une valeur de d^{Gm} donnée ne semblent pas varier beaucoup. Ce résultat indique que la croissance d'un peuplement ne dépend pas de l'âge du peuplement, mais de la taille moyenne des arbres dans le peuplement.

[Traduit par la Rédaction]

Introduction

In Finnish forest management planning, describing the current forest structure comprises two stages: firstly, the diameter distribution of the stand is estimated, and secondly, the heights of trees are predicted. Because height is quite expensive to measure in practice, one usually has only a few or even no height measurements available from a stand. However, one may have old height measurement(s) from a previous inventory or inventories. A good height–diameter (*H–D*) model is able to predict the *H–D* pattern even if no height measurements are available, and if measurements are available, a good model uses them effectively even if they are from different points in time. Some stand characteristics other than tree height may also improve the prediction of the *H–D* pattern of the stand, and a good model utilizes these variables if they are measured.

Many comparisons of different functions have been carried out to find an appropriate function for the *H–D* relationship in a stand (Curtis 1967; Arabatzis and Burkhart 1992; Huang et al. 1992; Lynch and Murphy 1995; Fang and Bailey 1998). However, no function has been found to be superior. A commonly used function for the *H–D* curve is the Korf function (see, e.g., Parresol 1992; Flewelling and De Jong 1994; Lappi 1997). Lynch and Murphy (1995) selected this model function from among the several functions they tested. Omule and MacDonald (1991) and Flewelling and De Jong (1994) have discussed whether the *H–D* curves from different points in time may cross and have proposed different constraints for such crossings. However, no theoretical justifications for these constraints have been shown, and as Lappi (1997) and Lynch and Murphy (1995) have pointed out, crossover is a natural consequence of different growth rates in different diameter and height classes. The model of Lappi (1997) uses the same model function as that of

Received 5 June 2003. Accepted 5 September 2003. Published on the NRC Research Press Web site at http://cjfr.nrc.ca on 21 January 2004.

L. Mehtätalo. Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, Fin-80101 Joensuu, Finland (e-mail: Lauri.Mehtatalo@metla.fi).

 doi: 10.1139/X03-207

Flewelling and De Jong (1994), but has no constraints for the curves. It has all the features of a good model listed above and will be the starting point of this study. This approach has been used recently by Hökkä (1997) and Jayaram and Lappi (2001); however, in those studies the data were cross sectional.

In previous studies with longitudinal data (e.g., Curtis 1967; Lappi 1997; Eerikäinen 2003; Flewelling and De Jong 1994), the development of an *H–D* curve is usually associated with stand age, i.e., the growth pattern of a stand is assumed to depend on stand age. However, the age at which maturity is reached varies among stands. This is true especially with shade-tolerant tree species, because as a consequence of improved light circumstances, after a long stay as undergrowth in a mature forest those trees may suddenly begin to grow like young saplings. This indicates that the parameter defining the growth rate of an individual tree may not be tree age but tree size. Hence, it would be sensible not to link the height growth pattern to stand age but to mean size of trees in the stand. Another problem with shade-tolerant tree species is that the age distribution of trees in a stand may be very wide, and the definition of stand age may be ambiguous. In this article, I show how an *H–D* curve of a shade-tolerant tree species can be modeled from longitudinal data.

The aim of this study was to develop an *H–D* model for shade-tolerant Norway spruce (*Picea abies* (L.) Karst.) growing in Finland. The model should be able to (*i*) predict the *H–D* pattern of a forest stand when no tree heights are measured, (*ii*) use measured diameter(s) and height(s) to improve the prediction, (*iii*) use old measurements to improve the prediction, (*iv*) predict the future *H–D* curve of a stand, and (*v*) use different combinations of stand characteristics in the prediction of an *H–D* curve. The model could be used in practice for predicting tree heights in forest management planning and for optimizing the number of height measurements in field surveys.

Data

The modeling data are a subset of a larger data set, collected from permanent sample plots by the Finnish Forest Research Institute (Gustafsen et al. 1988). The sample stands in the original data were selected randomly from the sample plots of the seventh National Forest Inventory, which are situated on mineral soils and forestland. The data cover the whole area of Finland. In each sample stand, three fixed-radius sample plots were established, and various characteristics were recorded, including the diameter, height, and species of each tree. These same measurements were taken one to three times at 5-year intervals. In addition, when establishing the plots, the height growth and diameter growth over the last 5 years were measured in some stands. Hence, the number of measurement occasions for each plot of the modeling data is one to four.

In this study, the three sample plots were combined to obtain data for a certain stand. The modeling data were constructed from stands including, on average, 10 or more Norway spruces per measurement occasion. In the actual modeling data, only Norway spruces were included, but basal area and median diameter of basal area weighted diameter distribution (d^{Gm}) were calculated from all trees belonging to the dominant story of the stand. The total number of stands in the data was 249, and each stand was measured, on average, 3.3 times. The total number of height measurements in the data was 18 056, and the number of trees in a stand on a certain measurement occasion varied between 3 and 49 (mean 22). The minimum, mean, and maximum values of the most important variables are shown in Table 1.

Table 1. Some characteristics of the modeling data.

Variable	*n*	Min.	Mean	Max.
T	249	3	65	155
d^{Gm}	249	1.1	19.5	35.1
G	249	0.1	19.6	41.5
D	18 056	0.3	16.1	49.9
H	18 056	1.4	13.5	32.7

Note: T is stand age at breast height (years), d^{Gm} is median of basal area weighted diameter distribution (centimetres), G is basal area (square metres per hectare), D is diameter at breast height (centimetres), and H is height (metres).

Model development

Basic model formulation

Development of the *H–D* model starts from an exponential function, known also as Korf's function (see Zeide 1993; Lappi 1997):

$$[1] \qquad H = a\ \mathrm{e}^{-bD^{-c}}$$

where H and D are tree height and diameter at breast height (1.3 m), respectively, and a, b, and c are parameters to be estimated. Note that setting $c = 1$ in eq. 1 gives eq. 6 of Curtis (1967), which is a very common function (either in its exponential or logarithmic form) in studies of *H–D* models (Zakrzewski and Bella 1988; Huang et al. 1992; Arabatzis and Burkhart 1992; Fang and Bailey 1998). Curtis (1967) and Arabazis and Burkhart (1992) suggest it for routine use, and Curtis (1967) remarks that a value of c other than 1 could provide a better fit in young stands.

Because diameter is measured at breast height, the *H–D* models are often estimated in the form where breast height (1.3 m) is subtracted from the left-hand side of the equation to get diameter and height to behave consistently when D = 0. Subsequently, the equation is linearized with a logarithmic transformation. The transformation $\ln(H - 1.3)$ is, however, highly variable for small trees. A more stable formulation (although not fully compatible with small diameters) is obtained by adding a small constant, λ, to the diameter D (see Lappi 1991, 1997). This constant can be interpreted as the expected difference between the diameter at ground level and that at breast height. The resulting inconsistency is not very serious because it occurs with small trees, which are not very important in the prediction of stand volume. The inconsistency can be corrected, for example, by substituting breast height for all predicted heights below breast height.

Equation 1 was linearized with respect to parameters a and b by logarithmic transformation. The linearized equation with the additional λ parameter is

[2] $\ln(H) = A - B(D + \lambda)^{-C}$

where $A = \ln(a)$, $B = b$, and $C = c$. Attempts to fit eq. 2 into the data showed that it is clearly overparametrized, since parameters B and C are strongly correlated. Furthermore, it cannot be linearized with respect to parameters C and λ. To overcome these problems, fixed values were given to those parameters. To find the values of λ and C, eq. 2 was fitted for each stand and measurement occasion using different combinations of λ and C. The value of λ giving lowest mean error variance over all stands and measurement occasions was selected as the fixed value of λ. For each stand and measurement occasion, the value of C giving the lowest error variance with the fixed λ was selected from among all the tested values. The fixed value of C was calculated as the mean of these measurement-occasion-wise optimal values. Some attempts were also made to estimate a trend curve for C as a function of stand age or d^{Gm}, as Lappi (1997) did, but no clear trend was found. To reduce the correlation, both between the estimates of A and B and between their estimation errors, the diameter was reparametrized as

$$[3] \quad x_{kti} = \frac{(D_{kti} + \lambda)^{-C} - (d_{kt}^{Gm} + 10 + \lambda)^{-C}}{(10 + \lambda)^{-C} - (30 + \lambda)^{-C}}$$

where $\lambda = 7$ cm, $C = 1.564$, and D_{kti} is diameter of tree i in stand k at time t. The model using this parametrization is

$$[4] \quad \ln(H_{kti}) = A_{kt} - B_{kt}x_{kti} + \varepsilon_{kti}$$

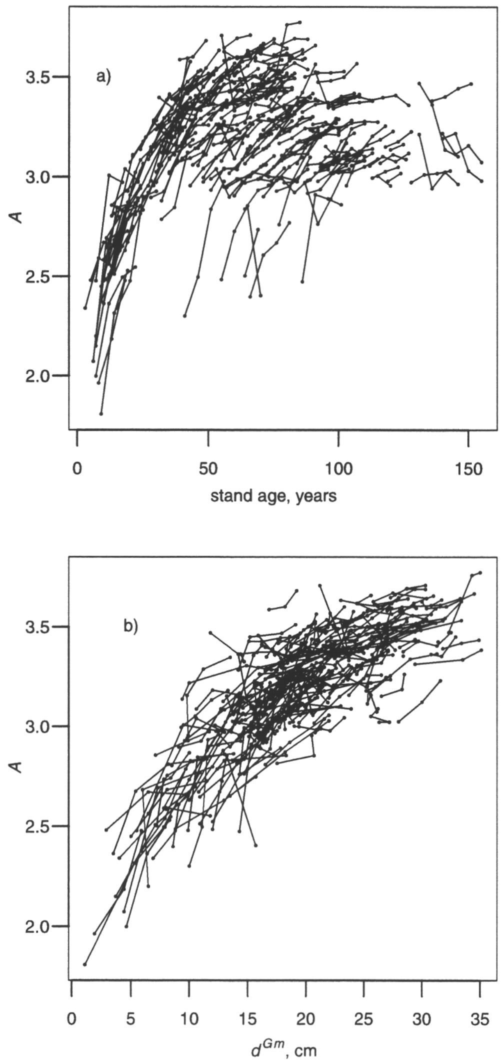
The parametrization (eq. 3) also provides interpretations for parameters A and B. A is the expected logarithmic height of a tree with a diameter of d^{Gm} + 10 cm, which is roughly the height of the trees with the largest D in the stand. (The mean difference between maximum D and d^{Gm} in the modeling data was 6.8 cm.) B is the expected difference in the logarithmic height between trees with diameters of 30 and 10 cm. Lappi (1997) used quite a similar parametrization, except that he had a fixed diameter instead of d_{kt}^{Gm} +10 in the numerator.

Stage of development

In this study, the term stage of development is used to describe the maturity of the stand. The stage of development is a nonquantitative, abstract variable defining at which state the stand is in its development process. The growth pattern of a stand depends strongly on the stage of development. Because the stage of development cannot be measured, it cannot be used as an independent variable in the *H–D* model. Some measurable stand characteristics are, however, strongly connected with the stage of development and can be used in the model as if they were equivalent to the stage of development. Such variables are, for example, stand age, d^{Gm}, and quadratic mean diameter.

Usually the models describing stand development are bound to stand age, i.e., stand age is used to describe the stage of development. However, with the modeling data in this study this does not work well. There are two main reasons for this. First, shade-tolerant tree species may regenerate and survive for dozens of years as undergrowth in a mature forest. Secondly, the stand development process from a sapling stand to a mature one takes longer on poor sites than on rich sites. These phenomena are visible when fitting eq. 2 for each stand and measurement occasion to study the trends of parameter A, which is interpreted as the asymptote of the *H–D* curve (the fixed values were used for λ and C). In Fig. 1*a* one can see old stands for which the asymptote of the *H–D* curve develops very steeply, as it would for a young sapling stand. This implies that the spruces in those stands have either grown very slowly during their first decades or that they arrived in the forest later than the trees used in the determination of the stand age. In addition, the age at which the trend in the asymptote of the *H–D* curve levels out (i.e., the age at which maturity is reached) varies

Fig. 1. Estimated asymptote of logarithmic *H–D* curve (parameter A in eq. 2) versus stand age (*a*) and weighted median diameter (d^{Gm}) of the stand (*b*) in the estimation data. The estimates of any one stand are connected by lines.



markedly among forests. When the asymptote is plotted against stand d^{Gm} (Fig. 1*b*), all stands seem to follow the same trend, except for a stand-wise shift in the level of the curve. Thus, the development of the *H–D* curves was linked to the stand d^{Gm}.

Trend functions for the parameters of the *H–D* curve

The first step in the analysis was to study how the parameters *A* and *B* develop during stand development. This was done by first estimating these parameters (*A* and *B*) separately for each stand and measurement occasion and then studying what kind of trend function would be appropriate to describe the trends of the parameters as a function of d^{Gm}. At this stage, the interest lay in finding the form of the trend functions and estimating their nonlinear parameters. These trends and parameter estimates would later be written into the original *H–D* equation, and all linear parameters would be estimated at once. The modeling was carried out with R, an open source environment for statistical computing (see http://www.r-project.org; Venables and Ripley 2002), where the package nlme (Pinheiro and Bates 2000) was used to fit nonlinear and linear mixed-effects models.

Equation 4 was fitted for each stand at each measurement time with ordinary least squares. (It would later be fitted using weighted least squares (WLS) because the residuals were heteroscedastic with respect to tree diameter; at this stage, however, it was ignored.) The obtained estimates of parameters *A* and *B* were plotted against d^{Gm}. Parameter *A* seems to have a clear sigmoidal trend, while parameter *B* appears to be linear with respect to d^{Gm} (Fig. 2). A four-parameter Champman–Richards equation (Richards 1959; Zeide 1993) was fitted for parameter *A*, and a simple linear model was fitted for parameter *B*:

$$[5] \qquad A_{kt} = p_{1a} + p_{2a}(1 - e^{-p_{3a}d_{kt}^{Gm}})^{p_{4a}} + \alpha_k + \alpha_{kt}$$

$$[6] \qquad B_{kt} = p_{1b} + p_{2b}d_{kt}^{Gm} + \beta_k + \beta_{kt}$$

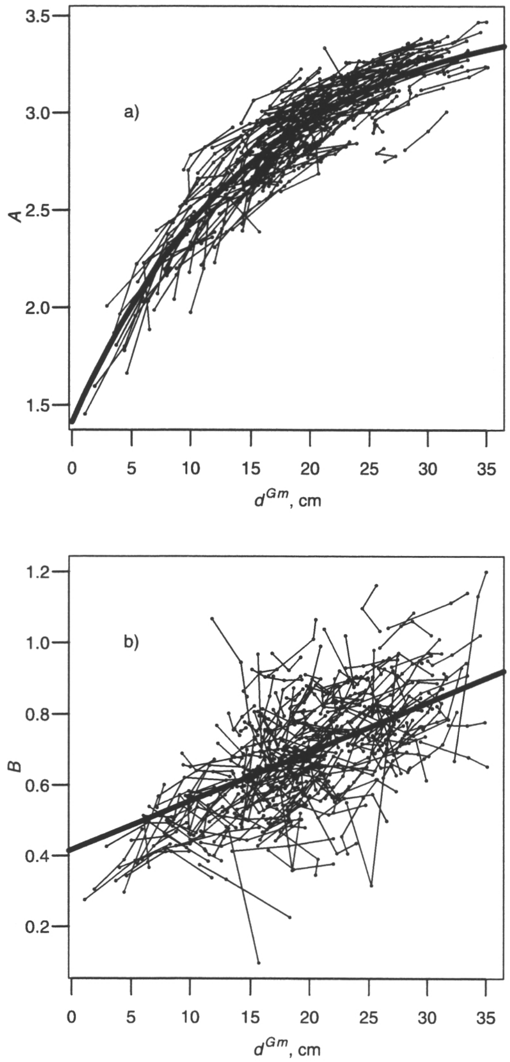
The *p* parameters are fixed. α_k and β_k are the stand-level random parameters, and α_{kt} and β_{kt} are the residuals, all of which have an expectation of 0 and constant variance. This notation is used for residuals because they will be later interpreted as measurement-occasion-level random parameters. It might have been reasonable to assume that parameters other than intercepts were also random, but to avoid convergence problems in the estimation of the final model, no additional random parameters were used.

The complete model

In the previous sections, the model was estimated in two stages. Firstly, eq. 4 was fitted for each measurement occasion to obtain the estimates of parameters *A* and *B*, and secondly, the trend functions (eqs. 5 and 6) of estimated parameters *A* and *B* were fitted. The next step in the analysis is to write the trend equations into eq. 4. The final model is obtained by fitting the resulting model to the original data set. The complete model equation is

$$[7] \qquad \ln(H_{kti}) = (p_{1a} + \alpha_k + \alpha_{kt}) + p_{2a}z_{kt} - (p_{1b} + \beta_k + \beta_{kt})x_{kti} - p_{2b}d_{kt}^{Gm}x_{kti} + \varepsilon_{kti}$$

Fig. 2. The weighted least squares estimates of parameter *A* (*a*) and *B* (*b*) in eq. 4 plotted against weighted median diameter (d^{Gm}) of the stand. Estimates of the same stand at different points in time are connected with lines. The continuous solid lines are the expectations of the trend equations (eqs. 5 and 6).



where x_{kti} is the parametrization (eq. 3) of tree diameter, and z_{kt} is the nonlinear part of eq. 5, i.e., $(1 - e^{-\hat{p}_{3a}d_{kt}^{Gm}})^{\hat{p}_{4a}}$. To retain the linearity of the model, the previously obtained estimates of nonlinear parameters were used as fixed constants, and only linear parameters were reestimated. The parameters are the same as in the trend functions shown previously: p_{1a}, p_{2a}, p_{1b}, and p_{2b} are fixed population-level parameters, α_k and β_k are random stand-level parameters, α_{kt} and β_{kt} are random measurement-occasion-level (i.e., time level) parameters nested within stands, and ε_{kti} is the tree-level residual. The random parameters and residual error are assumed to be normally distributed with an expectation of 0 and constant variance. The covariances $\mathrm{cov}(\alpha_k, \beta_k)$ and $\mathrm{cov}(\alpha_{kt}, \beta_{kt})$ may be nonzero, but $\mathrm{cov}(\alpha_k, \alpha_{k'}) = \mathrm{cov}(\alpha_{kt}, \alpha_{kt'}) = \mathrm{cov}(\beta_k, \beta_{k'}) = \mathrm{cov}(\beta_{kt}, \beta_{kt'}) = \mathrm{cov}(\varepsilon_{kti}, \varepsilon_{kti'}) = 0$ for $k \neq k'$, $t \neq t'$, and $i \neq i'$.

The terms in parentheses in eq. 7 are the intercepts of trend functions. The model assumes that the expected development of the parameters follows the fixed part of the models in eqs. 5 and 6. However, because of the stand-level random parameters α_k and β_k, in any one stand the level of the trend functions may differ from the expected trend. In addition, the measurement-occasion-level random parameters imply that at a certain point in time the parameter values may further deviate from the expected trend function of the stand.

When the trend functions were estimated, the heterogeneity of residual variation was not taken into account. However, examination of the standardized residuals against tree diameter showed that the variance of β_{kti} is about constant only with trees with a diameter of less than 7.5 cm; for larger diameters a clear decreasing trend was seen. Similar trends in variances have also been found in previous studies (Lappi 1997; Jayaraman and Lappi 2001), and Lappi (1997) gives a good explanation for this: with large trees, physical stability requirements do not leave as much room for height variation as with smaller trees. The following power function was used to describe the variance of the error term:

$$[8] \qquad \mathrm{var}(\varepsilon_{kti}) = \sigma^2\{[\max(D_{kti}, 7.5)]^{-\delta}\}^2$$

where σ and δ are fixed parameters. With the software used it is possible to estimate the parameters of the variance function simultaneously with the other model parameters.

A primary estimate for the power parameter δ was obtained by fitting eq. 7 to the data. The estimate was $\delta = 0.53$, which homogenized the variance well. This estimate was used to predict the residual variance for each tree. The parameters (A and B) of the H–D curves of each stand and measurement occasion were then reestimated with WLS, using inverted predicted variances as WLS weights. Furthermore, the parameters of trend functions in eqs. 5 and 6 were reestimated using these new estimates of parameters A and B. The obtained nonlinear part of the trend function in eq. 5 was $z_{kt} = (1 - \mathrm{e}^{-0.0651 d_{kt}^{\mathrm{Gm}}})^{0.999}$. Finally, eq. 7 was fitted again using the obtained value of z_{kt}. At this stage, new estimates for the parameters of the variance function were also obtained.

Fitted models

One requirement for the model was that it should be able to utilize the measured stand variables when predicting the H–D curve. To avoid a situation in which some of the measured stand characteristics cannot be used in predicting the stand H–D curve, models using different levels of information were estimated. The additional predictors were assumed to affect the H–D curve by shifting the level of the trend functions, i.e., it was assumed that $p_{1a} = b_{0a} + b_{1a}x_{1a} + b_{2a}x_{2a} + \ldots$, and $p_{1b} = b_{0b} + b_{1b}x_{1b} + b_{2b}x_{2b} + \ldots$, where x_{1a}, x_{1b}, x_{2a}, x_{2b},... are the additional predictors, and b_{0a}, b_{0b}, b_{1a}, b_{1b},... are fixed parameters to be estimated. These equations were written into eq. 7, and the whole model was estimated using restricted maximum likelihood. Only significant predictors at the 1% level of significance were included in the final models; the predictors were found by backward elimination. The additional predictors tested were stand coordinates, altitude above sea level, cumulative temperature sum, soil type, dummy variable for thinned stands, basal area, age, basal area of spruces, and d^{Gm} of spruces.

In Table 2 the estimated models are presented from the crudest model (I) to the broadest model (V). Model I is the basic model with only d^{Gm} as a predictor. In model II, site fertility class and geographical variables are included. These variables take into account the large geographic variation in the data: the maximum distance between stands in the modeling data was 845 km, the altitude above sea level varied from 10 to 360 m, and the cumulative temperature sum varied from 725 to 1342. The geographical variables are always known in practice (they can be derived from the stand location, for example, using generally available maps), and thus they are also used in all other models. A commonly assessed stand characteristic is also the site fertility class, which is in the model as a dummy variable for mesic and poorer sites. These variables do not cause any harm if the H–D curve of a stand is predicted for several points in time, since their values do not change over time and thus need not be measured at each point in time.

In addition to the predictors of model II, models III–IV include different combinations of stand variables. In models III and IV, stand characteristics are assumed to be measured only as mean values of the dominant story, not by tree species. The difference between models III and IV is that stand age is used in model IV but not in model III. In model V the stand characteristics (G and d^{Gm}) are also assumed to be measured by tree species, but as seen in the models, these variables provide only a small amount of additional information: basal area of spruces was not significant at all, and d^{Gm} of spruces was a significant predictor only for the A parameter.

The estimated parameters of the variance function vary only slightly, and the residual variation in each model is about the same. The differences among models appear in how the explained variation is shared between the fixed and random parts of the models. Model 0 was estimated to demonstrate and compare this division. The only fixed parameters in model 0 are parameters p_{1a} and p_{1b}, i.e., the trend functions of A and B are assumed to be constant, and all variation of the variables A and B is described by the random part of the model. In the other models, this variation is shared between fixed predictors and random parameters. The proportion explained by fixed parameters of a given model can be calculated as $R^2 = 1 -$ (variance of random parameter in the given model/ variance of random parameter in model 0). For example, for stand-level random parameter α_k in model I, it is calculated as $R^2_{\alpha_k,\mathrm{I}} = 1 - \mathrm{var}(\alpha_{k,\mathrm{I}})/\mathrm{var}(\alpha_{k,0})$. This statistic has been calculated for both stand-level and measurement-occasion-level random parameters in Table 2. One should note, however, that the value of R^2 does not behave consistently in all cases, because the correlation between random parameters is high, and the estimated share of the variation between stand and measurement occasion varies slightly among models.

In model I, the estimated trend functions explain 85% of the stand-level variation and 95% of the measurement-occasion-level variation of the maximum height in the stand (parameter A). With parameter B the proportion is clearly

Table 2. Estimated parameters of eq. 7 with different predictor combinations.

	Population level (fixed part)				Stand level	
Model	p_{1a}	p_{2a}	p_{1b}	p_{2b}	SD (α_k)	SD (β_k)
0	2.91		0.692		0.278	0.136
I	1.40	2.15	0.384	0.0157	0.108	0.0958
II	9.45 – 0.941yk – 0.119alt – 0.121dd + 0.0324soil	2.12	0.656 – 0.0281dd	0.0171	0.0891	0.0867
III	8.39 – 0.81yk – 0.101alt – 0.0999dd + 0.46G + 0.0213thin + 0.0367soil	1.82	0.646 – 0.0262dd – 0.305G	0.0196	0.0869	0.0842
IV	11.79 – 1.24yk – 0.127alt – 0.138dd + 0.369G – 0.706d^{Gm} + 0.0210thin + 0.228T	1.92	2.33 – 0.225yk – 0.0379dd – 0.338G + 0.00128t	0.0167	0.0813	0.0800
V	11.81 – 1.24yk – 0.128alt – 0.139dd + 0.36G – 1.03d^{Gm} + 0.224T + 0.332$d^{\mathrm{Gm,s}}$ + 0.0201thin	1.93	2.33 – 0.226yk – 0.0381dd – 0.339G +0.128t	0.0167	0.0812	0.0800

Note: Only significant parameters at the 1% level are included in the models. The predictors are as follows: yk, the north coordinate in the Finnish G, total basal area of the stand (100 m^2/ha); d^{Gm}, median of basal area weighted diameter distribution of the stand (100 cm); $d^{\mathrm{Gm,s}}$, d^{Gm} of spuces (100 cm); which takes a value of 1 with stands on mesic or poorer sites. σ and δ are coefficients of the variance function (eq. 8).

lower, which is due to the large within-stand variation of parameter B when compared with that of parameter A (Fig. 2). One can see two clear improvements in the calculated R^2 values among the models. The first one occurs between models I and II, i.e., when geographical variables and site fertility class are included in the model. In this case, only the stand-level R^2 increases markedly because the additional fixed predictors in model II are constant in a certain stand, and thus they do not explain the variation at the measurement-occasion level. The second improvement becomes evident when the most commonly measured stand characteristics are included, i.e., between models II and III. In this case, the improvement occurs also at measurement-occasion-level R^2 because values of the additional predictors vary among measurement occasions. Stand age also improves the fit slightly, and it should be used when it has been measured. On the other hand, measurement of stand characteristics by tree species is clearly not useful.

Application of the model

Calibrating the *H–D* curve with measured heights

When the model is applied in predicting the *H–D* curve of a stand, both the fixed and random parts can be utilized. The prediction of the fixed part is obtained simply by placing the measured values of the predictors in the appropriate model of models I–V. If no tree heights have been measured, the expected value 0 is used for all random parameters, and the prediction of the fixed part is the predicted *H–D* curve.

If tree heights have been measured, they can be used to predict the random parameters of the stand and measurement-occasion levels. The random parameters of a linear mixed model are predicted using the best linear unbiased predictor (see Searle 1971; McCulloch and Searle 2001), which has been applied, for example, by Lappi (1991, 1997) in the context of linear mixed-effects models in forestry. Assume that we have n measured trees, possibly from different points in time. The measured log-transformed heights are in vector $\boldsymbol{y}_{n\times 1}$ and their expectation in $\boldsymbol{\mu}_{n\times 1}$. The measured heights are described by the model

$$[9] \qquad \boldsymbol{y} = \boldsymbol{\mu} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{e}$$

where $\mathbf{Z}_{n\times m}$ is the model matrix of the random part of the model, $\boldsymbol{b}_{m\times 1}$ is the vector of the random parameters, and $\boldsymbol{e}_{n\times 1}$ is the vector of prediction errors. In addition, we define the variance–covariance matrices of the random parameters and residuals as var($\boldsymbol{b}$) = $\mathbf{D}_{m\times m}$ and var($\boldsymbol{e}$) = $\mathbf{R}_{n\times n}$, respectively. The number of random parameters to be predicted (m) depends on how many measurement occasions we have in the stand. For example, if we are predicting the random parameters of eq. 7 in stand k at measurement times t_1 and t_2, m = 6 and $\boldsymbol{b}' = (\alpha_k, \beta_k, \alpha_{kt_1}, \beta_{kt_1}, \alpha_{kt_2}, \beta_{kt_2})$. A numerical example of predicting the random parameters is presented in Lappi (1991).

When random parameters are predicted, the estimates of matrices **D** and **R** are needed. The estimate of **D** (denoted by $\hat{\mathbf{D}}$) can be built up from the estimated standard deviations and correlations of the random parameters (Table 2). The estimated error variances of the measured trees can be calculated using the variance model in eq. 8, with the estimated parameters δ and σ from Table 2. These variances are placed on the diagonal of matrix $\hat{\mathbf{R}}$. Because prediction errors of different trees are assumed to be uncorrelated, the nondiagonal elements of $\hat{\mathbf{R}}$ are zeros. The best linear unbiased predictor of the random parameters is calculated as

$$[10] \qquad \hat{\boldsymbol{b}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

When the random parameters are calculated, the estimated variance–covariance matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{D}}$ are substituted for matrices **R** and **D**, and the prediction of the fixed part is substituted for the vector $\boldsymbol{\mu}$. For further details see, for example, Lappi (1997) and Searle (1971, pp. 458–462).

To obtain calibrated *H–D* curves, the predictions of the random parameters from the vector $\hat{\boldsymbol{b}}$ are placed in eq. 7. Note that the predicted stand-level random parameters can also be used to predict the *H–D* curve at a development stage for which we do not have measurements. In this case, we just place the predicted stand-level random parameters in eq. 7 and use the expectation 0 for the measurement-occasion-level parameters.

			Measurement-occasion level					Variance function	
corr(α_k, β_k)	R_A^2	R_B^2	SD (α_{kt})	SD (β_{kt})	corr(α_{kt}, β_{kt})	R_A^2	R_B^2	σ	δ
0.718			0.0793	0.0247	0.346			0.408	–0.539
0.269	0.848	0.501	0.0168	0.0223	–0.681	0.955	0.188	0.401	–0.534
0.592	0.897	0.591	0.0170	0.0225	–0.642	0.954	0.171	0.402	–0.535
0.698	0.902	0.615	0.0139	0.0211	–0.484	0.969	0.271	0.402	–0.535
0.657	0.914	0.652	0.0117	0.0222	–0.49	0.978	0.196	0.404	–0.537
0.653	0.915	0.652	0.0114	0.0222	–0.517	0.979	0.192	0.404	–0.537

uniform coordinate system, 1000 km; alt, altitude above sea level (100 m); dd, cumulative temperature sum (average of the years 1951–1980) (100 × dd); T, stand age at breast height (100 years); thin, dummy variable indicating if the stand has been thinned within the last 10 years; soil, dummy variable

Utilizing the longitudinal character of the model

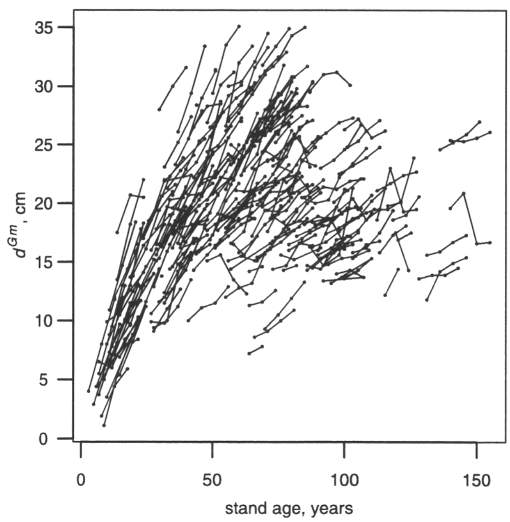
When the longitudinal character of the model is applied in practice, the fixed predictors of the model need to be known at each point in time under consideration, i.e., both at the point(s) used in predicting random effects and at the point at which the *H–D* curve is predicted. The longitudinal character of the model can be utilized in two different ways: (*i*) using height measurements from several points in time to predict the *H–D* curve at a certain point in time and (*ii*) making predictions into the future. In case *i*, the use of the model is straightforward: just place the measured stand variables into the appropriate model and, if heights have been measured, predict the random parameters using eq. 10. In case *ii*, the situation is a bit more complicated, because usually one wants to predict the *H–D* curve for some particular point in time, not for the time when d^{Gm} reaches a certain value. Therefore, we need to predict d^{Gm} first and use it in the prediction of the *H–D* curve.

Figure 3 shows that the variation in d^{Gm} in stands at any particular age is considerable. Among-stand variation is, however, greater than within-stand variation. In addition, in any one stand, the change in d^{Gm} is stable with respect to stand age. This is a typical situation, in which a longitudinal analysis can be applied (Diggle et al. 1994), and it was used to model the development of d^{Gm} as a function of stand age. In any one stand the growth rate seems to be almost linear except for a slightly concave trend. Hence, a simple power function seems to fit the data well. The power function was linearized with a logarithmic transformation to obtain the model

$$[11] \qquad \ln(d_{kt}^{\mathrm{Gm}}) = u + v\ln(T_{kt}) + u_k + e_{kt}$$

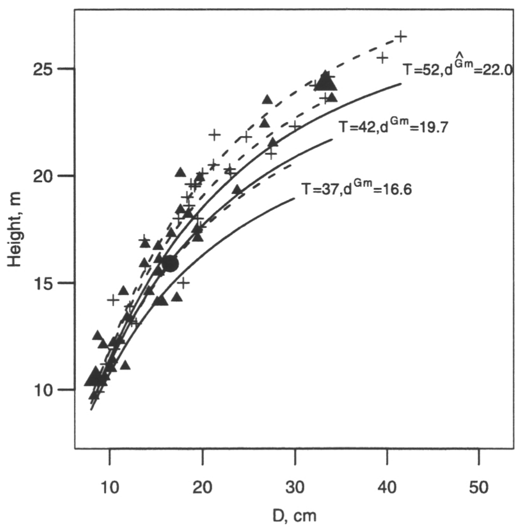
where u and v are fixed population-level parameters, u_k is a stand-level random parameter, and T_{kt} is the breast-height age of stand k at time t. The parameters u_k and e_{kt} are assumed to be independent and normally distributed with expectations of 0 and variances σ_u^2 and σ_e^2, respectively. The transformation linearized the data rather well, with the exception of a few very variable young stands. The likelihood ratio test showed no significant difference between the model in eq. 11 and a model without intercept u; thus, the intercept was dropped from the model. The parameter estimates were v = 0.727, $\sigma_u^2 = 0.372^2$, and $\sigma_e^2 = 0.0751^2$. The estimated standard deviations show that the major proportion of the random part is due to the among-stand variance component as opposed to the within-stand component.

Fig. 3. Stand weighted median diameter (d^{Gm}) versus stand age in the modeling data. The lines connect measurements of any one stand.



In practice, stand age and d^{Gm} are always known at least for one point in time, and they can be used to predict the localized stand-level d^{Gm}–age curve. We just need to predict the stand-level random parameter u_k with eq. 10 using the observed age(s) and d^{Gm}(s) of the stand. The localized curve is then obtained by writing the prediction of the random effect u_k into eq. 11, and the obtained curve is used to predict the d^{Gm} at a certain future point in time. Furthermore, the predicted d^{Gm} is used in predicting the *H–D* curve at the time point under consideration.

Fig. 4. An example of predicted *H–D* curves using model II. The solid lines are the predicted population curves at different ages, with no measured heights used. The broken lines are the calibrated curves obtained using three height measurements: height of weighted median tree at age 37 (●) and two height measurements at age 42 (▲). The small symbols show the observations for the stand at age 42 (▴) and at age 52 (+) (to make the plot more legible, the observations at age 37 are not shown). In the prediction of the 52-year curve, the predicted d^{Gm} was used; the true value at this age was 27.4 cm.



Application example

As an application example, take a look at a 42-year-old stand selected from the modeling data. It was assumed that stand d^{Gm} and the height of two sample trees (one small and one large) had been measured at the age of 42 years. Furthermore, the diameter and height of the tree with the median basal area had been measured at the age of 37 years. Using these measurements, the predictions of the stand *H–D* curve at ages 37, 42, and 52 years were made.

To exemplify making predictions about future development, stand d^{Gm} at age 52 was predicted using the model in eq. 11. The random effect u_k was predicted with eq. 10 using the known d^{Gm}'s at ages 37 and 42. Equation 10 was used also to predict the random effects of the *H–D* curve (i.e., the stand effects and measurement-occasion-level effects for ages 37 and 42 years) with the three measured sample trees. The calibrated curves were calculated by placing the predicted random effects in eq. 7 and using an expectation of 0 for the effects that are not predicted.

The curves obtained both by using and not using the measured heights are shown in Fig. 4. The calibrated curves match the observed heights very well, while the uncalibrated population-level curves, at least for this stand, clearly underestimate height. In this case, the predicted d^{Gm} at age 52 (22.0 cm) was clearly below the true value (27.4 cm), but the produced prediction of the *H–D* curve was good, whereas using the true value would have produced clear overestimates for heights.

Discussion

In this study, d^{Gm} was used to describe the stage of development of a stand instead of using the traditional stand age. This definition produced quite nice trend functions for the parameters of the *H–D* curve as a function of d^{Gm}, while the use of stand age would have led to considerable problems with the trend functions. This is because the development of the *H–D* curve depends strongly on the development of d^{Gm} (Fig. 1), but stand age does not explain d^{Gm} well (Fig. 3). When the development of the *H–D* curve was linked to d^{Gm}, the scatterplots in Figs. 1*a* and 3 did not harm the estimation of the model. The use of d^{Gm} instead of stand age makes the model efficient in the cases where d^{Gm} is known both at the current point in time and at the points at which the height measurements are made. This is because in this case, the stand-level trends in the parameters are precisely predicted and the current stage of development is known. It should be noted that d^{Gm} is measured even in the most parsimonious inventories in Finland, and it is even more often known than stand age. The use of d^{Gm} instead of stand age also eliminates the problem of how to define stand age in heterogeneous forests.

With Norway spruce, which is a rather shade-tolerant tree species, the difference between stand age and stage of development is evident (Fig. 1*a*). One reason for the difference is that shade-tolerant trees may stop their growth in shelter but later, when the amount of light increases, continue their growth as if they were young trees. Another reason is that site properties affect the age at which maturity is reached, i.e., trees on poorer sites grow more slowly but for a longer time than trees on rich sites. The second argument is true also with shade-intolerant tree species (see, e.g., Fig. 3*b* in Lappi 1997). Thus, it could be useful to also link the *H–D* curves of those tree species to mean tree size. One should note that the variable used as mean size need not necessarily be d^{Gm}; one can use any variable describing the mean tree size in the stand.

A problem caused by the definition used for the stage of development is that we cannot straightforwardly predict the *H–D* curve at some particular point of time in the future. Instead, the predictions are made at the point in time at which d^{Gm} has reached a particular value. If one wants to make the predictions for a particular point in time, the development of d^{Gm} needs to be predicted first. The known stand age(s) and d^{Gm}(s), however, make it possible to estimate a stand-level d^{Gm}–age curve by calibrating eq. 11, which can be used to predict d^{Gm} at any point in time.

When the stage of development is defined using d^{Gm}, it may in some cases move "backward". Take, for example, a stand where there are small spruces in a mature forest. Because the spruces are in a stand at a "late" stage of development, the predicted *H–D* curve develops like that of a mature forest. When the sheltering trees disappear, e.g., because of a harvest, d^{Gm} decreases, and the stand returns to an "earlier" stage of development. Therefore, the predicted *H–*

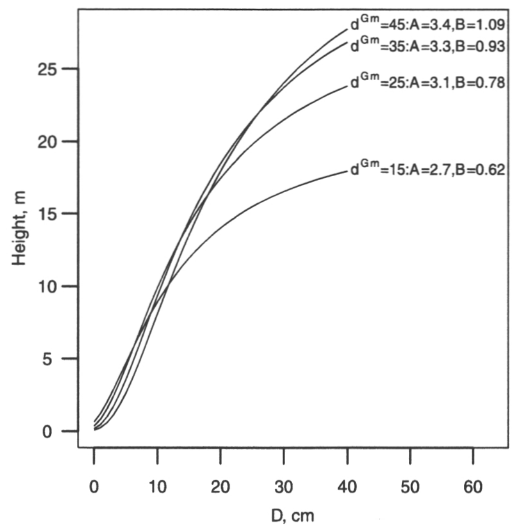
D curve begins to develop as for a young forest, which seems to be realistic according to Fig. 1*a*.

The approach used in this study was a modified version of the analysis of Lappi (1997). Using this approach it was possible to estimate a model that fulfills the requirements set at the beginning of this article, and it seems to be an appealing approach for modeling the *H–D* curve from longitudinal data. The estimation of model parameters with modern statistical software is easy, and imbalanced data and missing observations are no longer harmful. An important property of the random-parameter approach is that the measured heights can be used to localize (calibrate) the curve for a new stand and at a given point in time. One should note that localizing can be done even using just one measured tree of any diameter from any point in time, and no representative sample is needed, as is the case, for example, in the model of Lynch and Murphy (1995). Furthermore, in calibration, the within-stand and among-stand variation observed in the modeling data are used in a statistically correct and well-grounded manner.

The same precision in the prediction can be achieved by utilizing either the fixed part or the random part of the model. This means that one can either measure all possible stand characteristics and use model V, or, more likely, measure some sample trees and d^{Gm} and use the more parsimonious model I or II. In the second approach, the prediction error decreases, while the number of trees increases, and with a sufficient number of height sample trees, both approaches result in equally accurate (but not the same) predictions. However, when applied outside the estimation data, the first approach is more likely to produce biased predictions than the second one. This is why Lappi (1997) recommends the second approach if the prediction is equally accurate with both approaches. The second approach appears more straightforward, because it does not utilize somewhat fuzzy correlations between stand characteristics and the *H–D* curve, but improves the prediction of the *H–D* curve with measured heights and diameters. The second approach is useful in optimizing collection of forest data under some accuracy constraints, because the precision of the prediction can be improved by increasing the number of measurements. Naturally, the approaches can also be combined, i.e., models III–V can be calibrated with measured heights, if all the fixed predictors are known at each point in time.

The data were grouped according to the stand and the measurement occasion, where the measurement-occasion level was nested within stands. Lappi (1997) used an additional tree-level grouping to take into account the fact that the same trees are measured several times and the prediction errors of the same tree at different time points are correlated. The tree-level random parameter was nested within stands but crossed with the random parameter of the measurement-occasion level. The random parameter of the tree level was also tested in this study, but estimating the model became so slow that it was abandoned. This is not likely to have a considerable negative effect on the fit, because in the modeling data, the number of measurement occasions of one tree was low (on average 3.05) when compared with the number of trees (5919). In applications, the tree-level random parameter would be used only if the same tree were measured at different points in time. This is, however, very exceptional in practical inventories, and even if it happens, no one knows that it was the same tree because the sample trees are not marked.

Fig. 5. Expected *H–D* curves at different stages of development obtained with model I.



The trends of the parameters are associated with the way in which stands develop. Parameter *A* was interpreted as the height of the largest trees in the stand, and the trend of *A* is hence closely related to the height growth of the (largest) trees in the stand. Growth in parameter *B*, on the other hand, moves the curve to the right in the coordinate, which means that it is associated with diameter growth. This phenomenon can be seen in Fig. 5, where predictions of the expected *H–D* curves at different stages of development are plotted: in early stages of development, the expected curve is both moving to the right, and the asymptote is rising, while at late stages of development, the asymptote is rising only slightly, but the curve is still moving to the right as rapidly as before. The convex trend function of parameter *A* thus implies that height growth is most rapid in early stages of development and decreases as the stand gets older, while linear growth in *B* indicates that diameter growth continues steadily up to a fairly old age. These trends agree with our knowledge on tree growth pattern and indicate that the trends of *A* and *B* are logical.

Acknowledgements

This work was part of the project "Statistical modeling for forest management planning", which was carried out at the Finnish Forest Research Institute and funded by the Academy of Finland (decision No. 73392). The author thanks Dr. Juha Lappi for the many discussions we had, all of which helped to clarify the issues related to this study, and Professors Matti Maltamo and Annika Kangas for their comments, which enabled me to improve the manuscript. In addition, I thank the two anonymous referees and the Associate Editor of the Journal for the suggestions that helped me to improve the manuscript. Finally, I thank Dr. Lisa Lena Opas-Hänninen for revising my English.

References

Arabatzis, A.A., and Burkhart, H.E. 1992. An evaluation of sampling methods and model forms for estimating height–diameter relationships in loblolly pine plantations. For. Sci. **38**: 192–198.

Curtis, R.O. 1967. Height–diameter and height–diameter–age equations for second-growth Douglas-Fir. For. Sci. **13**: 365–375.

Diggle, P.J., Liang, K.-Y., and Zeger, S.L. 1994. Analysis of longitudinal data. Oxford Statistical Science Series. Oxford University Press, Oxford, U.K.

Eerikäinen, K. 2003. Predicting the height–diameter pattern of planted *Pinus kesiya* stands in Zambia and Zimbabwe. For. Ecol. Manage. **175**: 355–366.

Fang, Z., and Bailey, R.L. 1998. Height–diameter models for tropical forests on Hainan Island in southern China. For. Ecol. Manage. **110**: 315–327.

Flewelling, J., and de Jong, R. 1994. Considerations in simultaneous curve fitting for repeated height–diameter measurements. Can. J. For. Res. **24**: 1408–1414.

Huang, S., Titus, S.J., and Wiens, D.P. 1992. Comparison of nonlinear height–diameter functions for major Alberta tree species. Can. J. For. Res. **22**: 1297–1304.

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

Hökkä, H. 1997. Height–diameter curves with random intercepts and slopes for trees growing on drained peatlands. For. Ecol. Manage. **97**: 63–72.

Jayaraman, K., and Lappi, J. 2001. Estimation of height–diameter curves through multilevel models with special reference to even-aged teak stands. For. Ecol. Manage. **142**: 155–162.

Lappi, J. 1991. Calibration of height and volume equations with random parameters. For. Sci. **37**: 781–801.

Lappi, J. 1997. A longitudinal analysis of height/diameter curves. For. Sci. **43**: 555–570.

Lynch, T.B., and Murphy, P.A. 1995. A compatible height prediction and projection system for individual trees in natural, even-aged shortleaf pine stands. For. Sci. **41**: 194–209.

McCulloch, C., and Searle, S.R. 2001. Generalized, linear and mixed models. John Wiley & Sons, New York.

Omule, S.A.Y., and MacDonald, R.N. 1991. Simultaneous curve-fitting for repeated height–diameter measurements. Can. J. For. Res. **21**: 1418–1422.

Parresol, B.R. 1992. Baldcypress height–diameter equations and their prediction confidence intervals. Can. J. For. Res. **22**: 1429–1434.

Pinheiro, J.C., and Bates, D.M. 2000. Mixed-effects models in S and S-PLUS. Springer-Verlag, New York.

Richards, F.J. 1959. A flexible growth function for empirical use. J. Exp. Bot. **10**: 290–300.

Searle, S.R. 1971. Linear models. John Wiley & Sons, New York.

Venables, W.N., and Ripley, B.D. 2002. Modern applied statistics with S. 4th ed. Springer-Verlag, New York.

Zakrzewski, W.T., and Bella, I.E. 1988. Two new height models for volume estimation of lodgepole pine stands. Can. J. For. Res. **18**: 195–201.

Zeide, B. 1993. Analysis of growth equations. For. Sci. **39**: 594–616.

IV

HEIGHT-DIAMETER MODELS FOR SCOTS PINE AND BIRCH IN FINLAND

Lauri Mehtätalo

Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, Fin-80101 Joensuu, Finland.

Abstract. Height-Diameter (*H-D*) models for two shade-intolerant tree species were estimated from longitudinal data. The longitudinal character of the data was taken into account by estimating the models as random effects models using two nested levels: stand and measurement occasion level. The results show that the parameters of the *H-D* equation develop over time but the development rate varies between stands. Therefore the development of the parameters is not linked to the stand age but to the median diameter of the basal-area weighted diameter distribution (*DGM*). Models were estimated with different predictor combinations in order to produce appropriate models for different situations. The estimated models can be localized for a new stand using measured heights and diameters, presumably from different points in time, and the *H-D* curves can be projected into the future.

Key words: longitudinal analysis, random parameter, mixed model, stand development

Introduction

Many studies have presented models for the prediction of the height-diameter (*H-D*) relationship of a stand. Most of these models use a representative sample of height sample trees from the target stand (Curtis 1967, Arabatzis & Burkhart 1992, Huang et al. 1992, Lynch & Murphy 1995, Fang & Bailey 1998) and the main focus in these studies lies in finding the best functional form of the model. However, in many situations, the sample size needed is too large for practical purposes. This is because height measurements are time-consuming and in many situations, for example in a stand wise inventory for forest management planning, only one or a few sample trees from a stand can be measured. Therefore, in recent studies, models that can predict the *H-D* relationship of a stand using few or even no sample trees have been developed (Lappi 1997, Eerikäinen 2003, Mehtätalo 2004). In these models, the accuracy of the prediction can be improved by enlarging the number of height sample trees.

The *H-D* relationship of a stand is not stable but develops over time (Curtis 1967, Flewelling & de Jong 1994, Lappi 1997). However, Mehtätalo (2004) observed that with a shade-tolerant tree species (Norway spruce, *Picea abies*) the age at which maturity is reached varies from stand to stand and gave two reasons for this. First, after a long stay as undergrowth, shade-tolerant trees may suddenly begin to grow as rapidly as young saplings and secondly, the development of a stand from a sapling stage to a mature stand takes longer on poor sites than on rich sites, i.e. the development rate varies between stands. On the other hand, stands that are reaching maturity seem to have almost equal mean tree size. Thus, instead of linking the development of *H-D* relationship with stand age, Mehtätalo (2004) linked it with basal area weighted median diameter of the stand (*DGM*). The latter of the reasons given above for the variation in the age at which maturity is reached might hold also with shade-intolerant tree species. Therefore, this study presents an analysis similar to that in Mehtätalo (2004) but with two shade-intolerant tree species.

The aim of this study is to model the *H-D* relationship of Scots pine *(Pinus sylvestris)* and birch *(Betula pendula* and *Betula pubescens)* in Finland. The methodology is the same as was used in Mehtätalo (2004), but here it is applied to shade-intolerant tree species. Methodologically, the aim is to study if the development of the *H-D* curve of shade-intolerant tree species should also be linked with mean tree size rather than with stand age. Furthermore, the aim of this study is to show that the model formulation used in Lappi (1997) and Mehtätalo (2004) can be successfully applied with several tree species with only minor changes in the model formulation.

Data

The modeling data are a subset of a larger dataset collected by the Finnish Forest Research Institute (Gustafsen et al. 1988). The sample stands of the data were selected randomly from those sample plots of the 7th National Forest Inventory which are situated on mineral soils and forest land. Three fixed-radius sample plots were established in all stands. Each sample plot was measured 3 times with 5 year intervals. In addition, when establishing the plots, the growth of the trees over the previous 5 years was recorded in some stands. Thus, the number of measurement occasions for each stand varied from 1 to 4.

In this study, the three plots were combined to obtain the data of a stand. Only stands with on average more than 10 Scots pines / measurement occasion were selected to the Scots pine data and stands with more than 8 birches / measurement occasion to the birch data. All trees of other tree species than the one in question were removed from the data of any given tree species. However, before doing this, *DGM* and basal area were calculated from all trees belonging to the dominant story of the stand. The Scots pine data included 46338 observations from 497 stands (1774 measurement occasions) and the birch data 2979 observations from 61 stands (190 measurement occa-

Table 1. Some characteristics of the modeling data.

	Scots pine			birch		
	n=1774			*n*=190		
	min	mean	max	min	mean	max
stand age at breast height, years	1	52	166	6	58	126
y-coordinate, km	6652	7135	7568	6658	7111	7520
x-coordinate, km	204	479	716	214	478	654
altitude, m	5	151	320	2	146	300
temperature sum, dd	658	982	1339	696	998	1350
DGM, cm	2	15.6	38.6	2.5	16.5	32.1
basal area, m^2/ha	0.1	13.7	40.8	0.1	15.1	35.3
DGM of tree species, cm	2	15.8	38.5	2.4	13.4	31.8
basal area of tree species, m^2/ha	0.1	12.6	37.8	0.1	7.1	30.3

sions). Table 1 summarizes the modeling datasets of this study.

Model development

The model is the linearized form of the exponential function (Korf function), which in many studies has been found good for the description of the *H-D* relationship of a stand (Zakrzewski & Bella 1988, Huang et al. 1992, Arabatzis & Burkhart 1992, Fang & Bailey 1998). The model formula is

$$\ln(H) = A - B(D+\lambda)^{-C}, \qquad (1)$$

where H is tree height and D diameter at breast height and A, B and C and λ are parameters. The parameter λ is interpreted as the expected difference between diameter at ground level and that at breast height. It is an alternative for the more commonly used subtraction of the breast height from the height; for more discussion on the matter see Lappi (1991, 1997).

Trying to fit model (1) to the data showed that the model is clearly overparametrized. Thus, the first step in the analysis was to reduce the number of parameters to be estimated. To do this, model (1) was fitted separately for each stand and measurement occasion using different combinations of parameters C and λ. The value of λ giving the lowest mean error variance over all stands and measurement occasions was selected as the fixed value of λ. The selected values were 7 for Scots pine and 6 for birch. For each stand and measurement occasion, the value of C giving the lowest error variation was selected and these values were modeled as a function of *DGM* to fit a heuristic trend function for parameter C (see Lappi 1997). The trend function of C for Scots pine was C=0.9823+0.05753**DGM* but for birch no trend was found and thus the constant C=1.809 was used.

The model for tree i in stand k at time t is

$$\ln(H_{kti}) = A_{kt} - B_{kt}(D_{kti}+\lambda)^{-C_{kt}} + \varepsilon_{kti}, \qquad (2)$$

where parameters A_{kt} and B_{kt} need to be estimated. The next step in the analysis was to study whether *DGM* is a better descriptor of the stage of development than the stand age. If it is, then in the subsequent longitudinal analysis (Diggle et al. 2002) rather than using stand age, we use *DGM* as the variable describing the development of the stand.

Model (2) was fitted separately for each stand and measurement occasion. The obtained parameter estimates were plotted against stand age and *DGM* (Fig. 1). Only the estimates of parameter A are shown here; it is interpreted as the asymptote of the *H-D* curve of a stand, i.e. it is the maximum tree height of the stand. In any one stand, parameter A seems first to develop rather rapidly and later level out at some level, which can be interpreted as the maxi-

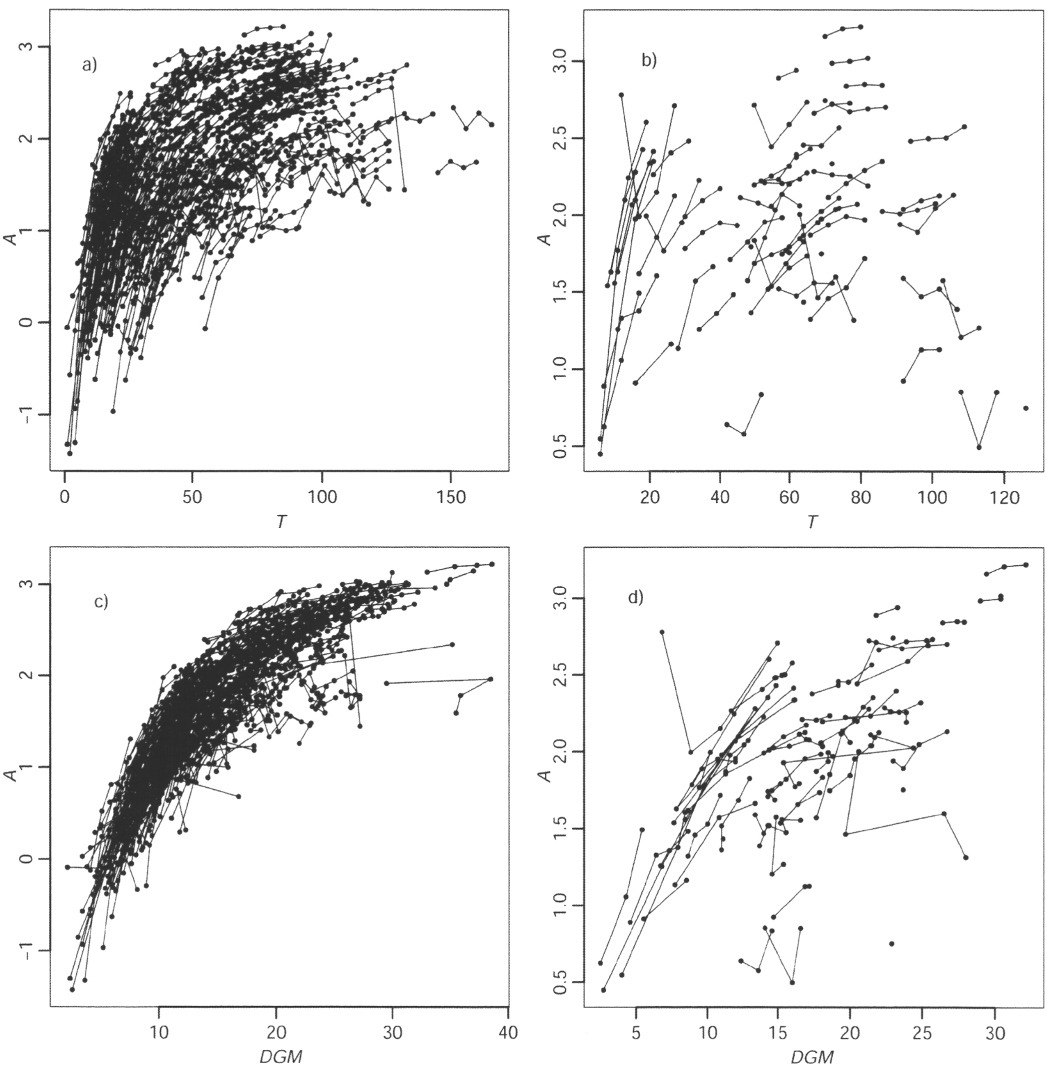


Figure 1. Estimates of parameter *A* of model (2) against stand age (a and b) and stand *DGM* (c and d) in Scots pine data (a and c) and birch data (b and d). The estimates of the same stand at different points in time are connected by lines.

mum height of a tree of that tree species growing at that site (Figures 1a and 1b). This level varies considerably between stands. Another, more interesting feature is that in those stands where the maximum height is low, the overall development of the stand takes longer than in those stands where the maximum height is high. In other words, the development rate of parameter *A* varies between stands, being higher in stands where trees get higher. This implies that two stands of equal age are not at the same stage of development. Plotting the parameters against *DGM* (Figures 1c and 1d) shows that the form of the trend as a function of *DGM* is quite similar in all stands except for a vertical shift in the level of the curve. This implies that stands with equal *DGM* are at the same stage of development, but, because of site properties, location, etc., the asymptote of the *H-D* curve is not equal in all stands with any given *DGM*. Thus, a good strategy in modeling is to link the *H-D* relationship with *DGM* in the model and take the vertical shift in the parameters into account with stand-specific predictors and random parameters.

The estimates of A and B and their estimation errors are strongly correlated. To reduce these correlations, the diameter was reparametrized as

$$x_{kti} = \frac{(D_{kti} + \lambda)^{-C} - (DGM_{kt} + 10 + \lambda)^{-C}}{(10 + \lambda)^{-C} - (30 + \lambda)^{-C}}. \quad (3)$$

The model using this parametrization is

$$\ln(H_{kti}) = A_{kt} - B_{kt} x_{kti} + \varepsilon_{kti}. \quad (4)$$

Parametrization (3) provides interpretations for the parameters of (4): A is the expected logarithmic height of a tree with a diameter of DGM+10 and B is the expected difference in the logarithmic height between diameters of 30 and 10 cm.

The next steps in the analysis were as follows [for more in-depth description, see Mehtätalo (2004) and Lappi (1997)]:

1. Model 4 was fitted with OLS for each stand and measurement occasion.
2. Appropriate trend functions for the estimates of parameters A and B were searched for and fitted as multilevel random effects models using stand and measurement occasion levels.
3. The trend functions were written into Model 4. The nonlinear parameters were used as fixed constants and the linear parameters were re-estimated. An appropriate model for the residual variation as a function of tree diameter was defined and fitted concurrently.
4. Step 1 was repeated using WLS, where the weights were calculated as the inverse of the variance function obtained in step 3. Furthermore, steps 2 and 3 were carried out again.
5. The intercepts of the trend functions were assumed to depend linearly on some stand variables. Different predictor combinations were used to obtain an appropriate model for different practical situations. The assumed dependencies were written into the model, the model was fitted and nonsignificant predictors were dropped stepwise.

The analysis was carried out with the R-implementation of the S-language (Chambers 1998, Venables & Ripley 2002, see http://ww.r-project.org), where package nlme (Pinheiro & Bates 2000) was used for the random effects models.

Fitted models

For both tree species, an appropriate trend function for parameter A was the Chapman-Richards function (Richards 1959)

$$A_{kt} = p_{1a} + p_{2a}\left(1 - e^{-p_{3a}DGM_{kt}}\right)^{p_{4a}} + \alpha_k + \alpha_{kt}, \quad (5)$$

and for parameter B, a linear function of the form

$$B_{kt} = p_{1b} + p_{2b}DGM_{kt} + p_{3b}DGM_{kt}^{\;2} + \beta_k + \beta_{kt} \quad (6)$$

was used. In (5) and (6), p_{1a}, p_{2a}, p_{3a}, p_{4a}, p_{1b}, p_{2b} and p_{3b} are fixed parameters, α_k and β_k stand-level random parameters and α_{kt} and β_{kt} the residual errors, which are later interpreted as measurement occasion level random parameters. In the birch model, the parameter p_{3b} was 0.

The complete model for both tree species is

$$\begin{aligned}\ln(H_{kti}) = {} & (p_{1a}+\alpha_k+\alpha_{kt}) + p_{2a}z_{kt} - (p_{1b}+\beta_k+\beta_{kt})x_{kti} \\ & - p_{2b}DGM_{kt}x_{kti} - p_{3b}DGM_{kt}^{\;2}x_{kti} + \varepsilon_{kti}\end{aligned}, \quad (7)$$

where z_{kt} is the nonlinear part of (5) and the parameters are as explained before. It is assumed that the random effects are normally distributed with a mean of 0 and constant variance. Covariances cov(α_k,β_k) and cov(α_{kt},β_{kt}) may be nonzero but all other covariances between the random effects and the error term are zero. The error term ε_{kti} is assumed to be normally distributed with a variance that depends on tree diameter according to the formula

Table 2. The parameter estimates for the model of Scots pine.

model	p_{1a}	p_{2a}	p_{1b}	p_{2b}	p_{3b}	s.d.(α_k)	s.d.(β_k)	cor(α_k,β_k)	s.d.(α_{kt})	s.d.(β_{kt})	cor(α_{kt},β_{kt})	σ	δ
I	1.552	1.644	0.6156	-0.02707	0.000935	0.1414	0.08945	0.282	0.02878	0.02094	-0.715	0.2585	-0.3724
II	2.283 -0.1025**yk* +0.2042**xk* -0.04661**alt* –0.03714**soil2*	1.631	-0.04113 +0.0917**yk*	-0.02751	0.000964	0.1356	0.08365	0.388	0.02866	0.02157	-0.713	0.258	-0.3716
III	1.688 -0.03007**alt* -0.02927**soil2* -0.5671**DGM* +0.8595**G* +0.03568**thin*	1.47	0.5214 -0.7423**G*	-0.00729	0.000537	0.1295	0.07864	0.624	0.02407	0.01424	-0.62	0.2587	-0.3727
IV	3.025 +0.2431**t* -0.1953**yk* -0.03463**soil2* –1.203**DGM* +0.7613**G* +0.03473**thin*	1.472	0.3147 +0.143**t* -0.7602**G* +0.03006**alt* +0.01621**dd*	-0.0121	0.000555	0.1207	0.07418	0.567	0.02192	0.01578	-0.742	0.2592	-0.3737
V	3.024 -0.1947**yk* +0.2431**t* +0.7668**G*- 1.187**DGM*- 0.0361**soil2* +0.035312**thin*	1.464	0.3297 +0.1301**t* -0.7649**G* +0.02835**alt* +0.01544**dd* +0.565*DGM_p	-0.01818	0.000584	0.1206	0.0748	0.567	0.0225	0.01465	-0.758	0.2597	-0.3745

Note: The predictors are as follows: *yk*, the north coordinate in the Finnish uniform coordinate system, 1000 km; *alt*, altitude above sea level, 100m; *dd*, cumulative temperature sum (average of the years 1951-1980), 100*dd; *G*, total basal area of the stand , 100 m^2/ha; *DGM*, median of basal area weighted diameter distribution of the stand, 100cm; DGM_p, *DGM* of pine, 100 cm; *t*, stand age at breast height, 100 years; *thin*, dummy variable indicating whether the stand has been thinned within the last 10 years and *soil2*, dummy variable which takes the value 1 with stands on sub-dry sites (fertility class 4). σ and δ are coefficients of the variance function, (8).

Table 3. The estimated models for birch. DGM_b is DGM of birch,100 cm. For other notations, see Table 2.

model	p_{1a}	p_{2a}	p_{1b}	p_{2b}	s.d.(α_k)	s.d.(β_k)	corr(α_k,β_k)	s.d.(α_{kt})	s.d.(β_{kt})	corr(α_{kt},β_{kt})	σ	δ
I	0.8241	2.493	0.1417	0.01369	0.2116	0.09047	0.526	0.02617	0.01516	-0.085	0.2975	-0.4535
II	6.434 -0.784*yk	2.454	-0.193 +0.243*xk +0.0226*dd	0.01322	0.1152	0.08599	0.527	0.02166	0.01258	-0.705	0.2986	-0.4546
III	-0.392 +0.08112*alt +0.1222*dd +0.3489*G +0.02684*$thin$	2.257	0.1485	0.01307	0.1168	0.08926	0.637	0.02185	0.01377	-0.503	0.2978	-0.4536
IV	7.435 -0.934*yk +0.02481*$thin$ −1.075*DGM +0.2831*G +0.2194*t	2.544	-0.4089 +0.1419*t +0.0455*dd +0.0553*alt	0.009574	0.1117	0.08485	0.603	0.01836	0.01469	-0.879	0.2975	-0.4536
V	5.421 -0.643*yk +0.02402*$thin$ +0.3152*G −1.504*DGM +1.524*DGM_b	2.445	0.1265 +0.01159* DGM_b	0.005007	0.1059	0.07691	0.564	0.01935	0.01177	-0.832	0.2945	-0.4494

Note: DGM_b is DGM of birch,100 cm. For other notations, see Table 2

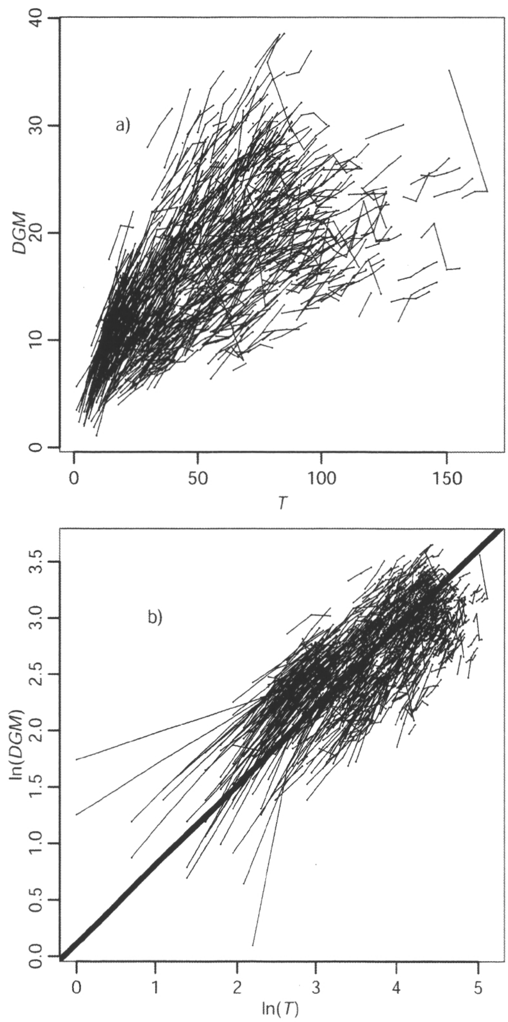


Figure 2. Stand *DGM* against stand age in a data including all stands of the original data. Subfigure a shows the relation on the arithmetic scale and subfigure b shows the relation between logarithmic *DGM* and age. The solid line in subfigure b is the expected development obtained with the fixed part of model (12).

$$\text{var}(\varepsilon_{kti}) = \sigma^2\left[\left(\max\left(D_{kti},\Delta\right)\right)^{-\delta}\right]^2, \qquad (8)$$

where Δ=4.5 for Scots pine and Δ=9.5 for birch.

The estimated nonlinear part of the trend function of *A* was

$$z_{kt} = \left(1-e^{-0.1046*DGM_{kt}}\right)^{1.823} \qquad (9)$$

for Scots pine and

$$z_{kt} = \left(1-e^{-0.06371*DGM_{kt}}\right)^{0.5981} \qquad (10)$$

for birch. In step 5, the fixed parameters p_{1a} and p_{1b} were written as linear combinations of the predictors used in the model, i.e. $p_{1a} = b_{0a} + b_{1a}x_{1a} + b_{2a}x_{2a} + \ldots$ and $p_{1b} = b_{0b} + b_{1b}x_{1b} + b_{2b}x_{2b} + \ldots$, where x_{1a}, x_{1b}, x_{2a}, x_{2b},... are the additional predictors and b_{0a}, b_{0b}, b_{1a}, b_{1b},... are their coefficients.

The estimates of the fixed parameters and variances of the random parameters are in Tables 2 and 3. For both tree species, 5 models with different predictor combinations were estimated. Model I uses only *DGM* as predictor. In model II, variables describing the geographical location are included. Models III and IV include, in addition to the predictors of model II, stand characteristics measured from the whole growing stock and furthermore, model V includes those measured by tree species. The difference between models III and IV is that model IV includes stand age. Only predictors with statistically significant coefficients were included. The significance level used was 1% in Scots pine models and 5% in the models for birch. The significant predictors were searched stepwise, refitting the model and eliminating the least significant parameter until all remaining coefficients were significant.

One can see that the variances of the random parameters decrease when the number of predictors increases. This happens because the variation explained by the additional fixed predictors belongs to the random part in the more sparse models.

Prediction of *DGM* in the future

In the prediction of the *H-D* curve of a stand at a given point in time, the *DGM* of the stand at that point in time needs to be known. With current and past points in time this is not a problem, since in Finnish inventories *DGM* is a very commonly measured mean stand characteristic. However, in order to be able to predict the *H-D* curve in the future, a model for the *DGM* was estimated. Since the *DGM* used in the models is the *DGM* of all trees species belonging to the dominant canopy layer, one and the same model can be used with Scots pine, birch and Norway spruce. To construct such a model the modeling data comprised both the full data used in this study and that used in Mehtätalo (2004).

As seen in Fig. 2a, the *DGM* of a stand with a given age varies considerably, especially with older stands. However, the change in *DGM* is stable with respect to stand age. Thus, if we have an observation of stand *DGM* at some age, we can predict the *DGM* at some other age rather well. The nonlinear trend in Fig. 2a was linearized by taking logarithms from *DGM* and age to obtain the model

$$\ln(DGM_{kt}) = u + v\ln(T_{kt}) + u_k + v_k\ln(T_{kt}) + e_{kt}\,, \qquad (11)$$

where T_{kt} is age of stand k at time t, u and v are fixed parameters, u_k and v_k are stand level random parameters and e_{kt} is the residual. The random parameters and residual error were assumed to be normally distributed with a constant variance. The parameter estimates obtained were u=0.1113(s.e.=0.05411) and v=0.6999 (s.e.=0.01342), var(u_k)= 1.060^2, var(v_k)= 0.2681^2, cov(u_k,v_k)= 0.2693 and var(e_{kt})= 0.05280^2. The linearized model seems to fit the data well (Fig. 2b). When utilizing the model to predict the *DGM* of a stand with a given age, the stand effects u_k and v_k are predicted using the best linear unbiased predictor (e.g. Searle 1971: 458-462, see Lappi 1993, 1997 and Mehtätalo 2004).

Application example

To demonstrate the use of the estimated models, *H-D* curves were predicted with each of the models for one stand selected from the modeling data. The stand was a 35-year-old mixed-species forest with a *DGM* of 12 cm and a basal area of 22m^2/ha, of which 17.7m^2/ha was Scots pine and 3.4m^2/ha birch.

Fig. 3 shows the fixed part predictions obtained with each of the models in Tables 1 and 2. The predictions obtained with the various models differ slightly from each other. However, for both tree species all models give too low heights in this stand.

Fig. 4a demonstrates the effect of localization on the predictions. Models II and V for Scots pine were localized for the sample stand by predicting the stand and time effects of the *H-D* models using one measured height sample tree from the stand. The prediction was carried out with the best linear unbiased predictor; the equations have been presented many times before (Lappi 1993, 1997, Mehtätalo 2004) and are thus not presented here. The localized models give much better height predictions than the fixed parts only. Note that in Fig. 4a the localized models are very close to each other even if the fixed part predictions are not. This is because the information of one measured sample overrides the information of the additional fixed predictors of model V. This demonstrates the somewhat self-evident fact that it is more efficient to improve the prediction of the *H-D* curve of a stand by measuring heights and diameters than by measuring the covariates of model V.

Fig. 4b demonstrates the projection of the *H-D* curve of a stand into the future. The measurements were made at the age of 35 years and the *H-D* curve is projected to the age of 45 years. This required knowledge about the *DGM* at the age of 45 years, which was predicted using model (12). The random effects of model (12) were predicted with BLUP using the known *DGM* at the age of 35 years to obtain a stand-specific *DGM*-age –curve. It was used to predict the *DGM* of the stand at the age of 45 years. The projected *H-D* curves were calculated using the predicted *DGM*. The projections obtained with the localized model are again clearly better than the predictions of the fixed part only but they seem to be underestimates of the height. In fact, all localized models in Fig. 4 seem to give slightly underestimated heights. This is because the expected *H-D* curve is so far from the observations that one measurement does not move it far enough. Using more than one sample tree would reduce the bias of the predictions.

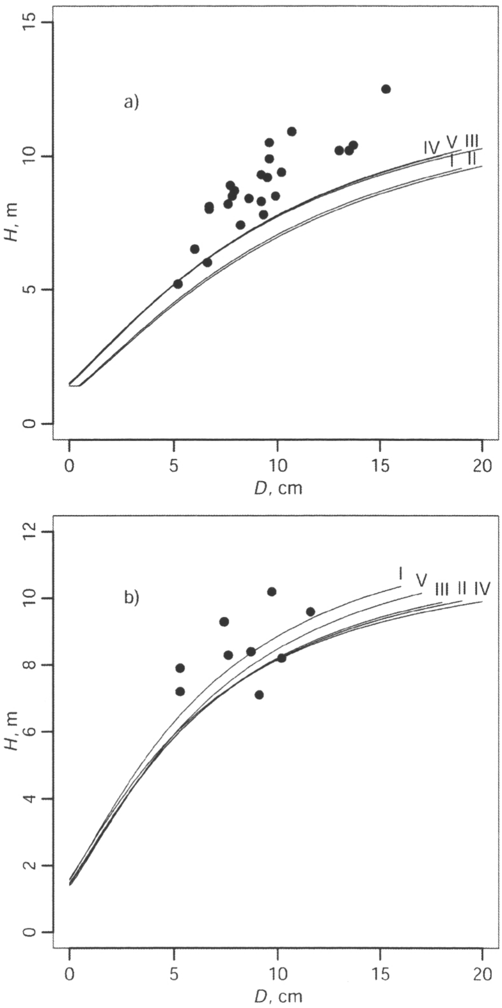


Figure 3 Predicted *H-D* curves of Scots pine (a) and birch (b) in a 35-year-old mixed pine-birch stand selected from the modeling data. The predictions are calculated with each of the models I-V using only the fixed part of the model. The marks show the observed heights and diameters.

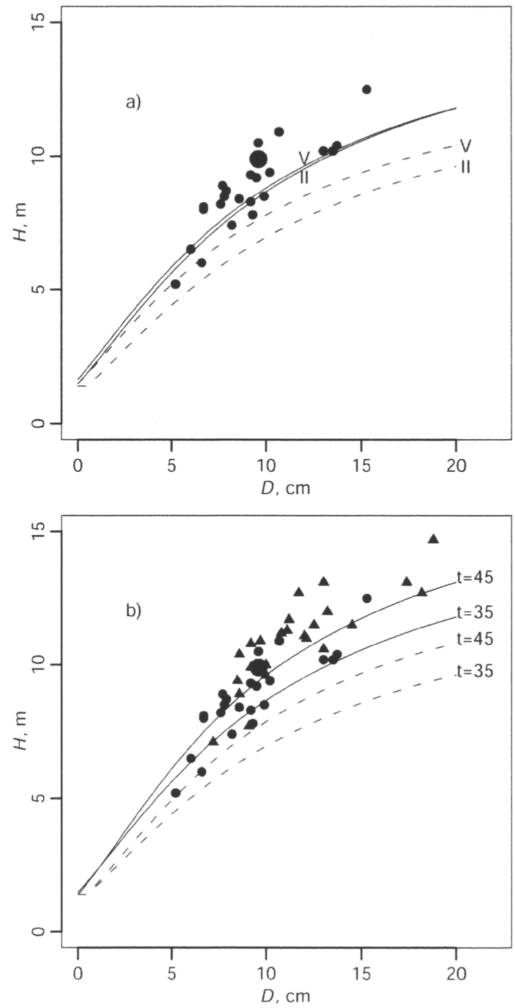


Figure 4. Predicted *H-D* curves of the Scots pines in the stand of Fig. 3. Dashed lines are the predictions of the models obtained using the fixed part only and solid lines are localized using one measured height-diameter pair at the age of 35 years (the big solid ball in the figures). Subfigure a shows the predictions obtained using models II and V at the age of 35 years. In subfigure b, the *H-D* curves after 10 years growth are predicted with model II. Small symbols are the observations (● at the age of 35 years and ▲ at the age of 45 years). The *DGM* of the stand at the age of 35 years was 9.6 cm and the predicted *DGM* at the age of 45 years was 11.5 cm.

Discussion

In this study *H-D* models for Scots pine and birch were estimated for practical use in Finland. The models can be used to predict the *H-D* relationship of a stand with known *DGM*. If other stand characteristics than *DGM* are measured, they can be used through selecting from models I,…,V the model that best suits the situation. Measured height-diameter pairs can be used in localizing the model for a target stand and, due to the longitudinal character of the model, information from any point in time, i.e. any stage of development, can be utilized.

Furthermore, the *H-D* curve of a stand in the future can be predicted, but this requires the prediction of *DGM* at that point in time. These properties of the model are discussed in Lappi (1997) and Mehtätalo (2004).

Mehtätalo (2004) observed that the development of the *H-D* curve of a shade-tolerant tree species (Norway spruce) depends on mean tree size in the stand rather than on stand age. The present study continued this work and showed that the observation made with shade-tolerant tree species is true also with shade-intolerant tree species. The reason for this is that the site properties affect the development rate of a forest stand, so that stands on poor sites develop more slowly and for longer than stands on rich sites. Because of this, the models predicting the development of the *H-D* curve of a stand perform better when mean diameter of the stand is used as the variable describing the maturity of the stand instead of stand age. This effect on the performance is probably also true of models predicting other things than *H-D* curves. Hence, when modeling the development of any stand characteristics, for example, diameter distribution and stand growth, the use of stand age as the only variable determining the stage of development of the stand should be viewed critically.

However, the method for taking into account the effect of *DGM* on the development of *H-D* curves which has been presented here is not the only correct one; other approaches may also lead to equally good results. For example, the development of parameters *A* and *B* can be linked with stand age and the *DGM* can be used in the prediction of the random effects as Lappi (1997) did in his models 6 and 7. Thus, if both *DGM* and age are known, it is not obvious that the development of *H-D* curves should be linked with *DGM* and linking it with age may work as well, if the *DGM* is taken into account in the models. However, using age only will not lead to as good models as will the use of *DGM* only and thus, if one of them must be chosen, it would be preferable to use the *DGM.*

Acknowledgements

This work was part of the project "Statistical modeling in forest management planning", which was funded by the Academy of Finland (decision number 73392) and carried out at the Finnish Forest Research Institute. The author would like to thank Dr Juha Lappi for comments and criticism on the manuscript and Dr Lisa Lena Opas-Hänninen for revising my English.

Literature cited

Arabatzis, A. A. & Burkhart, H. E. 1992. An evaluation of sampling methods and model forms for estimating height-diameter relationships in loblolly pine plantations. For. Sci. 38: 192-198.

Chambers, J. M. 1998. Programming with data. Springer, New York.

Curtis, R. O. 1967. Height-diameter and height-diameter-age equations for second-growth douglas-fir. For. Sci. 13: 365-375.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. 2002. Analysis of longitudinal data. Second edition. Oxford Statistical Science Series. Oxford university press, Oxford, U.K. 379 p.

Eerikäinen, K. 2003. Predicting the height-diameter pattern of planted Pinus kesiya stands in Zambia and Zimbabwe. For. Ecol. Manage. 175: 355-366.

Fang, Z. & Bailey, R. L. 1998. Height-diameter models for tropical forests on Hainan island in southern China. For. Ecol. and Manage. 110: 315-327.

Flewelling, J. & de Jong, R. 1994. Considerations in simultaneous curve fitting for repeated height-diameter measurements. Can. J. For. Res. 24: 1408-1414.

Huang, S., Titus, S. J. and Wiens, D. P. 1992. Comparison of nonlinear height-diameter functions for major Alberta tree species. Can. J. For. Res. 22: 1297-1304.

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.

Lappi, J. 1991. Calibration of height and volume equations with random parameters. For. Sci. 37: 781-801.

Lappi, J. 1997. A longitudinal analysis of height/diameter curves. For. Sci. 43: 555-570.

Lynch, T. B. & Murphy, P. A. 1995. A compatible height prediction and projection system for individual trees in natural, even-aged shortleaf pine stands. For. Sci. 41: 194-209.

Mehtätalo, L. 2004. A longitudinal height-diameter model for Norway spruce in Finland. Can. J. For. Res. 34(1): 131-140.

Pinheiro, J. C & Bates, D. M.. 2000. mixed-effects models in S and S-PLUS. Springer-Verlag, New York, USA. 528 p.

Richards 1959. A flexible growth function for empirical use. Journal of Experimental Botany 10: 290-300.

Searle, S. R. 1971. Linear models. John Wiley & Sons, New York, USA. 532 p.

Venables, W.N. & Ripley, B D. 2002. Modern applied statistics with S. Fourth edition. Springer-Verlag, New York, USA. 495 p.

Zakrzewski, W. T. & Bella, I. E. 1988. Two new height models for volume estimation of lodgepole pine stands. Can. J. For. Res. 18: 195-201.

V

AN APPROACH TO OPTIMIZING FIELD DATA COLLECTION IN AN INVENTORY BY COMPARTMENTS

Lauri Mehtätalo[1] and Annika Kangas[2]

[1] *(Corresponding author), Researcher, Finnish Forest Research Institute, Joensuu Research Centre, P.O. Box 68, FIN-80101 Joensuu, Finland*

[2] *University of Helsinki, Department of Forest Resource Management, P.O. Box 27, FIN-00014 University of Helsinki, Finland*

Abstract. This study presents models for the expected error of the total volume and saw timber volume due to sampling errors of stand measurements. The measurements considered are horizontal point sample (HPS) plots, stem numbers from circular plots, sample tree heights, sample order statistics (i.e. quantile trees) and sample tree heights from the previous inventory. Different measurement strategies were constructed by systematically varying the numbers of these measurements. A model system developed for this study was utilized in a dataset of 170 stands to predict the total volume and saw timber volume of each stand with each measurement strategy. The errors of these volumes were modeled using stand characteristics and the numbers of measurements as predictors. The most important factors affecting the error in the total volume were the numbers of HPS plots and height sample trees. In addition, the number of quantile trees had a strong effect on the error of saw timber volume. The errors were slightly reduced when an old height measurement was used. There were significant interactions between stand characteristics and measurement strategies. Thus the optimal measurement strategy varies between stands. It was demonstrated how constrained optimization can be used to find the optimal strategy for any one stand.

Key words: forest planning, response surface, optimization, quantile, diameter distribution, *H-D* curve, prediction

Introduction

Forest management planning in Finland is based on inventory by compartments of the target area. In the inventory, basal area, basal area median diameter (i.e., the median diameter of the basal area weighted diameter distribution, *DGM*), height of the *DGM*-tree, stand age and site fertility class of each stand are estimated with the aid of a few subjectively located horizontal point sample (HPS) plots. Using these data, estimates of the diameter distribution and height-diameter curve (*H-D*-curve) of the stand are produced. These estimates are used to build up a set of representative trees (i.e. a sample from a diameter distribution), which are then used in the simulation of alternative management schedules for the stand.

The system described above produces, in most cases, a reasonable description of a stand. There are, however, many things that are not taken into account or that could be done better in the system. One weakness is that, due to the subjective location of sample plots and visual assessment of stand characteristics, the sampling errors cannot be calculated using the standard formulas from sampling theory. Thus, the accuracy of stand measurements is not known and cannot be taken into account in the predictions. Furthermore, no assessment of the accuracy of the predictions can be given. Because the system includes chains of models with nonlinear transformations, random errors of the predictors may cause bias in the predictions (Gertner 1991, Kangas 1997, Kangas 1999). Another weakness is that the combination of measurements is fixed. This combination may not be optimal and thus reallocation of the measurement resources (i.e., the measurement time) to different measurements could improve the accuracy of the stand description and growth estimates. For example, Kangas and Maltamo (2000c) suggested that substituting height sample trees for HPS plots might improve the accuracy of volume predictions.

Kangas and Maltamo (2002) studied the usefulness of different measurements in the estimation of total volume. They utilized a calibration (or adjustment) algorithm (Kangas and Maltamo 2000a) to predict the stand description using several measurement strategies and estimated models for the prediction variance of stand volume as a function of basal area, *DGM* and stem number of the stand. The best measurement strategy varied between stands and depended on stand characteristics. However, their model did not take into account the sampling and measurement errors of the stand characteristics. According to previous studies, the standard errors (including sampling and measurements errors) of partly subjective and visual stand assessments may even be more than 30% of the true values of the stand characteristics (Mähönen 1984, Pigg 1994, Kangas et al. 2002, Kangas et al. 2004, Haara and Korhonen 2004). In this study, in order to be able to control the errors of stand measurements, sample plot measurements are assumed to be accurate and the sample plots are located randomly in the stand. Thus, the stand measurements include only sampling errors, which are controlled by varying the number of measurements carried out in a stand.

Response surfaces (Khuri and Cornell 1987) can be used to study the effect of one or more experimental factors on the outcome of a complex system. For example, they are used in studying the effect of errors in predictors on the response of a process-based forest growth model (Gertner et al. 1996). The idea is to vary systematically the values of the experimental factors and calculate the outcome of the system using each combination of them. The response surface is a regression model that has the experimental factors as predictors and the outcome as the dependent variable. The values of the experimental factors are selected so that the model matrix is orthogonal. The approach of this study resembles the response surface approach. However, the analysis is not carried out in a single stand but in a dataset consisting of several stands and the stand variables are also treated as experimental factors. Thus, the model matrix of the surface is not orthogonal. However, the estimated surface can be used in the prediction of responses, in a similar manner as the ordinary regression model.

This study presents a model that can be used in the optimization of stand measurements in an inventory which is carried out by compartments and where sample plots are located objectively. The model predicts the expected error in the predicted total volume and saw timber volume of the stand using the stand characteristics and the accuracy of different stand measurements as predictors. The model can be used in comparing different measurement strategies and in finding the optimal one for a single stand. The limitations of the traditional calculation system made it necessary to build up a new one for this study. The aims of this study were: (a) to build up a new system for producing a stand description for forest planning, i.e. a system that is able to utilize different kinds of information measured with different levels of accuracy, (b) to estimate a model for the prediction variance of total volume and saw timber volume using stand characteristics and the accuracy of the stand measurements as predictors, and (c) to show examples of how this model can be used in anticipating the prediction variance of a stand description, i.e. in allocating the measurement time to different measurements optimally.

The system for producing the stand description

A model system for the prediction of stand description was developed. The input of the system includes the values of the stand variables (basal area, *DGM*, age, site fertility class, information about thinning during the past ten years, stand coordinates, temperature sum, and altitude), the variance-covariance matrix of their errors (including sampling and measurement errors), any number of height sample trees from any points in time (i.e., trees with known diameter and height) and any number of sample order statistics (i.e., trees with known diameter and rank at the sample plot). The sample order statistics are later called quantile trees. The output of the system is a stand description that includes the basal area, basal area diameter distribution and *H-D* curve of the stand; this output can be used in the calculation of volume, saw timber volume, dominant height, etc.

The R-implementation of the S language was used in implementing the system on a computer (R Development Core Team 2003). However, some parts of the system (numerical integration, constrained nonlinear optimization) could not be done with standard R-functions and IMSL-subroutines were used (IMSL 1997). They were linked to R.

The production of the stand description comprises the following four stages. Since all parts of the system have been published elsewhere, the algorithms are not presented here in detail. It is not necessary to understand the algorithms in order to grasp the concept as a whole and therefore skipping the following four subsections will not prevent the reader from understanding the rest of the article.

Stage 1: Predicting the expected diameter percentiles

The expected diameter percentiles are predicted with the models of Kangas and Maltamo (2000b). The models use a percentile-based approach (Borders et al. 1987), predicting the logarithmic 0th, 10th, …, 80th, 90th, 95th and 100th percentiles of the basal area diameter distribution. The fixed predictors of the model are *DGM*, age, basal area and dummy variable for site fertility. The model for the diameter percentiles of stand *m* is of the form

$$\ln(\mathbf{d}_m) = \mathbf{B}\mathbf{x}_m + \mathbf{e}_m = E(\ln \mathbf{d} | \mathbf{x}_m) + \mathbf{e}_m , \qquad (1)$$

where $\mathbf{B}$ is the matrix of fixed parameters, $\mathbf{x}_m$ is the vector of predictors in stand *m*, and $\mathbf{e}_m$ is the vector of residuals in stand *m* with an expectation of 0 and variance-covariance matrix $\mathbf{D}$. Assuming that the model is correct and the parameters are known, the fixed part of the model gives the conditional expectations of the logarithmic percentiles given $\mathbf{x}_m$, and the residuals $\mathbf{e}_m$ are the deviations of the true logarithmic percentiles of stand *m* from their conditional expectations, later called stand effects. The conditional expectations of the percentiles of stand *m* are predicted using model (1). The continuous distribution function is obtained by linear interpolation on these predictions.

The measurement errors in the estimation data were so small that their effect on the parameter estimates is neglected. However, when utilizing the models in real forest inventories, the measurement and sampling errors are much larger. Because of the nonlinear transformations, these errors cause bias in the predicted diameter percentiles and affect the variance-covariance matrix of the error term, which is needed later. The effect of errors in predictors on the predicted distribution is shown in the Appendix, as are the derivations of the corrected predictions of the percentiles and the error variance-covariance matrix. The formulas given in the Appendix (Equations A9 and A10) were used for prediction.

Stage 2: Localizing the diameter distribution for a stand

If one has measured the diameter of a sample tree and, in addition, knows its rank r in a sample of size n, the measured diameter is the r^{th} order statistic of the sample, $D_{r:n}$. It can be shown (Mehtätalo 2004d) that an order statistic of an i.i.d. diameter sample is an unbiased estimate of a certain, say $100p^{\text{th}}$, diameter percentile of the stand, where the value of p depends on the diameter distribution of the stand (we will return to this later). In other words, the diameter of a quantile tree, $D_{r:n}$, is a measured diameter percentile of the stand. The measured percentiles can be utilized in localizing the expected percentiles of the stand by predicting the stand effects $\mathbf{e}_m$ of model (1). The localization utilizes the variance-covariance matrix of the measurement error (i.e. sampling error) of the measured percentiles and the variance-covariance matrix of stand effects. The former is derived using the theory of order statistics (Reiss 1989: 21, 30-31, see Mehtätalo 2004d) and the latter was obtained in stage 1.

The predicted percentiles follow model (1). Furthermore, the measured percentiles follow the model

$$\ln(\mathbf{d}^*_m) = E(\ln \mathbf{d}^* | \mathbf{x}_m) + \mathbf{e}^*_m + \boldsymbol{\varepsilon}_m, \qquad (2)$$

where the logarithmic measurements are in vector $\ln(\mathbf{d}^*_m)$, their conditional expectations in vector $E(\ln\mathbf{d}^*|\mathbf{x}_m)$, the stand effects in vector $\mathbf{e}^*_m$ and the measurement errors in vector $\boldsymbol{\varepsilon}_m$. The measured diameters estimate the $100\times\mathbf{p}^{*\text{th}}$ percentiles of the stand. The elements of $\mathbf{p}^*$ are calculated as $p^* = F^{-1}[E(D_{r:n})]$, where F is the cumulative diameter distribution function based on the conditional expectations of the percentiles and $E(D_{r:n})$ is the expectation of the measured order statistic, which depends on the diameter distribution of the stand [for derivation of $E(D_{r:n})$, see Mehtätalo (2004d)]. At this step, F is used as the diameter distribution of the stand (we will return to this later). Vector $E(\ln\mathbf{d}^*|\mathbf{x}_m)$ is obtained by interpolation of the predicted logarithmic percentiles for the values of $\mathbf{p}^*$. Vectors $\mathbf{e}_m$, $\mathbf{e}^*_m$ and $\boldsymbol{\varepsilon}_m$ will be predicted under the assumptions $E(\mathbf{e}_m)=E(\mathbf{e}^*_m)=E(\boldsymbol{\varepsilon}_m)=\mathbf{0}$, $\text{var}(\mathbf{e}_m)=\mathbf{D}$, $\text{var}(\mathbf{e}^*_m)=\mathbf{D}^*$, $\text{var}(\boldsymbol{\varepsilon}_m)=\mathbf{R}$, $\text{cov}(\mathbf{e}_m,\mathbf{e}^*_m)=\mathbf{C}$ and $\text{cov}(\mathbf{e}^*_m, \boldsymbol{\varepsilon}_m)=\mathbf{0}$, where matrices $\mathbf{D}$, $\mathbf{D}^*$, $\mathbf{C}$ and $\mathbf{R}$ are known: matrix $\mathbf{D}$ is the corrected variance-covariance-matrix of the stand effects of model (1) (Equation A10) and matrices $\mathbf{D}^*$ and $\mathbf{C}$ are derived from it with linear interpolation. The matrix $\mathbf{R}$ includes variances and covariances of measured order statistics, which can be derived using their probability distributions (Reiss 1989: 21, 30-31, see Mehtätalo 2004d).

The aim is to predict the unobserved random vector $\mathbf{e}_m$ using the observed random vector $\ln(\mathbf{d}^*_m) - E(\ln\mathbf{d}^*|\mathbf{x}_m)$. The prediction is based on the theory of linear prediction (McCulloch and Searle 2001: 169, Searle et al. 1992: 269-275). The best linear unbiased predictor of vector $\mathbf{e}_m$ is:

$$\hat{\mathbf{e}}_m = \mathbf{C}(\mathbf{D}^* + \mathbf{R})^{-1}\left[\ln\mathbf{d}^*_m - E(\ln\mathbf{d}^*|\mathbf{x}_m)\right] \qquad (3)$$

and the variance-covariance matrix of the prediction error is

$$\text{var}(\hat{\mathbf{e}}_m - \mathbf{e}_m) = \mathbf{D} - \mathbf{C}(\mathbf{D}^* + \mathbf{R})^{-1}\mathbf{C}'. \qquad (4)$$

The localized logarithmic percentiles are obtained by adding the predicted stand effects to the conditional expectations of the percentiles as in

$$\hat{\ln}(\mathbf{d}_m) = \mathbf{B}\mathbf{x}_m + \hat{\mathbf{e}}_m \,. \qquad (5)$$

Because we have obtained a improved prediction of the diameter distribution of the stand, new estimates of $E(D_{r:n})$ and **p*** can now be calculated using the localized diameter distribution (Equation 5). Therefore, the localization is repeated until the iteration has no considerable effect on the predicted distribution.

In some cases, the localized percentiles are not obtained because the distribution of some iteration step is not monotone or the iteration does not converge (see Mehtätalo 2004d). In these cases, the localization is carried out using a reduced set of quantile trees, including the same trees as before but dropping one tree randomly.

Stage 3: Ensuring compatibility with measured stand variables

The predicted diameter distribution is compatible with the measured basal area and *DGM*. If additional stand characteristics dependent on the diameter distribution are measured, the compatibility is no longer guaranteed. Mehtätalo (2004a) presented an algorithm for ensuring compatibility of predicted percentiles, basal area, *DGM* and stem number. In this study, the stem number of trees with a diameter of 5 cm or more was known. Thus, a modification of the algorithm of Mehtätalo (2004a) was needed in order to take into account the fact that the stem number was measured only above a fixed threshold diameter of 5 cm.

Let $\hat{d}_1,...,\hat{d}_M$ be the predicted $100^*p_1,...,100^*p_M{}^{\text{th}}$ diameter percentiles from stages 1 or 2 and $\hat{F}$ the cumulative distribution function obtained from them with linear interpolation, where $p_1, p_2, ... p_{M-1}, p_M$ are the fixed values given in stage 1 (i.e., 0, 0.10,…,0.95,1.0). Furthermore, let k be the index of the smallest percentile above 5 cm. The modified algorithm adjusts only percentiles $\hat{d}_k,...,\hat{d}_M$. The adjusted combination of percentiles and stand variables is the combination that is compatible with and as close as possible to the original combination with regard to a specified distance measure. The distance (Equation 6a) is defined as the weighted sum of squared deviations from the predicted percentiles and measured stand variables, where the weights are the inverse errors of the percentiles and stand variables. The compatibility requirement is included in the algorithm as a constraint (Equation 6d) in which the stem number based on the adjusted percentiles is equated to the adjusted stem number. Additional constraints (Equations 6b, 6c, 6e and 6f) are needed to ensure monotony of the adjusted percentiles and non-negativity of the percentiles and stand variables. The adjusted set of percentiles and stand variables is found by solving the following optimization problem:

minimize

$$z = \frac{s_k^2}{\sigma_k^2} + \sum_{i=k}^{M-1} \frac{\left(s_{i+1} - s_i\right)^2}{\sigma_{i,i+1}^2} + \frac{s_{50\%}^2}{\sigma_{50\%}^2} + \frac{s_G^2}{\sigma_G^2} + \frac{s_N^2}{\sigma_N^2} \tag{6a}$$

subject to

$$\hat{d}_i + s_i + \varepsilon \le \hat{d}_{i+1} + s_{i+1}, \text{ for } i = k, k+1, \ldots, M-1, \tag{6b}$$

$$\hat{d}_k + s_k + 0.01 \ge 5, \tag{6c}$$

$$\frac{4\left(\hat{G}_{d\ge5} + s_G\right)}{100\pi} \left\{ \frac{p_k - \hat{F}(5)}{\left(\hat{d}_k + s_k\right) - 5} \left(\frac{1}{5} - \frac{1}{\left(\hat{d}_k + s_k\right)} \right) + \sum_{i=k}^{M-1} \left[\frac{p_{i+1} - p_i}{\left(\hat{d}_{i+1} + s_{i+1}\right) - \left(\hat{d}_i + s_i\right)} \left(\frac{1}{\left(\hat{d}_i + s_i\right)} - \frac{1}{\left(\hat{d}_{i+1} + s_{i+1}\right)} \right) \right] \right\} = \hat{N}_{d\ge5} + s_N, \tag{6d}$$

$$\hat{G}_{d\ge5} + s_G \ge 0, \tag{6e}$$

$$\hat{N}_{d\ge5} + s_N \ge 0, \tag{6f}$$

where s_i is the deviation of the adjusted $100{*}p_i^{\text{th}}$ percentile from its predicted value, σ_i^2 the variance of its prediction error, $\sigma_{i,i+1}^2 = \text{var}\left(\hat{d}_{i+1} - \hat{d}_i\right)$, ε a constant determining the minimum allowed distance between adjacent percentiles, $\hat{G}_{d\ge5}$ the measured basal area of trees above 5 cm calculated as $\left[1 - \hat{F}(5)\right]\hat{G}$, $\hat{N}_{d\ge5}$ the measured stem number of trees above 5 cm, s_N and s_G the deviations of adjusted basal area and stem number from their measurements and σ_N^2 and σ_G^2 their error variances, respectively. The adjusted percentiles are calculated as $\hat{d}_i + s_i$ for $i=k,\ldots,M$ and adjusted basal area and stem number as $\hat{G} + s_G$ and $\hat{N}_{d\ge5} + s_N$, respectively.

This stage produces an adjusted set of percentiles and stand variables. The algorithm also adjusts the values of stand variables, and the distribution is scaled to the adjusted basal area instead of the original measured one.

Note that if the measurements of quantile trees or stem number are not available, stages 2 and 3 can be omitted without causing any problems.

Stage 4: Predicting the H-D curve

A height model of Mehtätalo (2004c) is used to predict the *H-D* curve of the stand. The model is called model V in Mehtätalo (2004c) and it is an application of model V of Mehtätalo (2004b) for Scots pine. The predictors are basal area, *DGM*, stand age, site fertility class, stand coordinates, altitude, cumulative temperature sum and a dummy variable indicating whether the stand has been thinned during the past 10 years. The model was estimated from longitudinal data applying a mixed model approach. With this model, the *H-D* curve can be predicted using commonly measured stand variables. The *H-D* model can be localized into a new stand using any number of measured heights and diameters. The height measurements may be from several points in time, for example, height measurements from the previous inventory can be used. The localization is based on predicting the stand and time-effects of the models using the standard linear prediction theory (see Lappi 1991, Lappi 1997, Mehtätalo 2004b).

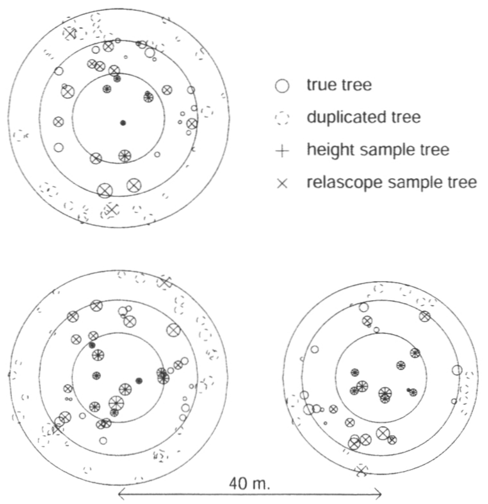


Figure 1. The map of sample plots and trees in one stand. The diameter of the mark is proportional to that of the tree.

Data

The primary dataset of the study

This study used a permanent dataset collected by the Finnish Forest Research Institute between the years 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 measured 1-3 times with 5-year intervals. However, only the first and third measurement occasions were used in this study. 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. In addition, the height was recorded from approximately one third of the trees, i.e. those trees which belong to the innermost circle of the plot (see Figure 1). This study used only the Scots pine trees of the data.

The dataset of this study included those stands where no regeneration cuts were made between the first and third measurement occasions, the total volume of Scots pine was more than 10m^3/ha, the total basal area of Scots pine was more than 4 m^2/ha and a HPS plot using a basal area factor of 1 could be established on all plots of the first and third measurement occasions (see below). The total number of stands in the data was 170. The stand description was predicted for the third measurement occasion using new measurements simulated using the data of the third occasion and old measurements simulated using the data of the first occasion.

Data preparation

In order to calculate the (assumed) true stand volume and saw timber volume of a stand, the heights of all sample trees were needed. The unknown heights were predicted using the model of Mehtätalo (2004c) (model V), which was localized for each stand using the measured height sample trees of the third measurement occasion. The saw timber volume was defined as the total volume of those parts of stems that include at least a 4-meter long log with a minimum top diameter of 15 cm. In the calculation of volume, the taper curve functions based on diameter and height were used (Laasasenaho 1982).

The within-stand variances of *DGM*, basal area and stem number were needed in order to calculate the sampling errors of the stand measurements in the simulation. They were estimated from the measurements generated on the three sample plots of the stand. The generated measurements of basal area and *DGM* were made from HPS plots using a basal area factor of 1 and the measurements of stem number from a circular plot with a radius of 3.99 m. It was assumed that the true values of *DGM* and basal area could be obtained using all sample trees of the stand, i.e. it was assumed that the expectations of the measurements were known. The within-stand sampling variance of random variable *X* was calculated using the formula

$$\operatorname{var}\left(\hat{X}\right) = \frac{\sum_{i=1}^{3}\left[\hat{X}_i - E(X)\right]^2}{3}, \qquad (7)$$

where $E(X)$ is calculated from all sample trees of the stand and $\hat{X}_i$ is the measurement generated on the i^{th} sample plot. The covariance of sampling errors of variables X and Y was calculated correspondingly as

$$\operatorname{cov}\left(\hat{X},\hat{Y}\right)=\frac{\sum_{i=1}^{3}\left[\hat{X}_i-E(X)\right]\left[\hat{Y}_i-E(Y)\right]}{3}. \quad (8)$$

The radius of the HPS plot determined by the thickest tree of the circular plot was often greater than the radius of the circular plot. In this case, all trees that would have belonged to the HPS plot were not present in the data, i.e. the circular plot should have been larger in order to include all trees of the HPS-plot. The area of the missing surface was calculated by subtracting the area of the circular sample plot, A_f, from the area determined by the maximum diameter of the plot, D_{max},

$$A_s=\pi\left(D_{\max}/2\right)^2-A_f, \quad (9)$$

where A_s is the area of the missing surface. To obtain unbiased measurements on the HPS plots, trees were generated on the surface. The stock density and diameter distribution on the surface were assumed to be similar to those on the fixed plot and the tree locations were assumed to be random. These criteria were fulfilled by duplicating trees of the circular plot on the missing surface and placing them at random locations as follows (see Figure 1). A random number (from $U(0,1)$) was generated for each tree on the circular plot. If it was less than the ratio A_s/A_f, the tree was included in the surface. If A_s was greater than A_f, for one or more plots in a stand, the stand was not included in the data. The distances of the duplicated trees from the center of the plot were generated using the formula $\sqrt{\left(r_{D\max}{}^2-r_f{}^2\right)U(0,1)+r_f{}^2}$, where $r_{D\max}$ and r_f are the radii corresponding to the areas in Equation 9 and $U(0,1)$ is a random number from the uniform distribution between 0 and 1.

The result of the above calculations was a stand specific dataset including the true values of the stand characteristics and the within-stand variances and correlations of basal area, *DGM* and stem number on the first and third measurement occasion. In addition, the data included the (assumed) true total and saw timber volumes of the third measurement occasion. A short summary of the data is given in Table 1 and histograms of estimated within-stand standard deviations and correlations in Figure 2.

In addition to the stand specific dataset, datasets of potential new and old sample trees were generated. The dataset of potential new sample trees included those trees of the third measurement occasion that belonged to the HPS plots established at the centers of the circular plots. However, only the true height sample trees were included, i.e. the duplicated trees and trees with unknown height were deleted after determination of the ranks and the total number of trees on the HPS plot. For each potential sample tree, diameter, height, rank on the HPS plot and total number of trees on the plot were saved. The dataset of potential old sample trees was generated in a similar manner from the data of the first measurement occasion but only heights and diameters of the plot specific basal area median trees were included.

Table 1. Summaries of the data.

	min.	mean	max.
stand age, years	22	72	180
basal area, m^2/ha	4.5	15.7	28.5
stem number, 1/ha	123	1052	3463
DGM, cm	6.3	16.8	32.0
total volume, m^3/ha	19.7	107.1	308.8
saw timber volume, m^3/ha	0	47.4	254.4

Note. *DGM* is the basal area weighted median diameter of the stand.

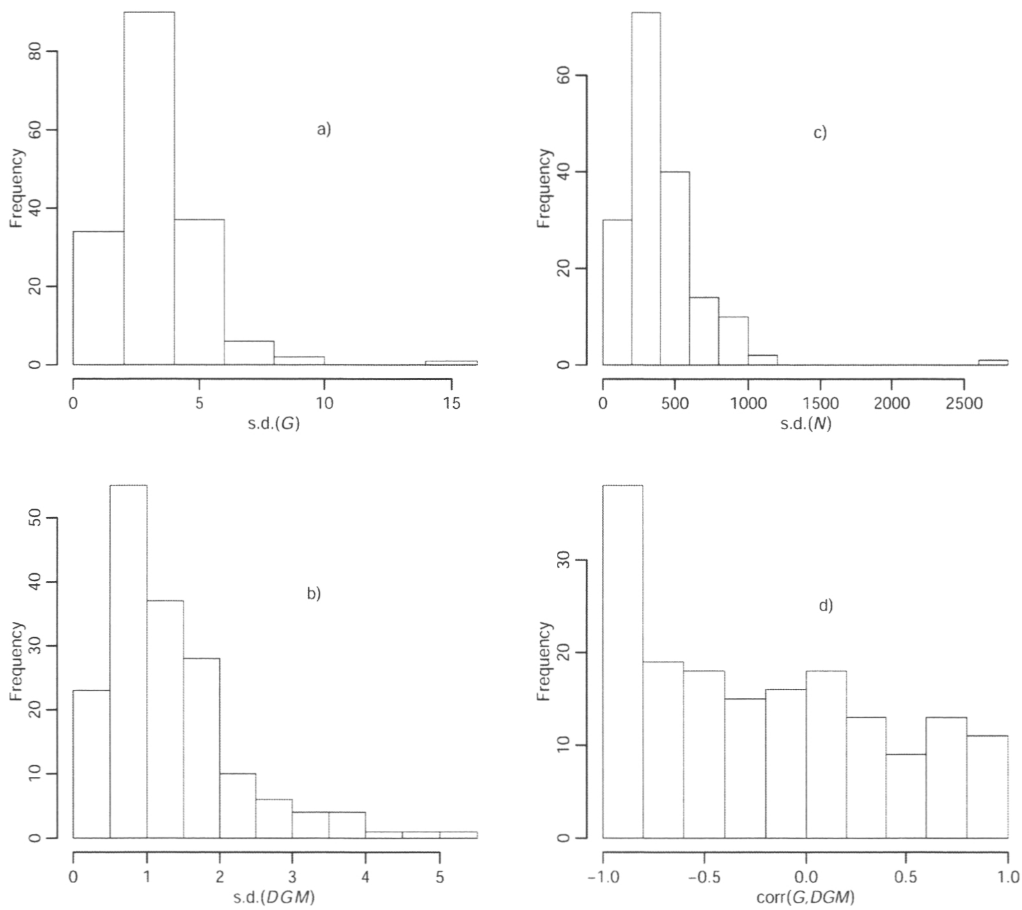


Figure 2. Histograms of the estimated within-stand standard deviations of basal area (a), *DGM* (b), stem number (c) and the within-stand correlations of basal area and *DGM* (d) on the third measurement occasion.

Construction of the data for the response surface models

The next step in the analysis was to define the measurement strategies and simulate realizations of them for each stand. These realizations were then used to study the accuracy of predicted stand description with different measurement strategies.

The simulated inventory of a stand comprised various amounts of the following measurements:

1. Measurement of a basal-area and *DGM* from an HPS plot using a basal area factor of 1 (G-plot). The basal area is then the number of trees on the plot and the basal area median diameter (*DGM*) of the plot is determined as the median diameter of the sample.
2. Measurement of the stem number by counting the number of trees above 5 cm from a plot with a radius of 3.99 m (N-plot). The N-plots were assumed to be from different locations than the G-plots (i.e. no correlation between the sampling errors of N- and G-plots was assumed) in order to derive more information from the N-plots and to avoid problems in the simulation of different numbers of G- and N-plots. [see Husch et al. (1982, p. 258-260) for previous work on combining measurements from HPS-plots with measurements from fixed-area plots.]
3. Measurement of diameter and height of one of the trees on the G-plot (H-tree).

4. Measurement of diameter and rank of one of the trees on the G-plot (Q-tree).
5. In addition, it was assumed that one 10-year-old height measurement from the *DGM*-tree might be available from the stand (old H-tree).

A measurement strategy of the stand defines how many G-plots, N-plots, H-trees, Q-trees and old H-trees are measured from a stand. A total of 384 (4×3×4×4×2) measurement strategies were constructed by varying systematically the number of measurements carried out in a stand between the assumed minimum and maximum values that would be used in practice. The number of G-plots was 1, 3, 5 or 7, the number of N-plots 0, 2 or 4, the number of H-trees 0, 2, 4 or 6, the number of Q-trees 0, 2, 4 or 6 and the number of old H-trees 0 or 1.

When simulating a realization of a measurement strategy, the measurements of H- and Q-trees were selected randomly from the set of potential new sample trees and the old H-tree measurement was selected randomly from the set of potential old sample trees of the stand. The H- and Q-trees were the same trees whenever possible, which minimized the number of diameter measurements in the inventory. Since the required amounts of G- and N plot measurements could be more than three, the true sample plots were not used, a Monte Carlo approach was used instead. The measurements were generated by adding errors from a multinormal distribution to the true values of the stand characteristics. The plot measurements were assumed to be accurate and the generated measurements included only a sampling error component. The error variances were obtained by dividing the within-stand variance obtained with Equation 7 by the number of plots. The correlation between *DGM* and basal area measurements was the within-stand correlation based on Equation 8 and their correlation with stem number was zero since the measurements were assumed to be from different locations.

When an old height measurement was used in the prediction of the *H-D* curve, old *DGM* and basal area measurements were needed in the calculation of the fixed part of the height model. The old measurements were generated in a similar manner as the new ones, but assuming that the number of sample plots in the previous inventory was 5, i.e. the variance of the generated measurements was the within-stand variance of the first measurement occasion divided by 5. The information about other stand variables (age, site fertility, information about thinning, stand coordinates, altitude, temperature sum) was assumed to be as accurate as in the data of this study and no measurement errors were added to them.

Ten realizations of each measurement strategy were produced for each stand and the diameter distribution and *H-D* curve were predicted using the system described previously. Using these predictions, the total volume and saw timber volume were calculated for each realization. The root mean squared errors (*RMSE*) of these characteristics were calculated from the 10 realizations for each stand and measurement strategy, thus resulting in 65280 (=170 stands×384 strategies) *RMSE* values of total and saw timber volumes. The *RMSE* of variable *V* was calculated using the equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{10}\left(\tilde{V}_i - V_{true}\right)^2}{10}}, \tag{10}$$

where $\tilde{V}_i$ is the i^{th} observation and V_{true} is the true value of variable *V*.

Modeling the errors

The *RMSE* of volume and saw timber volume were modeled using the number of measurements (G-plots, N-plots, H-trees, Q-trees and old H-trees) and stand characteristics (basal area and *DGM*) as predictors. The *RMSE* using measurement strategy *j* in stand *k* was assumed to follow the model

$$y_{kj} = \mathbf{x}_{kj}\mathbf{b} + u_k + v_j + e_{kj}, \tag{11}$$

where y_{kj} is the *RMSE*, $\mathbf{x}_{kj}$ includes the fixed predictors and $\mathbf{b}$ the fixed parameters, u_k is a random effect for stand *k*, v_j a random effect for strategy *j* and e_{kj} the residual error for strategy *j* in stand *k*. The variance of the residual error, var(e_{kj}), seemed to increase as the predicted value increased and it was taken into account by using weights in the estimation of the model. The weights were obtained from the inverse of a variance function that was fitted using an approach similar to that of Lappi (1997). The mixed model approach was used to take into account that the observations of a stand are correlated, as are the observations of a strategy. The model was estimated using the MIXED procedure of SAS (Littell et al. 1996).

Assuming that the sample plots and sample trees are a random sample from the population, their effect on the *RMSE* should be proportional to $1/\sqrt{\text{number of measurements}}$. Plotting the observations against the number of measurements showed that this assumption holds rather well in the data and transformations of this form were used.

First, a main effects model including only significant predictors (using 1% level of significance) was fitted. Plotting the residuals in different groups of fixed predictors showed that the stand variables had significant interactions with the amounts of measurements. In addition, the number of old and new height measurements interacted. The cross products were included in the model and the non-significant ones were dropped one by one using backward elimination until all terms of the model were significant. The estimated models are in Table 2 and Figures 3 and 4 show predictions of the models using different measurement strategies. In the estimation of the model for *RMSE* of total volume (RV_{total}), all 170 stands were used. Since the volume of saw timber was often zero in stands with a small *DGM*, the models for *RMSE* of saw timber volume ($RV_{saw\ timber}$) used only stands with a *DGM* of more than 13 cm. To provide an idea of the total variation in the data, a variance component model with the intercept as the only fixed predictor was fitted to the data. The estimated variance components of this model are shown in the last three lines of Table 2. However, note that the total variation of single realizations is greater than the variance component because the observations of the data are based on ten realizations.

The estimated coefficients (Table 2) show that RV_{total} and $RV_{saw\ timber}$ decrease with increasing amounts of measurements and decreasing basal area and *DGM*. The old height measurement decreases the RV_{total}, except for situations where the number of new height measurements is high and the basal area is low, and it decreases the $RV_{saw\ timber}$ if the number of new sample trees is three or less. These trends are realistic and in accordance with our prior hypotheses. The consistency of the model coefficients was considered by plotting the model predictions against *DGM* and basal area using different values of other predictors. The plots showed logical behavior with different values of stand variables and stand measurements. Examples of these plots are shown in Figures 3 and 4.

In addition to basal area and *DGM*, the most important factors affecting RV_{total} were the number of G-plots and H-trees (Figure 3). The number of old H-trees had a slight effect on the predicted RV_{total} but the numbers of Q-trees and N-plots were not significant predictors at all. The number of N-plots did not affect $RV_{saw\ timber}$ significantly either. However, the number of Q-trees was one of the most important predictors in the model of $RV_{saw\ timber}$, along with the numbers of G-plots and H-trees (Figure 4). Note that when stand *DGM* is small, the measurement of one Q-tree may decrease the $RV_{saw\ timber}$ even more than the measurement of seven G-plots. The effect of an old H-tree was, again, significant but slight. The main result with respect to the use of an old height sample tree is, however, that it reduces RV_{total} (except for stands with low basal area and some new H-trees) and $RV_{saw\ timber}$ (if the number of new sample trees is three or less). Thus, by using old height sample trees, an improvement on the predictions of total and saw timber volume can be obtained with little effort.

Table 2. Models of the root mean squared errors of volume (left) and saw timber volume (right).

	RV_{total}	$RV_{sawtimber}$
Fixed part		
Intercept	-0.2195 (0.3805)	-0.5369 (0.6725)
$1/\sqrt{nG}$	-42.333 (1.5905)	-6.2779 (0.1468)
$1/\sqrt{nH+1}$	-5.5433 (0.2565)	-0.2364 (0.1581)
$1/\sqrt{nQ+1}$	-	3.5707 (0.1394)
nHo	0.7473 (0.1305)	0.3144 (0.09612)
$nHo/\sqrt{nH+1}$	-0.8203 (0.1647)	-0.6574 (0.1495)
G^2	-	0.008175 (0.001856)
$DGM/\sqrt{nG}$	1.9529 (0.0503)	-
$G/\sqrt{nG}$	0.29 (0.01683)	-
$1/(DGM*\sqrt{nG})$	334.35 (11.8928)	-
$DGM^2/\sqrt{nG}$	-	0.0373 (0.000409)
$G^2/\sqrt{nG}$	-	0.003378 (0.00035)
$DGM/\sqrt{nH+1}$	0.2932 (0.01432)	-
$G/\sqrt{nH+1}$	0.3961 (0.0119)	-
$DGM^2/\sqrt{nH+1}$	-	0.005581 (0.000355)
$G^2/\sqrt{nH+1}$	-	0.003074 (0.000305)
$DGM^2/\sqrt{nQ+1}$	-	-0.00575 (0.000356)
$G^2/\sqrt{nQ+1}$	-	0.004386 (0.000305)
*nHo*G*	-0.02214 (0.005164)	-
Random part		
var(u_k)	0.07444 (0.01359)	0.08025 (0.009348)
var(v_j)	23.0597 (2.5172)	13.561 (1.7549)
var(e_{kj})	0.9366 (0.005216)	0.7806 (0.00513)
Total variation		
var(u_k)	14.6026 (1.0674)	1.8388 (0.1375)
var(v_j)	34.2693 (3.7352)	44.0895 (5.6703)
var(e_{kj})	1.0061 (0.005593)	0.9406 (0.006179)

Note: The estimated standard errors are in parentheses. The total variation is obtained from a model with the intercept as the only fixed predictor. The notations are as follows: *nG*=number of G-plots; *nH*=number of H-trees; *nHo*=number of old H-trees (0 or 1); *nQ*=number of Q-trees; *G*=basal area, m^2/ha; *DGM*=basal area median diameter, cm.

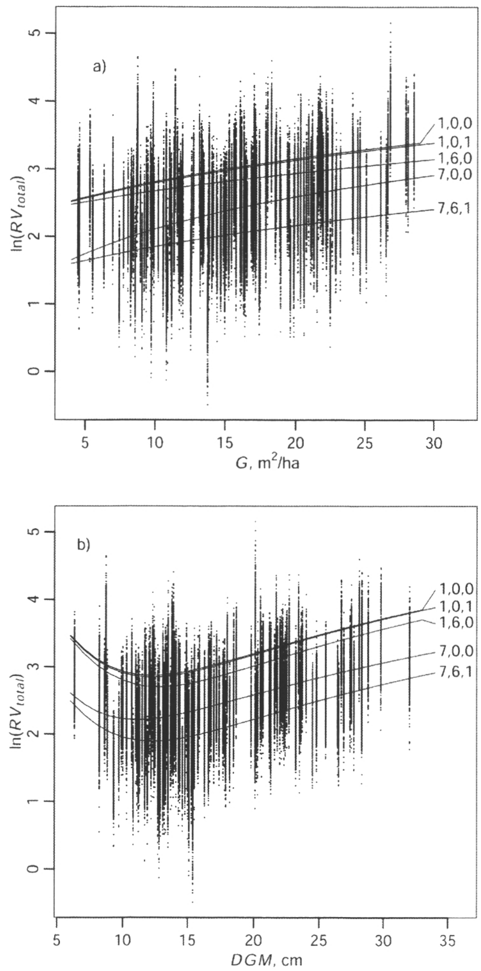


Figure 3. The observations and predictions of the logarithmic *RMSE* of total volume using different measurement strategies against basal area in a stand with a *DGM* of 17 cm (a) and against *DGM* in a stand with a basal area of 16 m2/ha (b). The numbers in the legends indicate the number of G-plots, H-trees and old H-trees, respectively. The logarithmic scale is used to make the plot more legible.

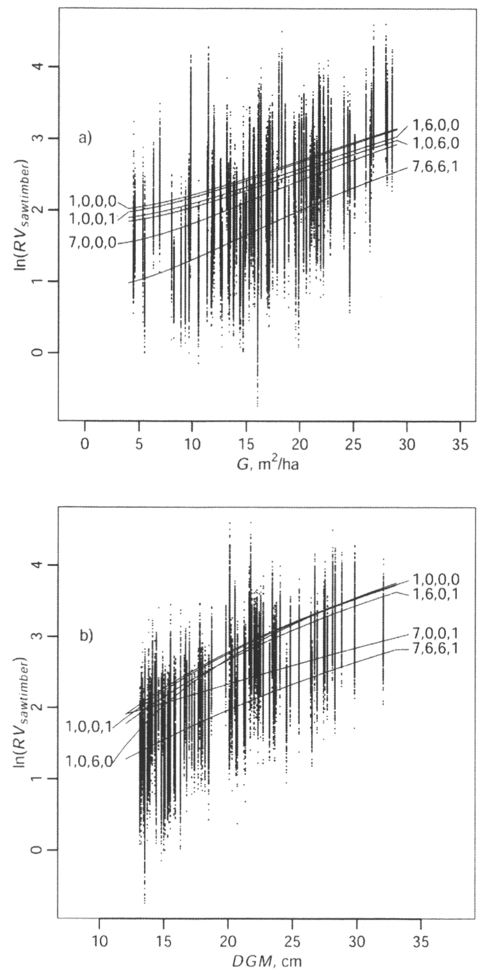


Figure 4. The observations and predictions of the logarithmic *RMSE* of saw timber volume using different measurement strategies against basal area in a stand with a *DGM* of 17 cm (a) and against *DGM* in a stand with a basal area of 16 m2/ha (b). The numbers in the legends indicate the number of G-plots, H-trees, Q-trees and old H-trees, respectively. The logarithmic scale is used to make the plot more legible.

Figures 3 and 4 show that approximately the same accuracy can be reached with various combinations. For example, in a stand where *DGM* is around 20 cm and basal area is 16 m^2/ha, one G-plot and six H-trees lead to approximately as accurate a prediction of saw timber volume as one G-plot and six Q-trees (Figure 4b). In addition, these figures show that the interaction terms have a great effect on the models. The interaction results in the predicted accuracies of different strategies crossing each other when plotted against basal area and *DGM*. This phenomenon is especially strong when the predictions of $RV_{saw\ timber}$ are plotted against *DGM* (Figure 3b) and it means that the measurement strategy of a stand should depend on the stand characteristics. For example, if the *DGM* is small, substituting the measurement of Q-trees for G-plots could reduce the error of predicted saw timber volume. The estimated models make it possible to define optimization problems, where the expected sampling error of the stand description is minimized subject to given budget constraints, as shown in the next section.

Utilizing the models

Definition of the optimization problem

This section demonstrates how the estimated models (Table 2) can be used to optimize data collection in a forest inventory. In the optimization, the accuracy of the obtained stand description is maximized (i.e. the *RMSE* is minimized) with respect to time constraints.

Define the measurement strategy as a vector having 3 elements, $\mathbf{s}'=(s_1,s_2,s_3)$, where the elements s_1, s_2 and s_3 are the numbers of G-plots, H-trees and Q-trees measured from the target stand, respectively. The state variables are in vector $\mathbf{r}$, which includes the stand characteristics (*DGM* and basal area) and the information about the old height measurement. The number of old H-trees is in vector $\mathbf{r}$ because the decision whether to measure it or not is not made during the inventory; it either is available or not and is always used when available. Furthermore, define a cost vector $\mathbf{c}'=(c_1,c_2,c_3)$, where c_1, c_2 and c_3 are the times required for measurement of one G-plot, H-tree and Q-tree, respectively. The maximum time used for measurement of the target stand is defined by the constant $c_{\max}$. The optimization problem is defined as follows:

$$\min z=y(\mathbf{s},\mathbf{r}) \qquad (12a)$$

subject to

$$\mathbf{s}'\mathbf{c}-c_{\max}\leq 0 \qquad (12b)$$
$$\mathbf{s}'-(1,0,0)\geq \mathbf{0}. \qquad (12c)$$

To carry out the optimization, the state vector $\mathbf{r}$ needs to be known. Thus, before carrying out the optimization, one G-plot needs to be measured in order to estimate the *DGM* and basal area of the stand. Therefore, the minimum number of G-plots in the optimization problem is 1. If additional G-plots are measured during the inventory, the optimization is carried out again using the improved estimate of the state vector $\mathbf{r}$ based on all G-plot measurements available.

The unit for the time can be, for example, hours or minutes. However, a more convenient approach is to define it as a function of the time needed for some basic measurements, for example, as the time needed for the measurement of one G-plot. The time requirements of other measurements are then defined as proportions of the basic measurement. Furthermore, the time requirement and maximum available time may also be a function of the state vector $\mathbf{r}$. For example, the measurement time of a G-plot and Q-tree may depend on the basal area of the stand and more time can be spent inventorying a mature stand than a young one.

Application examples

The optimization problem (Equations 12a-12c) was solved in order to find an optimal measurement combination in four hypothetical stands with different *DGM*s and basal areas. No old H-tree was assumed to be available. The optimization was carried out both with respect to the total volume and the saw timber volume. The cost vector was defined as $\mathbf{c}'=(1,0.3,0.5)$, i.e. the measurement of an H-tree takes 30% and the measurement of a Q-tree takes 50% of the time needed to measure one G-plot. The time available for measurements was the time of 5 G-plots, i.e. $c_{\max}$ was 5. Tables 3-5 show real number solutions of the optimization problems but near-optimal integer solutions can be obtained by rounding the figures of the real solution to the nearest integers. Another possibility would have been to use the complete enumeration of integer solutions, which is a realistic alternative because the number of possible integer solutions is rather small.

When the *RMSE* of total volume is minimized, the measurement resources are distributed between measurements of G-plots and H-trees, but when *RMSE* of saw timber volume is minimized, the resources are distributed between G-plots, H-trees and Q-trees (Table 3). The values of stand characteristics have a strong effect on how the resources should be distributed between different measurements. In particular, a slight growth in *DGM* causes remarkable changes in the optimal combination.

In order to show an example where the cost vector and maximum measurement time depend on the state vector, the optimization was carried out again using a cost vector of

$$\mathbf{c'}=(0.05\times G,\ 0.1+0.01\times DGM,\ 0.033\times G) \quad (13)$$

and $c_{\max}$ of $0.25\times G$, where G is the basal area of the stand. The time unit is now the time required for the measurement of one G-plot in a stand with a basal area of 20 m^2/ha. The solution differs from that in Table 3, but the trends are quite similar in both solutions. However, no conclusions about the optimal combination should be drawn on the basis of these results because of the ad hoc definitions of the cost vectors.

If both total volume and saw timber volume should be predicted accurately, both variables need to be taken into account in searching for a single solution. In this case, the objective function can be defined as a weighted average of the RV_{total} and $RV_{saw\ timber}$ (Table 5).

Table 3. The solution of the optimization problem (Equations 12a-12c) using costs determined by vector $\mathbf{c'}$=(1,0.3,0.5) and a $c_{\max}$ of 5 in the four hypothetical stands.

target variable	*DGM*	*G*	s_1	s_2	s_3	$y(\mathbf{s},\mathbf{r})$
RV_{total}	15	15	3.97	3.42	0	8.88
$RV_{saw\ timber}$	15	15	2.48	2.92	3.29	5.57
RV_{total}	25	15	4.02	3.27	0	15.6
$RV_{saw\ timber}$	25	15	4.25	2.5	0	13.01
RV_{total}	15	25	3.68	4.39	0	12.15
$RV_{saw\ timber}$	15	25	2.4	3.2	3.27	11.16
RV_{total}	25	25	3.84	3.88	0	18.92
$RV_{saw\ timber}$	25	25	3.88	2.62	0.67	19.12

Table 4. The solution of the optimization problem (Equations 12a-12c) using costs determined by Equation (13) and a $c_{\max}$ of $0.25\times G$ in the four hypothetical stands.

target variable	*DGM*	*G*	s_1	s_2	s_3	$y(\mathbf{s},\mathbf{r})$
RV_{total}	15	15	3.96	3.11	0	8.98
$RV_{saw\ timber}$	15	15	2.46	2.61	2.46	5.79
RV_{total}	25	15	4.01	2.11	0	16.25
$RV_{saw\ timber}$	25	15	4.26	1.59	0	13.34
RV_{total}	15	25	3.76	6.21	0	11.55
$RV_{saw\ timber}$	15	25	2.45	4.61	3.94	10.77
RV_{total}	25	25	3.85	4.12	0	18.77
$RV_{saw\ timber}$	25	25	3.86	2.76	0.91	18.96

Table 5. The solution of the optimization problem (Equations 12a-12c) using the objective function $z=0.2\times RV_{total}+0.8\times RV_{saw\ timber}$, costs determined by Equation (13) and a $c_{\max}$ of $0.25\times G$ in the four hypothetical stands.

DGM	*G*	s_1	s_2	s_3	$y(\mathbf{s},\mathbf{r})$
15	15	3.03	2.78	1.53	6.72
25	15	4.19	1.74	0	13.93
15	25	2.88	5.07	2.69	11.31
25	25	3.88	3.17	0.57	19.05

Discussion

The aim of this study was to predict the error of the saw timber volume and the total volume in an inventory where different numbers of G-plots, N-plots, H-trees and Q-trees are measured in a stand. The study showed that the error depends on stand variables and on the measurement strategy used. Furthermore, the amounts of measurements showed considerable interaction with the stand variables. This indicates that the most accurate stand description is obtained with different measurement strategies in different stands. The study also demonstrated how the optimal strategy for each stand can be found as a solution of an optimization problem where the expected error is minimized with respect to budget constraints. Instead of using variances, the numbers of single measurements were used as the accuracy measure of stand measurements. Thus, the solution of the optimization problem includes concrete numbers of sample plots and sample trees, making the results convenient and easily applicable in practice.

This study used the RMSE of total and saw timber volume as measures of the accuracy of stand description. Depending on the aim of the inventory, other measures can be used as well. These measures can be any variables that depend on the stand description. For example, goodness of fit statistics of diameter distribution or root mean squared errors of stem number, stand growth or dominant height could be used.

An alternative formulation of the optimization problem would have been to minimize costs with respect to some accuracy requirement. However, the solution of this problem would have required nonlinearly constrained optimization whereas the problem formulation used (Equations 12a-12c) could be solved with linearly constrained optimization.

The number of Q-trees had a very strong effect on the *RMSE* of saw timber volume in stands with a small *DGM* but in stands with a large *DGM* the effect was remarkably smaller. The explanation of this is that the Q-trees affect the predicted diameter distribution but not the basal area. The accuracy of total volume estimate depends much more on the accuracy of basal area and *DGM* than on the accuracy of the diameter distribution. For example in the tests carried out by Kangas and Maltamo (2000b, 2003), the accuracy of volume varied very little between the different distribution models. Even adjusting the distribution according to several different stand variables had no marked effect. The accuracy of the saw timber volume estimate, on the other hand, depends much more on the accuracy of the diameter distribution, especially in stands where only part of the trees are saw timber trees. Thus, the improved predictions of the diameter distribution do not have a great effect on the error of predicted volume and saw timber volume in stands where the *DGM* is large, but in stands with a smaller *DGM* the improved prediction of diameter distribution clearly decreases the error of saw timber volume.

The estimated models predict the error of the total and saw timber volumes in an inventory where the locations of sample plots are random and sample plot and tree measurements are accurate. If the measurements include, in addition to the sampling errors, visual assessment errors and measurement errors, the model predictions are downward biased. However, if the relations between sampling errors of different variables are approximately equal to the relations between total errors, the measurement errors may not have a strong effect on the optimal measurement strategy. In this study, the means of the within-stand standard deviations of basal area, *DGM* and stem number were 3.27, 1.34 and 408, respectively. In the study of Kangas et al. (2002), the total standard errors of these variables (including visual assessment errors, sampling errors and measurement errors) were 2.73, 3.67 and 422, respectively and other studies have reported the total errors of basal area and *DGM* to be between 2.8-5.5 and 2.3-2.6, respectively (Mähönen 1984, Laasasenaho and Päivinen 1986, Pussinen 1992, Pigg 1994, Haara and Korhonen 2004). Comparisons of the sampling errors of this study with the reported total errors show that the proportion of assessment and measurement errors of the total errors is approximately the same with basal area and stem number, but with *DGM* it seems to be much higher. Therefore, the optimal measurement strategies obtained using the estimated models are probably not optimal in a partly subjective and visual inventory. However, if the proportion of measurement error of *DGM* could be decreased to the same level as the proportion of measurement error of basal area and stem number, the models of this study could produce near-optimal solutions also in a partly subjective and visual inventory based on objective location of sample plots.

The current practice in Finland is to use subjectively located sample plots. This results in biased estimates of stand variables if the person carrying out the inventory does not have a realistic view of the stand. Furthermore, the sample variance of subjectively located plot measurements underestimates the real within-stand variance. Because the bias and underestimate of variance depend mostly on the person carrying out the inventory, they are hard to estimate. The system utilized in the prediction of stand description requires unbiased estimates of stand variables and estimates of their error variances. Thus, if the inventory based on subjective location cannot be shown to be much more efficient than the inventory based on objective location, the subjective location should be abandoned.

In this study, the algorithm of Mehtätalo (2004d) was used for the first time in localizing the diameter percentiles using true trees from true sample plots. Q-trees improved the predictions of diameter distribution markedly, as seen in the effect of number of Q-trees on the *RMSE* of saw timber volume. However, the measurement of diameter and the determination of rank were assumed to have been made without error. Field measurements of Q-trees are carried out in a similar manner as the measurements of *DGM*, i.e. by assessing the rank visually and measuring the diameter accurately (in fact, the measurement of *DGM* is a special case of the Q-tree measurement). As discussed before, the measurements of *DGM* include a great deal of measurement errors and the same supposedly holds for Q-trees, too. Thus the real gain derived from Q-tree measurements is probably not as great as the model shows. The accuracy of the measurements of Q-trees should be studied, as should the usefulness of the localization with Q-trees with erroneous ranks.

The data used in this study were quite limited. In particular, there were only a few stands with large *DGM* in the data, because the radii of the circular plots were so small that an HPS plot could not be established in stands with large diameters. Therefore, the models should not be extrapolated to stands outside the range of the data. A safe approach would be to use a *DGM* of 32 cm and a basal area of 28.5 m^2/ha in optimization if the true values are higher than these. Since the numbers of measurements are included in the model in a theoretically well-grounded form, the extrapolation of the number of measurements should not be as risky as the extrapolation of *DGM* and basal area. However, solutions where the number of G-plots is more than 7 or the number of Q- or H-trees is more than 6 should not be accepted unreservedly.

Acknowledgements

This work was funded by the Academy of Finland (decision numbers 73392 and 200775). The authors wish to thank Professor Juha Alho, Dr Juha Lappi, Professor Matti Maltamo and the group of researchers on forest assessment at the University of Joensuu for comments on the earlier drafts of the manuscript. We also thank Dr Lisa Lena Opas-Hänninen for revising the language of the article.

References

Borders, B.E., Souter, R.A., Bailey, R.L. and Ware, K.D. 1987. Percentile-based distributions characterize forest stand tables. For. Sci. **33**: 570-576.

Gertner, G. 1991. Prediction bias and response surface curvature. For. Sci. **37**: 755-765.

Gertner, G., Parysow, P. and Guan, B. 1996. Projection variance partitioning of a conceptual forest growth model with orthogonal polynomials. For. Sci. **42**: 474-486.

Gustafsen, H.G., Roiko-Jokela, P. and Varmola, M. 1988. Kivennäismaiden talousmetsien pysyvät (INKA ja TINKA) kokeet. Suunnitelmat, mittausmenetelmät ja aineistojen rakenteet. Finnish Forest Research Institute, Res. Pap. 292.

Haara, A. and Korhonen, K.T. 2004. Kuvioittaisen arvioinnin luotettavuudesta. Manuscript. (In Finnish)

Hines, W.W, and Montgomery, D.C. 1980. Probability and statistics in engineering and management science. John Wiley & Sons. New York.

Husch, B., Miller, C.I. and Beer, T.W. 1982. Forest mensuration, 3rd edition. John Wiley & Sons, New York.

IMSL 1997. Fortran subroutines for mathematical applications. Math/Library. Vols 1 and 2. Visual Numerics.

Johnson, N.L. and Kotz, S. 1972. Distributions in statistics: continuous multivariate distributions. John Wiley & Sons. New York.

Kangas, A. 1997. On the prediction bias and variance in long-term growth projections. For. Ecol. Manage. **96**: 207-216.

Kangas, A. 1999. Methods for assessing uncertainty of growth and yield predictions. Can. J. For. Res. **29**: 1357-1364.

Kangas, A., Heikkinen, E. and Maltamo, M. 2002. Puustotunnusten maastoarvioinnin luotettavuus ja ajanmenekki. Metsätieteen aikakauskirja 3/2002: 425-440. (In Finnish.)

Kangas, A., Heikkinen, E. and Maltamo, M. 2004. Accuracy of partially visually assessed stand characteristics – A case study of Finnish inventory by compartments. Can. J. For. Res. **34**: 916-930.

Kangas, A. and Maltamo, M. 2000a. Calibrating predicted diameter distribution with additional information. For. Sci. **46**: 390-396.

Kangas, A. and Maltamo, M. 2000b. Percentile-based basal area diameter distribution models for Scots pine, Norway spruce and birch species. Silva Fenn. **34**: 371-380.

Kangas, A. and Maltamo, M. 2000c. Performance of percentile based diameter distribution prediction and Weibull method in independent data sets. Silva Fenn. **34**: 381-398.

Kangas, A and Maltamo, M. 2002. Anticipating the variance of predicted stand volume and timber assortments with respect to stand characteristics and field measurements. Silva Fenn. **36**: 799-811.

Kangas, A. and Maltamo, M. 2003. Calibrating predicted diameter distribution with additional information in growth and yield predictions. Can. J. For. Res. **33**: 430-434.

Khuri, A. I. and Cornell, J.A. 1987. Response surfaces, designs and analyses. Marcel Dekker, Inc. New York.

Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Comm. Inst. For. Fenn. **108**: 1-74.

Laasasenaho, J. and Päivinen, R. 1986. On the checking of an inventory by compartments. Folia Forestalia 664. (In Finnish with English summary)

Lappi, J. 1991. Calibration of height and volume equations with random parameters. For. Sci. **37**: 781-801.

Lappi, J. 1997. A longitudinal analysis of height/diameter curves. For. Sci. **43**: 555-570.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. 1996. SAS system for mixed models. SAS Institute Inc, Cary, NC.

McCulloch, C. E. and Searle, S. R. 2001. Generalized, linear and mixed models. John Wiley & Sons, New York.

Mehtätalo, L. 2004a. An algorithm for ensuring compatibility between estimated percentiles of diameter distribution and measured stand variables. For. Sci. **50**: 20-32.

Mehtätalo, L. 2004b. A longitudinal height-diameter model for norway spruce in Finland. Can. J. For. Res. **34**: 131-140.

Mehtätalo, L. 2004c. Height-diameter models for Scots pine and birch species in Finland. Submitted Manuscript (Silva Fennica).

Mehtätalo, L. 2004d. Localizing predicted diameter distribution with sample information. Revised Manuscript (Forest Science).

Mähönen, M. 1984. Kuvioittaisen arvioinnin luotettavuus. University of Helsinki, M. Sc. thesis. (In Finnish)

Pigg, J. 1994. Keskiläpimitan ja puutavaralajijakauman sekä muiden puustotunnusten tarkkuus Metsähallituksen kuvioittaisessa arvioinnissa. University of Helsinki, M. Sc. thesis. (In Finnish)

Pussinen, A. 1992. Ilmakuvat ja Landsat TM – satelliittikuva välialueiden kuvioittaisessa arvioinnissa. University of Joensuu, M. Sc. thesis. (In Finnish)

R Development Core Team 2003. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

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

Searle, S.R., Casella, G. and McCulloch, C.E. 1992. Variance components. John Wiley & Sons, New York.

APPENDIX

Expectation, variance and covariance of logarithmic measurements

Assume that we have measured the values of variables $X_1>0$ and $X_2>0$ and the unbiased measurements $\hat{X}_1$ and $\hat{X}_2$ are lognormally distributed with expectations X_1 and X_2 and variances $\text{var}(\hat{X}_1)=\sigma_1^2$, $\text{var}(\hat{X}_2)=\sigma_1^2$ and $\text{cov}(\hat{X}_1,\hat{X}_2)=\rho$. We want to know $E(\ln\hat{X}_1)$, $\text{var}(\ln\hat{X}_1)$ and $\text{cov}(\ln\hat{X}_1,\ln\hat{X}_2)$. It follows from the properties of a lognormal distribution (see, e.g., Johnson and Kotz 1972: 18-20, Hines and Montgomery 1980: 188-191) that the logarithmic measurements are multi-normally distributed with the expectations, variances and covariances of

$$E(\ln\hat{X}_1)=2\ln X_1-\frac{1}{2}\ln(\sigma_1^2+X_1^2), \qquad \text{(A1)}$$

$$\text{var}(\ln\hat{X}_1)=\ln(\sigma_1^2+X_1^2)-2\ln X_1 \qquad \text{(A2)}$$

and

$$\text{cov}(\ln\hat{X}_1,\ln\hat{X}_2)=\ln(\rho+X_1X_2)-\ln(X_1X_2). \qquad \text{(A3)}$$

Now assume that we have k unbiased log-normally distributed measurements in random vector $\hat{\mathbf{x}}=(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_k)'$ with the variance covariance matrix $\text{var}(\hat{\mathbf{x}})=\mathbf{C}$. Formulas (A1)-(A3) give the matrix results

$$E(\ln\hat{\mathbf{x}})=2\ln\mathbf{x}-\frac{1}{2}\ln\{[\mathbf{I}\otimes(\mathbf{C}+\mathbf{xx'})]\mathbf{1}_{k\times 1}\} \qquad \text{(A4)}$$

and

$$\text{var}(\ln\hat{\mathbf{x}})=\ln(\mathbf{C}+\mathbf{xx'})-\ln(\mathbf{xx'}). \qquad \text{(A5)}$$

Prediction error of models with erroneous logarithmic predictors

Assume a model of the form

$$\mathbf{y}_m=\mathbf{B}\ln\mathbf{x}_m+\mathbf{e}_m, \qquad \text{(A6)}$$

where $\mathbf{y}_m$ is a vector of interest variables from observation m, $\mathbf{B}$ is a matrix of parameters, $\mathbf{x}_m$ the vector of predictors and $\mathbf{e}_m$ the vector of residual errors. For example, vector $\mathbf{y}_m$ could include percentiles of a diameter distribution of stand m. Assume that the measured predictors are lognormally distributed with $E(\hat{\mathbf{x}}_m)=\mathbf{x}_m$ and $\text{var}(\hat{\mathbf{x}}_m)=\mathbf{C}$. We want to know how errors in the predictors affect the model predictions, $\tilde{\mathbf{y}}_m=\mathbf{B}\ln\hat{\mathbf{x}}_m$.

The expectation of the prediction is

$$E(\tilde{\mathbf{y}}_m)=E(\mathbf{B}\ln\hat{\mathbf{x}}_m)=\mathbf{B}E(\ln\hat{\mathbf{x}}_m). \qquad \text{(A7)}$$

Using Equation (A4) we get

$$E(\tilde{\mathbf{y}}_m)=\mathbf{B}\left(2\ln\mathbf{x}_m-\frac{1}{2}\ln\{[\mathbf{I}\otimes(\mathbf{C}+\mathbf{x}_m\mathbf{x}_m')]\mathbf{1}\}\right)$$
$$=\mathbf{B}\ln\mathbf{x}_m+\mathbf{B}\left(\ln\mathbf{x}_m-\frac{1}{2}\ln\{[\mathbf{I}\otimes(\mathbf{C}+\mathbf{x}_m\mathbf{x}_m')]\mathbf{1}\}\right). \quad \text{(A8)}$$

Thus, the predictions are biased, the last term of (A8) being the bias. The true value of the bias cannot be calculated since the expectations of the measurements are not known. However, approximate bias can be calculated by substituting the expectations of the measurements with the measured values. An approximately unbiased prediction with model (A6) using measurements (A7) is obtained by subtracting the approximate bias from the prediction:

$$\tilde{\mathbf{y}}_{unbiased}\approx\mathbf{B}\ln(\hat{\mathbf{x}}_m)-\mathbf{B}\left(\ln\hat{\mathbf{x}}_m-\frac{1}{2}\ln\{[\mathbf{I}\otimes(\mathbf{C}+\hat{\mathbf{x}}_m\hat{\mathbf{x}}_m')]\mathbf{1}\}\right). \quad \text{(A9)}$$

The variance-covariance matrix of the error term is

$$\mathrm{var}(\tilde{\mathbf{y}}_m)=\mathrm{var}(\mathbf{B}\ln\hat{\mathbf{x}}_m+\mathbf{e}_m)$$
$$=\mathbf{B}\,\mathrm{var}(\ln\hat{\mathbf{x}}_m)\mathbf{B}'+\mathbf{D}.$$

Using Equation (A5) we get

$$\mathrm{var}(\tilde{\mathbf{y}}_m)=\mathbf{B}[\ln(\mathbf{C}+\mathbf{xx}')-\ln(\mathbf{xx}')]\mathbf{B}'+\mathbf{D}$$
$$\approx\mathbf{B}[\ln(\mathbf{C}+\hat{\mathbf{x}}\hat{\mathbf{x}}')-\ln(\hat{\mathbf{x}}\hat{\mathbf{x}}')]\mathbf{B}'+\mathbf{D}. \quad \text{(A10)}$$

The first term of (A10) is the increase of the residual variance caused by the measurement errors of the predictors.

ISBN 951-40-1934-2
ISSN 0358-4283