Dissertationes Forestales 94

Estimating individual tree growth using non-parametric methods

Susanna Sironen
Faculty of Forest Sciences
University of Joensuu

Academic dissertation

To be presented with the permission of the Faculty of Forest Sciences of the University of Joensuu for public examination in Auditorium N100 of the University of Joensuu, Yliopistokatu 7, Joensuu, on 4th September 2009, at 12 o'clock noon.

Title of dissertation: Estimating individual tree growth using non-parametric methods

Author: Susanna Sironen

Dissertationes Forestales 94

Thesis Supervisors:
Prof. Annika Kangas
Department of Forest Resource Management, University of Helsinki, Finland
Prof. Matti Maltamo
Faculty of Forest Sciences, University of Joensuu, Finland

Pre-examiners:
Assistant Prof. Hailemariam Temesgen
College of Forestry, Oregon State University, USA
Dr. Jari Miina
Finnish Forest Research Institute, Joensuu Research Unit, Finland

Opponent:
Dr. Kari T. Korhonen
Finnish Forest Research Institute, Joensuu Research Unit, Finland

ISSN: 1795-7389
ISBN: 978-951-651-277-1 (PDF)

(2009)

Publishers:
Finnish Society of Forest Science
Finnish Forest Research Institute
Faculty of Agriculture and Forestry of the University of Helsinki
Faculty of Forest Sciences of the University of Joensuu

Editorial Office:
Finnish Society of Forest Science
P.O. Box 18, FI-10301 Vantaa, Finland
http://www.metla.fi/dissertationes

ABSTRACT

Sironen, S. 2009. Estimating individual tree growth using non-parametric methods. Dissertationes Forestales 94. 54 p. Available at http://www.metla.fi/dissertationes/df94.htm.

Information about the current condition, extent and quantity of forests that is provided by forest inventories, combined with forest growth models, is of great importance in forecasting the future development of forests. The ability to make reliable predictions has an important role as a tool of management planning, in evaluating silvicultural options, and ensuring sustainable forest management. In Finland, growth models are typically national models which may be markedly biased for a given stand or region. Non-parametric methods offer an alternative to the traditional regression methods. In non-parametric methods, the value of the variable of interest for a target observation is estimated often as a weighted average of the values of neighbouring reference observations, which are similar to the target observation in terms of the independent variables and weighted by their proximity to the target observation Locality can easily be described by non-parametric methods, if local data is available. The overall purpose of this thesis was to examine and evaluate different non-parametric methods as a method for individual tree growth estimation. One of the main focuses was to test non-parametric methods in order to reduce the regional biases associated in the growth estimates.

The study material comprised temporary local sample plot data from Kuusamo in north-eastern Finland and nationwide permanent inventory growth plot data (INKA). The tree species considered were Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* (L.) Karst.). The tested methods included k-nearest neighbour methods employing various distance functions and generalized additive models. The different topics analysed in this thesis include local non-parametric growth estimation methods, localizing the non-parametric growth estimates, simultaneous estimation of individual tree diameter and height increment, and the effects of correlated observations on non-parametric growth estimation methods.

The results showed that non-parametric methods are suitable for estimation of growth, although the performance of the different methods varied depending on the purpose and the data used. The non-parametric methods were capable of reducing the regional biases. The most promising alternative to the means of localization was the sub-setting of the reference data by selecting the neighbours from a circle around the target tree. The levels of accuracy achieved in the estimation of individual tree growth were at least as good as those obtained by the parametric models at the tree, stand and regional levels. The methods presented in this thesis could be implemented in practical planning systems, although several issues still require further study and consideration, especially the issues concerning silvicultural treatments.

Keywords: diameter increment, height increment, nearest neighbour, k-NN, generalized additive models, GAM

ACKNOWLEDGEMENTS

I am grateful for all the support I have received whilst researching and writing up this dissertation. First and foremost, I would like to thank my supervisors and co-authors Pof. Annika Kangas at the University of Helsinki and Prof. Matti Maltamo at the University of Joensuu. I am grateful for all the support, ideas and comments I have received from them during this long dissertation journey. Life is what happens when you are doing your dissertation, and they have been very patient and understanding.

In that same vein, I want to thank my other co-writer Prof. Jyrki Kangas at Metsähallitus, and Jouni Kalliovirta at Simosol Oy for making the calculations within the SIMO system. I wish to express my appreciation and thanks to the pre-examiners of this thesis, Dr. Jari Miina at Finnish Forest Research Institute and Prof. Hailemariam Temesgen at the Oregon State University, for their valuable comments. I also wish to thank all the language checkers and anonymous reviewers of the separate articles.

This work was carried out at the University of Joensuu, Faculty of Forest Sciences, and was mainly funded through the Graduate School in Forest Sciences. Financial support provided by Niemi Foundation (Niemi-säätiö), Metsämiesten Säätiö, and The Finnish Society of Forest Science (Suomen Metsätieteellinen Seura) are gratefully acknowledged. I have been lucky to have two comprehensive datasets for my studies. I would like to express my gratitude to Finnish Forest Research Institute for providing me the INKA data. Moreover, I would like to thank people at the Forestry Centre North Ostrobothnia (Metsäkeskus Pohjois-Pohjanmaa) in Kuusamo and people at the Kuusamo Common Forest (Kuusamon Yhteismetsä) for providing and collecting the Kuusamo data.

Although the Faculty of Forest Sciences has provided me great working facilities, it is the people who make the environment. I would like to thank all my colleagues, who have created an enjoyable working atmosphere. I owe special thaks to Annukka, who has been of great support and joy in our otherwise so “manly” research group. Moreover, I would like to thank all the people near me who have given me their support during the years.

I have been fortunate to have many special friends who have been irreplaceable for many years. Thanks especially go to Katja, Kirsi and Sanna - you have shared my moments of success, listened to my moaning, and supported me every step of the way.

Finally, I give my warmest thanks to my family. To my mother, who has always been there for me and given her unfailing support. To Jani, who besides everything else has given me invaluable help with my numerous programming problems. My heartfelt thanks go to Samuli and Eemeli for their patience; so many times have I heard the words “Mom, do you always have to work?” Although you may not understand it now, since you’d rather have me playing with you all the time, you’re the ones I’ve made this for. The acknowledgements would not be complete without thanks to my furry family members; Reetu and dearly beloved and forever missed Sokrates. You have taken care of my daily exercise and helped me to forget the work issues during the walks in the woods.

Joensuu, July 23

LIST OF ORIGINAL ARTICLES

This thesis is based on the following papers, which are referred to in the text by the Roman numerals I-VI.

I Sironen, S., Kangas, A., Maltamo, M. and Kangas, J. 2001. Estimating individual tree growth with the k-Nearest Neighbour and k-Most Similar Neighbour methods. Silva Fennica 35: 453-467. http://www.metla.fi/silvafennica/full/sf35/sf354453.pdf

II Sironen, S., Kangas, A., Maltamo, M.and Kangas, J. 2003. Estimating individual tree growth with nonparametric methods. Canadian Journal of Forest Research 33: 444-449. http://www.ingentaconnect.com/content/nrc/cjfr/2003/00000033/00000003/art00010

III Sironen, S., Kangas, A., Maltamo M. and J. Kalliovirta, J. 2008. Localization of growth estimates using non-parametric imputation methods. Forest Ecology and Management 256: 674-684. doi:10.1016/j.foreco.2008.05.013

IV Sironen, S., Kangas, A., Maltamo M. 2009. Predicting tree- and stand-level growth using simultaneous k-Nearest Neighbour imputation for diameter and height increment. Submitted manuscript.

V Sironen, S., Kangas, A., Maltamo M. 2009. Effect of reference data selection on the accuracy of non-parametric k-NN imputation for estimating individual tree growth. Submitted manuscript.

VI Sironen, S., Kangas, A., Maltamo M. 2009. Comparison of different non-parametric growth imputation methods in the presence of correlated observations. Revised manuscript.

Articles I-III are reprinted with permission.

In regard to the entire thesis, most of the work done in these papers was carried out by Susanna Sironen. The co-authors, Professor Annika Kangas and Professor Matti Maltamo (papers I–VI), participated in the planning of the papers and in the writing of the papers by commenting on the text and thereby improving the manuscripts. In papers I and II, Professor Jyrki Kangas contributed especially to planning the study and the fieldwork. In paper III, Jouni Kalliovirta calculated the parametric growth predictions within the SIMO system.

TABLE OF CONTENTS

ABSTRACT 3
ACKNOWLEDGEMENTS 4
LIST OF ORIGINAL ARTICLES 5
1 INTRODUCTION 7
1.1 Modelling growth 7
1.2 Non-parametric methods 9
1.3 Objectives of the thesis 12
2 STATISTICAL BACKGROUND OF NON-PARAMETRIC METHODS 13
2.1 Common features of non-parametric methods 13
2.2 Nearest neighbour methods 14
2.3 Generalized additive models 16
3 STUDY MATERIAL 18
3.1 Kuusamo data 19
3.2 INKA data 20
4 METHODS 22
4.1 Local non-parametric growth estimates (I and II) 22
4.2 Localization of the non-parametric growth estimates (III) 23
4.3 Predicting tree- and stand-level growth using simultaneous k-Nearest Neighbour imputation for diameter and height increment (IV) 24
4.4 Effect of reference data selection on the accuracy of non-parametric k-NN imputation for individual tree growth (V) 24
4.5 Comparison of different non-parametric growth imputation methods in the presence of correlated observations (VI) 25
4.6 Evaluation criteria 26
4.7 Comparisons to parametric growth predictions 26
5 RESULTS 27
5.1 Variable selection and size of the neighbourhood (I-VI) 27
5.2 Local non-parametric growth estimates (I and II) 28
5.3 Localization of the non-parametric growth estimates (III) 29
5.4 Simultaneous estimation of individual tree diameter and height increment (IV) 33
5.5 Effects of reference data selection on the accuracy of non-parametric growth estimation (V) 34
5.6 Comparison of different non-parametric growth imputation methods in the presence of correlated observations (VI) 36
5.7 Performance of the non-parametric methods under different growing conditions (I-VI) 37
6 DISCUSSION 39
6.1 Different non-parametric methods 39
6.2 Dependent and independent variables 41
6.3 Effects of correlated observations 42
6.4 Localization of the non-parametric growth estimates 43
6.5 Concluding remarks and need for future research 45
REFERENCES 48

1 INTRODUCTION

1.1 Modelling growth

In long-term management of forests, planning decisions concerning optimal silvicultural regimes are needed, and this requires a large amount of information on both the current and future condition, extent, and quantity of forests. In order to make informed decisions for using and managing forest resources, accurate growth estimates are very important, since forestry databases are normally projected using growth models. Growth models may be classified according to many criteria based on different characteristics of the models. In general, there are two ways to develop growth and yield models; methods based on the physiological mechanism of plant growth and empirical methods. The former are developed for understanding, and are difficult to apply practically due to the constraints of complicated environmental factors (e.g. Vanclay 1994). The latter are developed for prediction and may sacrifice specific detail of growth processes in order to achieve efficiency and accuracy in providing information for management planning (e.g., Vanclay 1994).

Traditionally empirical growth and yield models have been classified into whole stand models, diameter class models and individual tree models based on the level of detail they provide. Whole stand models predict growth for the entire stand and require stand-level information, such as basal area of the stand, stand age and site type (e.g. Clutter 1963, Sullivan and Clutter 1972, Nyyssönen and Mielikäinen 1978, Pienaar and Harrison 1989, Ochi and Cao 2003, Huuskonen and Miina 2007, Martínez Pastur et al. 2008). Whole stand models are easy to use in practical applications. They are easy to control and analyse, therefore their applicability is more easily distinguished (e.g. Gustavsen 1998). Inventory costs concerning the input data are low compared with individual tree growth models. Although the accuracy of whole stand models has been as good as that of individual tree models, they may not be reliable in mixed stands or suitable for uneven-aged stands (e.g., Gustavsen 1998, Hasenauer 2006, Mäkinen et al. 2008). Diameter class models are developed to predict stand growth rates by diameter classes and can be seen as a compromise between whole stand models and individual tree models (e.g. Adams and Ek 1974, Ek 1974, Solomon et al. 1995, Eerikäinen and Maltamo 2003). These models have some information regarding the structure of the stand and there are several techniques to estimate this structure, one of the most widely used being stand table projection, which produces histograms of stem diameters (e.g. Vanclay 1994). Furthermore, transition matrix models can be built for different entities as basic units (Buongiorno and Michie 1980). The largest group of transition matrix models are constructed with a single tree as the basic entity and diameter as the state-defining variable (e.g. Usher 1969, Haight and Getz 1987). Area transition matrix models instead apply forest area as the basic unit and the state-defining variables are related to area (e.g. Hool 1966, Sallnäs 1990).

Individual tree models predict the development of each individual tree in a stand and are usually based on a combination of tree- and stand-level information (e.g. Hynynen 1995, Monserud and Sterba 1996, Andreassen and Tomter 2002, Hynynen et al. 2002, Zhao et al. 2003). The individual tree growth model enables, in principle, predictions to be made regardless of species mixture, age distribution or applicable silvicultural system (Hasenauer 2006). However, the suitability also depends on the specific models, so that in practice the suitability is not quite so straightforward. Predictions can attain more accuracy, if tree-wise

input data is available instead of trees generated using a predicted distribution (e.g. Gustavsen 1998). However, forest growth simulators applying individual tree growth models may modify uneven-aged stands to make them resemble even-aged stands by removing small trees. Individual tree models may be further divided into distance-independent and distance-dependent models, the latter requiring mapped or simulated tree locations for incorporating competition information (Munro 1974).

Furthermore, two different modelling approaches for predicting individual tree diameter and height increment have been commonly used. One is the growth-potential independent approach, in which actual diameter or height increment rates are directly expressed as a function of tree and stand characteristics from the available data, including the competitiveness of the tree in the stand (Wykoff et al. 1982, Wykoff 1990, Monserud and Sterba 1996). The other is the growth-potential dependent approach, which assumes a species and site specific upper limit or limited growth (e.g. Hasenauer 2006). Models based on this approach must first select a function that defines the potential diameter or height growth of competition-free trees, then a modifier function is used to reduce this potential for each tree within a stand, according to the competition situation of the tree (Ek and Monserud 1974, Daniels and Burkhart 1975, Pretzsch et al. 2002). Base potential for diameter increment functions may be defined by the dimensions of open-grown trees, which may be considered as the potential dimensions that a tree may have (e.g. Hasenauer 1997). The potential height increment is often defined by mean dominant height increment derived from site index functions. The modifier that predicts the tree growth in relation to the potential growth is commonly expressed as a function of the individual tree's attributes, such as crown ratio, crown length and competition indices (e.g. Hynynen et al. 2002, Hasenauer 2006).

Primary applications for the information provided by growth models include inventory updating (Burkhart 1992), and when growth models are used for that purpose the time horizon in growth prediction is usually a few years (Hynynen 1995). Growth models are also used in long-term simulations. Long-term forecasts covering some decades are needed, for example, in forest management planning and in making decisions on forest policy strategies (Hynynen 1995). While considering long-term forecasts, the method applied should be capable of reliably predicting the effects of silvicultural treatments and practices applied today and in the future (e.g. Hynynen 1995). Furthermore, predictions of the impacts of changing climate somehow need to represent the key biological processes and take into account the effect of climatic and edaphic factors on the physiological process behind the growth of trees (Matala et al. 2006, Nuutinen et al. 2006).

Most growth models are constructed from several equations independently fitted to the data (Vanclay and Skovsgaard 1997). However, implicit assumption of independence does not apply from a biological point of view, and is inefficient from a statistical point of view (e.g. Borders and Bailey 1986, Hasenauer et al. 1998). Since cross-equation correlations and feedback mechanisms might exist among variables that are used to describe growth relationships, forest stand dynamics should be described by simultaneous systems rather than separate individual equations (e.g. Daniels and Burkhart 1988, Huang and Titus 1999). Simultaneous estimations of all model components minimizes overall model errors and provides a variance-covariance matrix for the model as a whole (Vanclay and Skovsgaard 1997). A considerable number of simultaneous stand-level equations based on mean and total stand characteristics have been constructed (e.g. Furnival and Wilson 1971, Borders and Bailey, 1986, Fang et al. 2001, Eerikäinen 2002), but few simultaneous individual tree models (Hasenauer et al. 1998, Huang and Titus 1999).

Models used for growth prediction are typically on a large scale. While these models may give accurate results for larger areas, they may be markedly biased for a given stand or region. Gustavsen (1998) found notable differences across the Forestry Centre regions in Finland when assessing the performance of national growth models; regional biases for five-year volume growth were as large as 7.97 m^3 ha^{-1} in North Karelia or 5.93 m^3 ha^{-1} in north-eastern Finland, whereas it was as low as -0.44 m^3 ha^{-1} in northern Häme. The accuracy of forestry databases can be effectively improved by accounting for regional differences in the associated growth predictions. This could be done by fitting local growth models, but this may be overly costly in practice. Another solution would be calibration of the national models for a particular area (Talvitie 2005). For example, linear prediction theory can be used to calibrate mixed models for a given stand (e.g. Lappi 1986, 1991, Kangas and Korhonen 1995). However, this approach is not suitable for regional adjustment unless an expected value of stand effect in a region can be estimated. Räty and Kangas (2007, 2008) tested selecting the localizing areas for general models based on local indices of spatial association and the method seemed to be useful and the localization removed the local bias associated in the global model.

Growth and yield models have been developed for many purposes and the choice of one from among the different approaches may simply be a matter of preference or convenience, because each approach can produce acceptable predictions if appropriately used considering the model's own function, the requirements of the application in question and the data available. Different kinds of problems may require different kinds of solutions and model types (Daniels 1993). In general, different model types should be applied to the kind of data from which the models are constructed in order to obtain reliable predictions (Gustavsen 1998). Currently in Finland, the growth models used in practice are usually individual tree growth models constructed with parametric regression technique, in which growth of trees is predicted with distance-independent models for tree basal-area growth and height growth (Hynynen et al. 2002). The former are growth-potential independent models and the latter are driven by the height development of dominant trees. An alternative to traditional regression models is to construct estimates applying different non-parametric methods.

1.2 Non-parametric methods

Intensive efforts have been devoted to non-parametric methods over the past few decades (Fan 2000). The progress in the field of non-parametric methods has been dynamic. Non-parametric methods allow data to search appropriate nonlinear forms that best describe the available data, and as non-parametric methods make fewer assumptions, their applicability is much wider than that of the corresponding parametric methods (Fan 2000). In non-parametric methods, the estimate for the target observation is a local estimate, for example, a local mean, of the values of neighbouring reference observations, each value weighted by its proximity to the target observation in the space of dependent variables (e.g. Härdle 1989, Altman 1992). Reference observations form a group of potential nearest neighbours and have information on both dependent variables and independent variables, while target observations have information only on independent variables. The neighbourhood is defined by these independent variables which are known in both data (e.g. Härdle 1989, Altman 1992, Korhonen and Kangas 1997).

The non-parametric approach subsumes many methods and variations on methods. Many useful techniques have been proposed for univariate smoothing, including kernel, local polynomials and splines (e.g. Rosenblatt 1956, Stone 1977). Univariate smoothing techniques can be extended in a straightforward manner to situations including many independent variables. However, such extensions are not useful due to the "curse of dimensionality", which means that when the number of independent variables increases, the solution of the estimation tasks rapidly becomes more complex (e.g. Bellman 1961, Friedman 1994). Many techniques have been proposed to overcome this problem, including generalized additive models (GAMs) (Hastie and Tibshirani 1986), which are broadly applied in ecological studies (Guisan et al. 2002). Forestry applications of GAMs include modelling forest characteristics and model-assisted estimation of forest resources (Frescino et al. 2001, Moisen and Frescino 2002, Opsomer et al. 2007, Zhang et al. 2008). In addition, nearest neighbour methods are suitable for multivariate settings. Nearest neighbour methods are used in numerous forestry applications, including generalization of sample tree information, estimation of diameter distribution, in remote sensing applications and many multivariate and multisource forest inventory applications (see Moeur and Stage 1995, Korhonen and Kangas 1997, Maltamo and Kangas 1998, Moeur and Hershey 1999, Holmström et al. 2001, Packalén and Maltamo 2007, LeMay et al. 2008, Temesgen et al. 2003, 2008). Beyond this thesis, non-parametric methods have been applied to estimation of stand-level volume growth for *Pinus kesiya* plantations by Maltamo and Eerikäinen (2001). In addition, Neurogenetic Algorithm System, an artificial neural network with genetic algorithm has been applied in individual tree growth modelling (Liao et al. 1998). Artificial neural networks are computer models that attempt to mimic the way in which the human brain performs a particular task. They are non-parametric methods; they do not assume any particular noise process and can learn linear and nonlinear processes directly from the data. In addition, random forests are among the recent additions to the non-parametric statistics and machine learning methods (e.g. Breiman 2001). Random forests can be used both for regression and classification, and they have shown to be effective in practical applications and their generalization properties are good.

Non-parametric methods have certain advantages over the traditional regression methods. Firstly, the model structure in non-parametric methods is not specified a priori. The non-parametric methods do not rely on any probability distribution or require any predefined information on the form of underlying function, thus they are very flexible (e.g. Härdle 1989). Secondly, non-parametric methods can retain more of the variance structure of the data. However, this cannot be guaranteed, if more than one neighbour is used in the estimation. Furthermore, to retain the full variation of the data, the occasions on which the same reference observation is imputed have to be restricted. Barth and Ståhl (2007) restricted the ordinary imputation in order that the method preserved the composition of the original data at the landscape level. This was obtained by imputing each observation in reference data into the target observations for a limited number of times. Each observation was represented in the reference data as many times as it should be found in the population level. Non-parametric estimates are formed from existing measured samples, hence the estimates are always within the bounds of biological reality and unrealistic growth estimates cannot occur (e.g. Moeur and Stage 1995). In certain applications; however, the non-parametric methods may produce combinations that do not exist within the realm of real values, if more than one nearest neighbour is averaged (e.g. LeMay and Temesgen 2005). In addition, careful attention must be paid if the data is not continuous, since non-

parametric methods may interpolate to areas not allowable in these kinds of situations (Maltamo and Eerikäinen 2001)

Some of the non-parametric methods are multivariate and make it possible to estimate many variables of interest simultaneously (Moeur and Stage 1995, Katila and Tomppo 2002, Packalén and Maltamo 2007). These applications consider mainly simultaneous estimations of stand-level characteristics, but the method can similarly be applied at individual tree level for many characteristics. Furthermore, non-parametric methods can effectively describe local conditions, if sufficient local data is available. Localization in the case of nearest neighbour estimation has been investigated in numerous studies by using different approaches (e.g. Tokola 2000, Maltamo et al. 2003, Koistinen et al. 2008). Localization of the non-parametric estimates may be obtained through a variety of methods, in particular, by including both variable-space and physical-space in the imputations. Previously, Katila and Tomppo (2001) have studied the inclusion of moving geographical reference areas both in horizontal and vertical directions in the nearest neighbour method.

Non-parametric methods need reference data in the database in the application phase as well; and such reference data should be of good quality and cover the whole range of possible values of the dependent variables (Moeur and Stage 1995). This also requires permission to use the reference database at the application phase. However, databases and estimates produced by non-parametric methods are easy to maintain and update when it is necessary to add or remove data (Maltamo and Eerikäinen 2001). According to McRoberts et al. (2002), non-parametric methods are unlike regression analyses in that inclusion of additional independent variables may actually increase residual uncertainty. This concerns only modelling data; however, and the adjusted R-square may decrease with increasing number of independent variables in the parametric methods as well. Nevertheless, because of the high flexibility of non-parametric methods, caution must be taken not to over-fit the data, that is, to apply an overly complex model to data so as to produce a good fit that likely will not be replicated in subsequent applications (e.g. Hastie and Tibshirani 1990). This does not pose such a problem when cross-validation is used and observations, for example, from the same stand as the target tree are excluded. However, if the weighting matrix applied in non-parametric methods is calculated on the basis of all possible variables and correlations, the results might be poorer in independent data due to over-fitting compared with fewer independent variables used. The possible nearest neighbours may also be more difficult to find, if there are several dependent variables (McRoberts et al. 2002). Additionally non-parametric methods do not automatically guarantee unbiased estimates as do the regression models in the modelling data (e.g. Korhonen and Kangas 1997).

In non-parametric methods observations are also assumed to be independent of each other. In practical situations the correlation among observations is a common occurrence. In general, correlation can have important consequences on the statistical properties of the estimator and on the selection of the smoothing parameter. The smoothing parameter is usually selected with some kind of data-driven method, such as cross-validation or plug-in methods. However, the presence of correlation among the errors may cause the commonly used automatic tuning parameter selection methods to break down (Opsomer et al. 2001). In forestry applications the correlations may especially pose a problem. The cross-validation method might give too optimistic results of the performance of the method, if the estimate is formed based on nearest neighbours from the same stand where the target tree is situated. Moreover, such data is not available in practical situations. Therefore observations from the same stand as the target tree are usually excluded from the pool of possible nearest neighbours (Packalén and Maltamo 2007). However, at tree level the nearest neighbours

may still all be selected from one particular stand, if the stand-level variables contain much weight in the distance function. Stand-level results may be affected if the errors of all individual trees point in the same direction, for example, if all the nearest neighbours are selected from a stand in a dry site, while the target tree is growing in a damp site.

1.3 Objectives of the thesis

The overall purpose of this thesis was to examine and evaluate different non-parametric methods as a method for growth estimation. One of the main focuses was to test non-parametric methods in order to reduce the regional biases associated in the growth estimates. The non-parametric methods were compared with parametric models at tree, stand and regional levels. This thesis has been implemented in a series of papers, designated I–VI. Each of the individual papers concentrated on one topic, but had some elements in common with other papers, and each of them provided some new information for the next paper. The specific aims of papers I–VI were as follows:

Paper I: To apply and test two different k-nearest neighbour methods for local conditions in north-eastern Finland and to construct five-year individual tree diameter increment estimates under bark for Scots pine and Norway spruce.

Paper II: To compare two different k-nearest neighbour methods and generalized additive models in constructing individual tree diameter increment estimates for local conditions in north-eastern Finland.

Paper III: To construct five-year individual tree diameter increment estimates over bark for Scots pine and Norway spruce applying k-nearest neighbour method, and to examine the localization of this method by including spatial neighbourhood in the imputation in order to obtain regionally unbiased growth predictions.

Paper IV: To simultaneously estimate individual tree diameter and height increment with k-nearest neighbour method. The performance of the method was analysed in different forest site types, within stands and in producing long-term growth forecasts.

Paper V: To test further the simultaneous k-nearest neighbour estimation method against independent test data as well as localization of the non-parametric methods and the effects of reference data size on the accuracy of growth estimates.

Paper VI: To apply different non-parametric methods to the estimation of individual five-year diameter increment estimates for Scots pine and Norway spruce, and compare the performance of different types of non-parametric methods when observations are correlated.

2 STATISTICAL BACKGROUND OF NON-PARAMETRIC METHODS

2.1 Common features of non-parametric methods

In non-parametric methods, the value of the variable of interest for a target observation is estimated often as a weighted average of the values of neighbouring reference observations, which are similar to the target observation in terms of the independent variables and weighted by their proximity to the target observation (Härdle 1989, Altman 1992). Unlike in regression analysis, where the whole data is used, the weighted averages are calculated and the local estimate is formed based on part of the data, with the neighbourhood size varying depending on the method and application used. The non-parametric estimator can be calculated as follows:

$$\hat{y}_i = \frac{\sum_{j=1}^{k} w_{ij} y_j}{\sum_{j=1}^{k} w_{ij}}, \qquad (1)$$

where k is the number of the nearest neighbours, w_{ij} is the weight of the reference tree j to the target tree i and y_j is the value of the variable for reference tree i. Applying non-parametric methods requires decisions regarding the distance function to be used to find the nearest neighbours, neighbourhood size and possible weighting function to define weighting of the reference trees.

Size of the neighbourhood is of critical importance in non-parametric methods. Applying more than one neighbour results in greater precision, but smoothing of the estimates and in particular, the bias of the extreme values of variables of interest may rise with increasing size of the neighbourhood (e.g. McRoberts et al. 2002). The neighbourhood can be selected on the basis of a fixed bandwidth, as with kernel estimators, or a variable bandwidth with a fixed number of nearest neighbours, as in k-nearest neighbour methods (k-NN) (Altman 1992). In fixed bandwidth methods the number of neighbours used varies according to the input space. However, a fixed bandwidth and a weighting function that progresses to zero at a finite distance can involve large variance in areas where the density of the data is low, in particular, on the edges of the dataset or between data clusters. In general, the variance is more stable with nearest neighbour bandwidth selection methods than with the fixed bandwidth approach (e.g. Atkeson et al. 1997).

Weighted averages are used to reduce the bias of the non-parametric estimators (Altman 1992). The weighting function should have its maximum value at zero distance and decrease smoothly as the distance increases (Cleveland and Loader 1994). Weighting functions such as tricube, Gaussian and quadratic are applied; however, one of the most common weighting functions is that based on inverse of the distance (Cleveland 1979, Atkeson et al. 1997). Then the weight of reference tree j for target tree i can be as follows:

$$w_{ij} = \frac{1}{1 + d_{ij}} \qquad (2)$$

2.2 Nearest neighbour methods

Nearest neighbour methods (k-NN) have been applied in the fields of non-parametric statistics and pattern recognition and they continue to be very popular because of their simplicity and suitability to many practical problems (Lin and Jeon 2002). A variety of distance functions have been proposed to be used in nearest neighbour methods. Three commonly used distance measures for continuous variables are based on the Minkowski distance of 1-norm, 2-norm and infinity norm (Batchelor 1978, Rao et al. 2008). These are called Manhattan distance, Euclidean distance and Chebychev distance, respectively. Other distance measures include Mahalanobis (Mahalanobis 1936), Quadratic, Canberra and Chi-Square (e.g. Diday 1974). The Most Similar Neighbour technique (MSN) is based on Mahalanobis distance, but employs weighting derived from canonical correlation analysis and uses single nearest neighbour (Moeur and Stage 1995). Furthermore, Gradient Nearest Neighbour is a specific combination of single nearest neighbour and distance metric based on canonical correspondence analysis (Ohmann and Gregory 2002).

The common terminology concerning the different variations of nearest neighbour methods is not yet stabilized. The k-Nearest Neighbour is the most general term, and it permits the use of various distance measures and any numbers of nearest neighbours. The k-NN method including many nearest neighbours and distance measure based on canonical correlation is often called k-Most Similar Neighbour method (k-MSN) (e.g. Packalen and Maltamo 2008). This is the situation in this thesis as well, when the particular study included more than one variation of the k-NN methods, since easily separable abbreviations were required.

Euclidean distance function can be applied with or without weighting the variables and is usually calculated without taking the square root. Non-weighted function gives equal weight to each of the independent variables. Weights for the variables may be achieved, for example, by applying grid search, non-linear optimization algorithm or genetic algorithm (e.g. Haara et al. 1997, Haara 2002, Tomppo and Halme 2004). The weighted Euclidean distance function can be defined as:

$$d_{ij}^2 = \sum_{l=1}^{p} c_l (x_{il} - x_{jl})^2 , \qquad (3)$$

where
x_{il} = value of the considered variable l for target tree i
x_{jl} = value of the considered variable l for reference tree j
c_l = coefficient for variable x_l
p = number of the variables.

Applying this kind of distance functions requires that variables of different ranges are first standardized, for example, by subtracting the mean of the variable and dividing it by the standard deviation of the variable, otherwise the variables that have large values receive more weight in the distance function. Manhattan distance (or Minskowski 1-norm distance) is based on absolute differences between the values of the considered variables. Similarly to the Euclidean distance function, variables require standardizing and can be non-weighted or weighted. The weights for the variables may be obtained by grid search (Korhonen and Kangas 1997) or by robust regression using least absolute deviations, for example. The weighted Manhattan distance function can be defined as:

$$d_{ij} = \sum_{l=1}^{p} c_l \mid x_{il} - x_{jl} \mid , \qquad (4)$$

where

x_{il} = value of the considered variable l for target tree i
x_{il} = value of the considered variable l for reference tree j
c_l = coefficient for variable x_l
p = number of the variables.

The Mahalanobis distance (Mahalanobis 1936) differs from the Euclidean distance in that it takes into account the correlation structure between the variables, whereas the Euclidean distance is blind to correlated variables and may weight a correlated variable more heavily than other variables even though it does not provide any new information. In the Mahalanobis distance, the inverse of the covariance function is inserted into the middle of the quadratic form in order to reduce the weights of highly correlated pairs (Theodoridis and Koutroumbas 2006). This method enables the coefficients of variables to be obtained directly from a linear regression analysis, and it is therefore computationally fast. The Mahalanobis distance function calculates the squared distance metrics between the target tree and the reference tree as follows:

$$d_{ij}^2 = (X_i - X_j)\,\beta \sum_{zz}{}^{-1} \beta' (X_i - X_j)' , \qquad (5)$$

where
X_i = independent variables of the target tree i
X_j = independent variables of the reference tree j
β = vector of the regression coefficients (or matrix in the case of many dependent variables)
$\sum_{zz}{}^{-1}$ = inverse of the variance of the dependent variable (or inverse of the variance-covariance matrix in the case of many dependent variables).

Distance metric derived from canonical correlation analysis is as follows:

$$d_{ij}^2 = \underset{1\times p}{(X_i - X_j)}\,\underset{p\times p}{\Gamma \Lambda^2 \Gamma'}\,\underset{p\times 1}{(X_i - X_j)'} , \qquad (6)$$

where
X_i = independent variables of the target tree i
X_j = independent variables of the reference tree j
Γ = matrix of canonical coefficients of the independent variables, $\underset{pxs}{\gamma_k}$
Λ^2 = diagonal matrix of squared canonical correlations, $\underset{sxs}{\lambda_k}$
s = number of the canonical correlations used
p = number of the independent variables.

The weighting matrix in the distance function is calculated on canonical correlation analysis by summarizing the relationships between dependent and independent variables simultaneously (Moeur and Stage 1995). In canonical correlation linear transformations (U_r and V_r) are formed from the set of dependent and independent variables, in such a way that the correlation between them is maximized

$$U_r = \alpha_r Y \text{ and } V_r = \gamma_r X \ , \qquad (7)$$

where U_r represents the canonical coefficients of the dependent variables ($r = 1 \ldots s$) and Vr the canonical coefficients of the independent variables ($r = 1 \ldots s$). There are s possible pairs of canonical variates (U_r and V_r) as the result of the analysis, where s is either the number of dependent or independent variables, depending on which is smaller. Canonical variates are ordered in such a way that canonical correlation between them is the largest for variate (U_1,V_1), second largest for (U_2,V_2) and so on. Thus, the predictive relationship between original variables is concentrated in the first few canonical variates and less important variates can be left out without loss of predictability (Moeur and Stage 1995). However, canonical correlation and linear regression give equivalent weighting when there is only one dependent variable, or when the full-rank coefficient matrix is used in both. If multiple variables are estimated simultaneously, canonical correlation formulation offers the possibility to restrict the distance function to use only the first significant canonical variates, which may guide the nearest neighbour selection towards the variables that are most useful (Moeur and Stage 1995, Crookston et al. 2002). On the other hand, if correlations between all dependent and independent variables are taken into consideration, the neighbourhood selection is guided towards the best all-around neighbour (Crookston et al. 2002). However, the correlation structure may be more straightforward to interpret and possible transformations for the independent variables easier to find with linear regression analysis (Maltamo et al. 2003).

2.3 Generalized additive models

Generalized additive models (GAM) are a method of fitting a smooth relationship between two or more variables through a scatterplot of data points (Hastie and Tibshirani 1986). The purpose of GAMs is to maximize the quality of prediction of a dependent variable from various distributions, by estimating unspecific functions of the independent variables which are connected to the dependent variable via a link function. GAMs are extensions of generalized linear models (GLM) (Hastie and Tibshirani 1987). The only underlying assumption in generalized additive models is that the functions are additive and that the components are smooth. However, the probability distribution of the dependent variable must still be specified, and in this respect, generalized additive models are parametric. Thus they are more aptly named semi-parametric models rather than non-parametric methods (Guisan et al. 2002). A wide variety of distributions for the dependent variable are allowed to be selected though, as well as many link functions (McCullagh and Nelder 1989, Hastie and Tibshirani 1990). Generalized additive models consist of a random component, an additive component and a link function relating these two components. GAMs have the form

$$\eta(X) = \alpha + \sum_{j=1}^{p} f_j(X_j), \qquad (8)$$

where $\eta(X)$ is a known link function and $Ef_j(X_j) = 0$. GAMs replace the linear form in parametric models $\Sigma\beta_j(X_j)$ by a sum of smoothing functions $\Sigma f_j(X_j)$ (Hastie and Tibshirani 1986). Instead of estimating single parameters like the regression weights in multiple regression, a general unspecific non-parametric function that relates the predicted dependent values to the independent values is specified. By assuming that the mean function is a sum of one-dimensional smooth functions, the curse of dimensionality can be avoided (Opsomer 2000b). Furthermore, the resulting one-dimensional additive fits are easily displayed and interpreted, unlike in unrestricted multi-dimensional smoothing (Opsomer 2002).

The estimation procedure for a GAM requires iterative approximation in order to find the optimal estimates. In general, there are two separate iterative operations involved in the algorithm, which are usually called the outer and inner loop. The purpose of the outer loop is to maximize the overall fit of the model by minimizing the overall likelihood of the data given the model. The purpose of the inner loop is to refine the scatterplot smoother. In particular, the estimation is based on a combination of local scoring algorithm and backfitting algorithm. The local scoring procedure uses a scatterplot smoother as a building block in the estimation of individual components of the additive model. Many different univariate and bivariate smoothing techniques can be used, for instance, running means, running lines, kernel, splines and locally weighted regression models.

In a standard approach inside each step of the local scoring algorithm a weighted backfitting algorithm is applied to the adjusted dependent variable until convergence. Then, based on the estimates from this weighted backfitting algorithm, a new set of weights is formed and the next iteration of the local scoring algorithm starts. During each iteration, an adjusted dependent variable and a set of weights are computed, and then the smoothing components are estimated using a weighted backfitting algorithm. At each step of the backfitting algorithm, partial residuals are defined and one component is estimated keeping the other components fixed. The partial residuals are obtained by removing the estimated functions or covariate effects of all other variables (Hastie and Tibshirani, 1987). The backfitting algorithm cycles through the partial residuals fitting the individual smoothing components to its partial residuals. The iterative procedures are repeated until convergence (Hastie and Tibshirani, 1986).

Locally weighted regression (LOESS) is a method for constructing an estimate from observed data by fitting a model in a local manner by defining a neighbourhood of the target observation in the space of the independent variables and weighting the points in the neighbourhood according to their distance from the target observation (Cleveland 1979). The observations close to the target observation x_i have large weight and observations far from x_i have small weight (Cleveland and Devlin 1988). Hence, the smoothing function is formed pointwise to a subset of the data by fitting a polynomial using weighted least squares. The local polynomials fitted to each subset of the data are usually of first or second degree (Cleveland and Devlin 1988). Many details of the locally weighted regression are flexible, such as the degree of the polynomials and the form of the weighting function. For instance, tri-cube weight function, where the weights are proportional to the cubic distance from the target observation x_i, may be applied. First, largest distances between observations x_i and x_j in the neighbourhood $N(x_i)$ are calculated by using

$$\Delta(x_i) = \max_{x_j \in N(x_i)} \left| x_i - x_j \right| . \quad (9)$$

Weights are calculated for every observation in the neighbourhood using tri-cube weight function:

$$W(z) = \begin{cases} (1-z^3)^3 & \text{, for } 0 \le z < 1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

and

$$z = \frac{\left| x_i - x_j \right|}{\Delta(x_i)} . \quad (11)$$

Thus, the observations outside of the set of nearest neighbours of the target observation x_i receive zero weight (Cleveland 1979). The final estimate is the predicted value from a weighted least squares fit of the dependent variable values of the reference observations on the neighbourhood $N(x_i)$.

Smoothing splines emerges as a solution to an optimization problem. A smoothing spline is the solution to the following optimization problem: among all functions $f(x)$ with two continuous derivatives, find one that minimizes the penalized residual sum of squares

$$\sum_{j=1}^{n} (y_j - f(x_j))^2 + \lambda \int_a^b (f''(t))^2 \, dt , \quad (12)$$

where λ is a fixed constant and $a \le x_1 \le ... \le x_n \le b$ (e.g. Silverman 1984, Härdle 1990). The first term in the equation measures closeness to the data, while the second term penalizes curvature in the function. It can be shown that there exists an explicit, unique minimizer, and that minimizer is a natural cubic spline with knots at the unique values of x_j. The parameter λ is the smoothing parameter. Large values of λ produce smoother curves, while smaller values produce wiggly curves. The boundary defined by a and b is arbitrary, as long as it contains the data. The smoothing spline is linear beyond the data points regardless of the values of a and b (e.g. Härdle 1990).

3 STUDY MATERIAL

The study material consisted of two different datasets. The first is a small dataset collected from temporary sample plots from the areas owned by Kuusamo Common Forest situated in Kuusamo in north-eastern Finland. This data was purposely collected to be used in estimating individual tree growth with non-parametric methods (I, II, IV). The second dataset is the nationwide permanent inventory growth plot database INKA, provided by the Finnish Forest Research Institute (Gustavsen et al. 1988) (III–VI). These data were applied

in a different manner in each of the separate papers. Growth estimates were constructed for (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* (L.) H. Karst.), except in paper II, where only Scots pines were used. Birch trees were excluded from all of these sub-studies because of limited numbers of birches in both datasets.

3.1 Kuusamo data

The Kuusamo data was collected from temporary sample plots from the areas owned by Kuusamo Common Forest situated in Kuusamo in north-eastern Finland. The sample plots were measured during the summer of 1999. Sampling included seven main strata according to stand register data: pine and spruce dominated damp forest site types, pine dominated dryish and dry forest site types, pine and spruce swamps and pine forests with low productivity. The Finnish forest site type classification, which originates from botany and was started by A.K. Cajander more than 100 years ago, is based on the assumption that the presence of different plant species is determined by the ecology of the habitat, and the habitat characterized by certain vegetation reflects the potential forest productivity of that site (e.g. Cajander 1909, Lindholm and Heikkilä 2006). Forest site type groups are poor dry (barren), dry, semi dry (dryish), mesic (damp), semi herb rich (rich) and herb rich (very rich) (Kalliola 1973). All the main strata were further divided into six 30-year age classes and two stands were supposed to be measured from each of these strata. The stands with notable damage or dominant height lower than 3 metres were not included in the data.

Two fixed-radius circular plots were systematically placed in each sample stand. The distance between the two plots was 40 metres apart from the centre of each other and the plot size varied from 100 m^2 to 700 m^2 according to the stand density. Tree species and diameter at breast height (DBH) were recorded for all tallied trees in these plots. From every plot, an average of nine sample trees were selected for more detailed measurements by establishing a circular subplot comprising a quarter of the area of the larger plot. Characteristics of the sample trees measured within the inner circles included tree height, length of the live crown, bark thickness and five-year diameter increment. Mean stand age was determined by measuring age from one-third of the sample trees. In addition, several variables describing the site and the growing stock were also registered for each stand. These variables included location, altitude, temperature sum, soil type, forest site type group and dominant tree species.

A total of 71 stands were measured, comprising 53 stands dominated by Scots pine and 18 stands dominated by Norway spruce. The whole measured data consisted of 4051 tally trees and 1308 sample trees, the latter including 941 Scots pines and 367 Norway spruces. Most of the pines were located in damp and dryish forest site types and the proportion of pines located in dry sites was low. Norway spruces were mainly located in damp sites. Most spruces belonged to mature forests and the proportions of other stages of stand development were small. The pines were more evenly distributed to different age classes. Mean age of the spruce stands was 109 years and pine stands 65 years. The average of the five-year stand-level volume growth was 12.8 m^3ha^{-1} with a standard deviation of 7.1 m^3ha^{-1}. The minimum, maximum and mean values of the most important tree- and stand-level characteristics are presented in Table 1 of paper I and in Table I of paper V.

Preparation of the Kuusamo data included calculation of tree and stand-level characteristics that had not been directly measured and variables that describe competition among the trees within a stand. Basal area of the stand (BA), basal area median diameter

(D_{gM}), dominant height (H_{dom}), relative size of a tree (d_{rel}) and basal area of the trees larger than the subject tree (BAL) were calculated based on the tally tree plots, including all tree species. Relative size of a tree was calculated by dividing DBH with D_{gM}. Data preparation also included back-calculations of all characteristics, since the data were collected from temporary sample plots. Tree diameter under bark for the sample trees at the beginning of the growth period was calculated by subtracting the measured five-year diameter increment and thickness of the bark from the measured tree diameter. Bark thickness and tree height at the beginning of the growth period were estimated with random parameter models applying the MIXED procedure in SAS (SAS 1992). In addition, simple regression models were separately constructed for every sample plot to calculate tree diameters at the beginning of the growth period for tally trees. Other tree and stand characteristics at the beginning of the growth period were calculated by means of estimated tree diameters and heights.

Sample trees from the Kuusamo data were used as study material in the non-parametric estimation of diameter increment for local conditions (papers I and II), and as independent test data in paper V. In paper II, the Kuusamo data was further divided into separate target data and reference data so that the proportions of different forest site types and age-classes were similar in both data.

3.2 INKA data

The INKA data consisted of a sample of the stands measured for the sixth National Forest Inventory (NFI6) in northern Finland and seventh National Forest Inventory (NFI7) in southern Finland, thus the stands were distributed extensively over the whole area of Finland (Fig. 1) (Gustavsen et al. 1988). Only stands on forest land with mineral soils were included and sapling stands, i.e., stands with dominant height below 5 metres, were excluded. However, the INKA data included a few stands in drained peatlands and swamps, and these stands were not excluded from these studies (e.g. Table 2 in IV). Healthy, single-storied stands with the proportion of the major three species being at least 50% of the total volume of the growing stock were included. The dominant tree species considered were Scots pine, Norway spruce or birch. Furthermore, the basic population was restricted to pine-dominated stands on dry and dryish sites and pine, spruce and birch-dominated stands on damp sites. The plots were established during the years 1976–1983 and have in most cases been re-measured twice at intervals of five years. The original INKA data consisted altogether of 828 measured stands. In each measured stand a cluster of three fixed-radius circular plots was established. The plots were located systematically 40 metres apart from the centre of each other. The size of the plot varied according to the stand density, in order that at least 100 trees were measured in each stand in northern Finland and 120 in southern Finland. In addition to stand descriptors, tree species and DBH were recorded for all trees on these plots. A smaller concentric circular sample plot equal in size to one-third of the tally tree plot was delineated within each plot, and tree height and crown length were recorded for all sample trees within these latter plots.

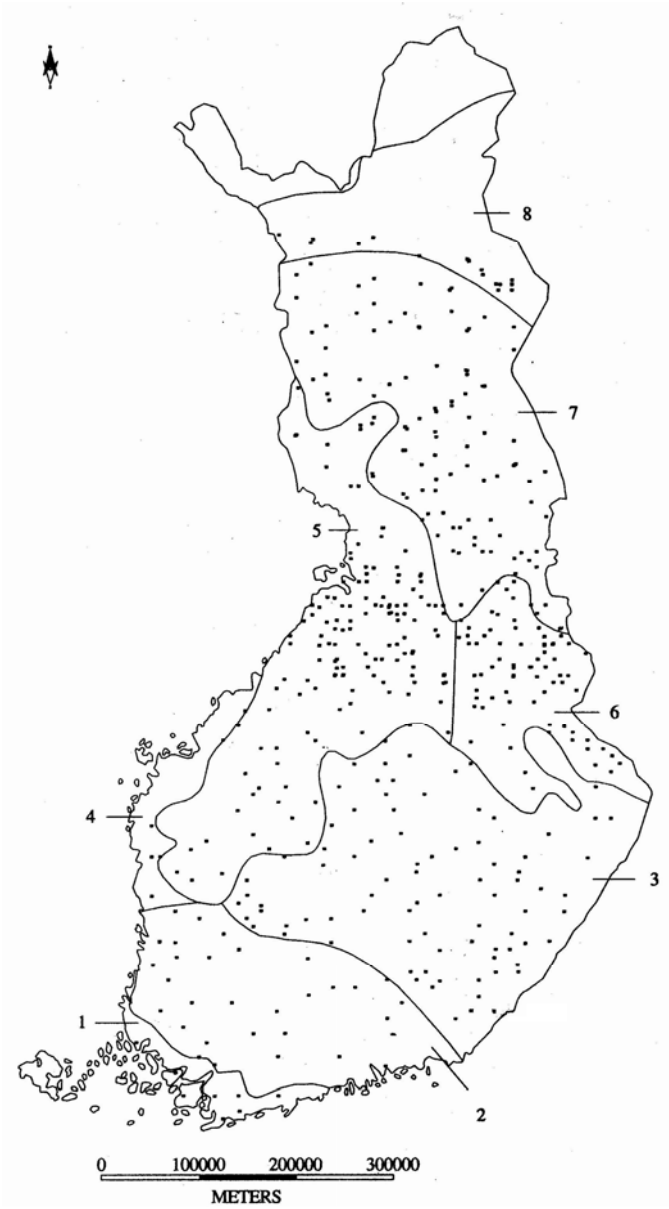

Figure 1. Locations of the original permanent inventory growth plots (INKA). The areas represent the eight vegetation zones into which Finland is divided.

Similarly to the preparation of the Kuusamo data, stand-level variables that had not been directly measured and variables that describe competition among the trees within a stand were calculated. Since the heights of all trees were needed to calculate tree volumes and dominant tree heights, for instance, random parameter height models were constructed for Norway spruce, Scots pine and birch based on measurements of trees on plots located in different stands (III). Diameter increments over bark and height increments were obtained

by means of two successive measurements. The minimum, mean and maximum values, as well as standard deviations, of the most important tree- and stand-level characteristics are presented in Table 2 of paper III and Table 1 in each of papers IV, V, VI. Note that these tables are calculated on the basis of the sub-data used in different papers, including those observations of the original INKA data that fulfil the requirements stated in each paper. The numbers of observations used in performing the imputations are presented in Table 1 and 2 of paper III and in Table 1 of paper IV, V and VI.

The study material of paper III comprised both tally trees and sample trees of INKA data from two growing periods. Each five-year growing period of a tree was used as one growth observation. In order to examine the effects of localization methods and regional accuracy, the data were further divided into eight subsets consisting of trees in the eight sub-boreal vegetation zones into which Finland is divided; hemi-boreal, south-western, Lake District, southern Ostrobothnia, Ostrobothnia, Kainuu, southern Lapland and Forest Lapland (Fig. 1 in III) (Kalliola 1973, Maltamo et al. 2003). The fourth and fifth papers (IV, V) used the sample trees of the INKA data for which tree diameter, tree height and length of the live crown had been recorded at the second and the third measurement occasion. Therefore the study material contained observations from one five-year growing period. The material for the sixth paper (VI) consisted of the tally trees and sample trees from the second and third measurement occasion located in south-western Finland, the Lake District and Ostrobothnia. These three vegetation zones form a large area in the middle of Finland. The coastal areas, northern Finland and Kainuu were excluded from the data.

4 METHODS

4.1 Local non-parametric growth estimates (I and II)

Local non-parametric diameter increment estimates for Kuusamo in north-eastern Finland were constructed using nearest neighbour methods and generalized additive models (GAM). The distance functions applied in the nearest neighbour methods included Manhattan distance (referred to here as k-NN Manhattan) and distance measure based on canonical correlation (referred to here as k-MSN). Optimal variables for the distance function, coefficients of the variables, number of the nearest neighbours and weighting parameter were determined using grid search when applying k-NN with Manhattan distance. Both five-year diameter increment under bark and bark thickness were estimated on the basis of the same neighbouring trees. However, the estimation was not performed simultaneously, thus the only variable whereby the weights were optimized was diameter increment. The GAMs for the five-year diameter increment under bark and for the thickness of the bark were first fitted to the reference data and then growth and bark estimates were predicted for the target data (II). Smoothing splines and locally weighted regression were tested as scatterplot smoothers.

4.2 Localization of the non-parametric growth estimates (III)

Individual tree diameter increment estimates were constructed with non-spatial k-NN method applying Mahalanobis distance, which was referred to as the basic k-NN method (BASIC). This method was then localized in various ways (Fig. 2). The first approach was to use geographical coordinates measured at sample plot level as independent variables (COORDINATES). The coordinates were measured from the plot centre and represent variation of growth in a larger area. The second approach was to restrict the spatial neighbourhood by using moving geographical areas (CIRCULAR). This involved selecting the nearest neighbours from a circle around the target tree, having first tested radiuses with varying sizes from 50 to 300 kilometres for this circle. Thus, in the localized methods real space is included in addition to variable space. The estimates were formed on the basis of the whole dataset, as were the basic non-spatial k-NN estimates. The growth estimates given by these localized methods were subsequently divided into vegetation zones and regional growth estimates were calculated.

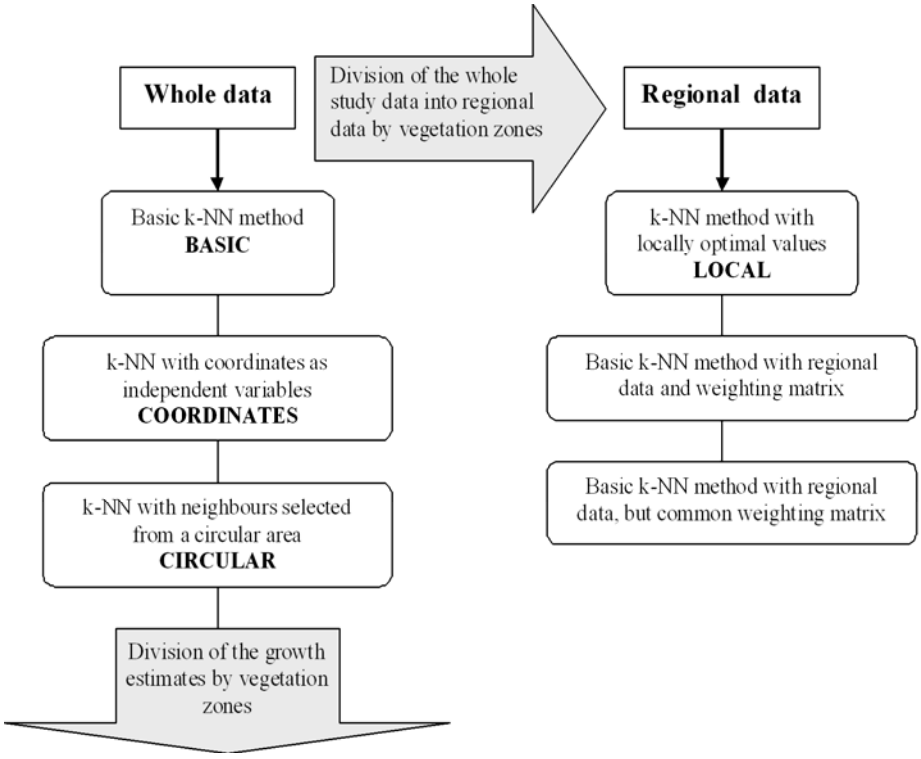

Figure 2. Different localization methods in Paper III.

Localization was also achieved by choosing the neighbours from a local database in order to be able to compare this approach with localizing methods based on the whole dataset. Local k-NN estimates were constructed in three ways (LOCAL). Separate local k-NN with optimal values for all parameters were first performed for each region based on the data divided into vegetation zones. Secondly, the basic k-NN, i.e., the same variables, was used but with regional weighting matrices and neighbours selected from the regional data. Finally, the basic k-NN method was used and the weighting matrices were common to the whole dataset, i.e., formed on the basis of the whole data, while the neighbours were still selected from the corresponding regional data.

4.3 Predicting tree- and stand-level growth using simultaneous k-Nearest Neighbour imputation for diameter and height increment (IV)

Individual tree diameter and height increments were constructed with the k-NN method using distance measure based on canonical correlation analysis. Performance of the k-NN method and parametric models were analysed and compared in more detail. First, the logic of the growth estimates produced in relation to the position of a tree in a stand was analysed by means of plotting the growth estimates versus relative tree size. The figures were created from six randomly selected stands dominated by the Scots pine or Norway spruce and representing various age classes. Secondly, the performance of the methods was analysed in various forest site-type groups by calculating RMSEs, biases, averages, standard deviations and variances of the growth estimates both at tree and stand level within each group. Finally, the performance of the k-NN method in forecasting long-term growth was analysed by producing growth estimates for an 80-year growing period at five-year intervals and then plotting the growth curves against growing period both at tree and stand level. Five young stands representing different forest site types were randomly selected from the INKA data for this analysis. Simulations were performed without thinning, including one thinning, and including two thinnings during the forecasting period. In addition, the simulations were performed applying thinning dummy as auxiliary variable in the k-NN method.

4.4 Effect of reference data selection on the accuracy of non-parametric k-NN imputation for individual tree growth (V)

The simultaneous k-NN method was tested against independent test data collected from Kuusamo in northeastern Finland. Evaluation was carried out by using the Kuusamo data as the target data for which the diameter and height increment estimates were imputed and the INKA data as the reference data. Furthermore, localization of the k-NN method and, in particular, the effects of local observations and size of the reference data on the accuracy of the growth estimates produced for the Kuusamo area were tested. This analysis was carried out by adding a varying number of local observations from the Kuusamo data to the reference data (Fig. 3).

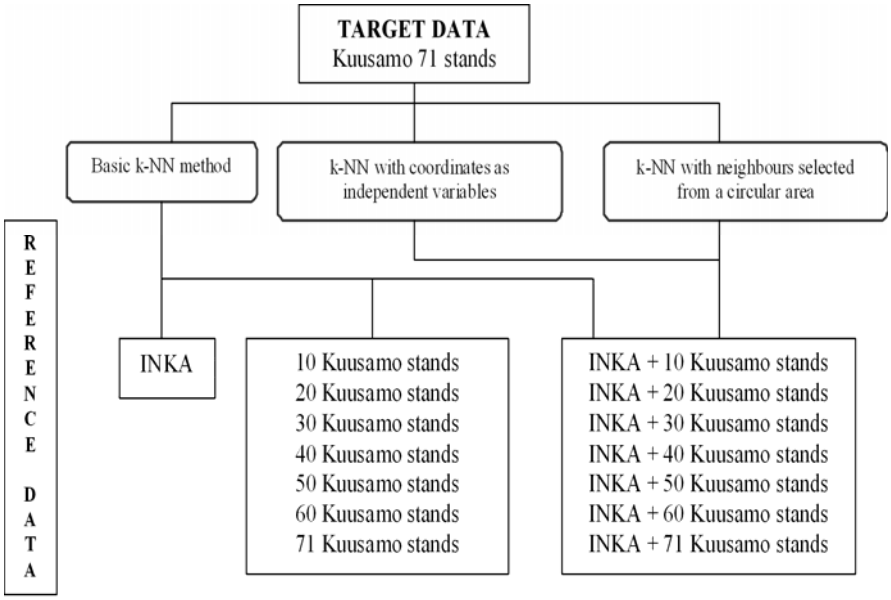

Figure 3. Different k-NN methods and reference datasets applied in Paper V.

4.5 Comparison of different non-parametric growth imputation methods in the presence of correlated observations (VI)

The applied non-parametric methods included two k-NN methods and generalized additive models (GAM). In the k-NN methods the distance measure was based on either Euclidean distance (referred to here as k-EUC) or canonical correlation analysis (referred to here as k-MSN). Examining the effects of correlated observations was implemented by defining different restrictions to the pool of possible reference trees in all of the abovementioned methods:

1. No restrictions to the possible reference trees.
2. Plot restriction: trees from the same plot as the target tree were excluded from the possible reference trees.
3. Stand restriction: trees from the same stand as the target tree were excluded from the possible reference trees.
4. One per plot restriction: in addition to stand restriction (3) only one tree from each original INKA plot was allowed to be in a group of possible nearest neighbours
5. One per stand restriction: in addition to stand restriction (3) only one tree from each original INKA stand was allowed to be in a group of possible nearest neighbours.

4.6 Evaluation criteria

The results were calculated by means of leave-one-out cross-validation, which can be used to estimate the generalization error of a given model or to choose from among several models the one that has the smallest estimated generalization error (Härdle 1989). In this method, each observation is used in turn as a target tree and predicted with the reference data excluding the observation itself. It was also determined that the nearest neighbours should not be selected from the same stand, i.e., from the same cluster of three plots, as the target tree on either of the measurement periods, since observations in the same stand are closely correlated and would be given too much weight in the calculations and might give excessively optimistic results or the performance of the model. The root mean squared error (RMSE) and mean of residuals (bias) were used as criteria for choosing the variables and assessing the reliability of the estimates. The RMSE weights the average goodness of the estimates, but penalizes biased estimates, since the squaring ensures that negative values do not cancel out positive ones. The root mean squared error was calculated as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \qquad (13)$$

where n denotes the number of observations, y the observed growth for observation i and $\hat{y}$ denotes the growth estimate for observation i. The bias was calculated as

$$\mathrm{bias} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n}. \qquad (14)$$

The relative RMSE and relative bias of the growth estimates were calculated by dividing the absolute values by the observed mean growth of the observations. The imputations were performed at the tree level, but then summarized to the stand level, and the evaluations were made both at tree and stand level. The evaluation concerned either individual tree diameter or height increment at the tree level, and volume growth at the stand level. The stand-level volume growth estimates were calculated by subtracting the true volume of a stand at the beginning of the growth period from the estimated volume at the end of the growth period. The stand volumes were obtained by summing the volumes of the trees in a particular stand. The volumes of the trees were calculated with volume functions based on the tree diameter and tree height developed by Laasasenaho (1982).

4.7 Comparisons with parametric growth predictions

Local nearest neighbour methods (I) were compared with a regression growth model constructed from the same study data as the non-parametric methods. The regression model was built with mixed model technique, because the observations were correlated due to the hierarchical structure of the study data (e.g. Lappi 1993). Furthermore, the stand-level volume growth estimates were compared to the volume growths produced within the

Monsu-forest planning programme (Pukkala 2000), which used single-tree regression growth models developed by Nyyssönen and Mielikäinen (1978).

In papers III, IV and V, parametric growth predictions calculated within the SIMO (simulation and optimization for next-generation forest planning) simulation framework (Tokola et al. 2006, Rasinmäki et al. 2007) applying MELA2002 models (Hynynen et al. 2002) were used for comparison purposes. The MELA models are developed to be applicable to management planning tools throughout the whole Finland, and special attention has been paid to the ability to predict the responses to silvicultural practices (Hynynen et al. 2002). MELA2002 models are developed from the permanent inventory growth plot data (INKA) and are applicable to all the tree species and forest site types occurring throughout Finland. The growth of trees, in particular, is predicted with distance-independent models for basal-area growth and height growth. The MELA models were applied with and without the self-thinning model (Hynynen 1993). Volumes for the trees were calculated similarly, since the SIMO framework utilizes the volume functions of Laasasenaho (1982) as well.

5 RESULTS

5.1 Variable selection and size of the neighbourhood (I-VI)

The choice of independent variables depended on the information that was available for every tree and how these variables were related to diameter increment, and also to the height increment when simultaneous estimation of them both was considered. The candidate independent variables were chosen from among the easily measured or traced tree and stand characteristics that describe tree size, phase of stand development, competitive situation of a tree in terms of distance-independent competition measures and the growing site. The tested tree-level variables included DBH, tree height, d_{rel} and BAL. In papers IV and V tree crown ratio (CR) was also included in the group of tested tree-level variables. The stand-level variables tested were stand age, H_{dom}, BA, D_{gM}, altitude, temperature sum and forest site type. Correlations between the dependent and independent variables varied from low to moderately high (Table 3 of III and Table 3 of IV). Stand age, BAL, CR, d_{rel} and BA mainly correlated well with diameter increment, depending somewhat on the data used in each paper. Height increment had the highest correlations with stand age and CR.

Transformations of the independent variables for linearizing the relationship between diameter increment and independent variables were tested in all of the papers, utilizing k-NN method based on distance measures derived either from linear regression or canonical correlation. Linear correlation is assumed between the variables in these methods, and the method is more efficient the more linear the relationship is (e.g. Maltamo et al. 2003). The tested transformations included second powers, square roots, natural logarithms and inversions of the variables, together with various ratios between the independent variables. Maltamo et al. (2003), for example, noted that their results were considerably improved when second powers of the dependent and independent variables were used. However, the non-linear relationships between the variables could not be described with these transformations. The only transformation that improved the results was the use of the

inverse of the stand age in the k-NN methods for spruces in papers III–VI, otherwise the effect of transformations was minor.

The performance of the nearest neighbour methods was greatly affected by the number of neighbours used in performing the imputations. This effect was tested in all of the papers by varying this value (*k*) from 1 to 20. The number of neighbours affected the accuracy more than the optimal variables in the distance function, if the most critical variables were included in the imputations. These variables included, for example, stand age or crown ratio, H_{dom} and BAL (III, IV, V and VI). The number of the nearest neighbours used had a fairly similarly effect on the relative RMSE of all the different k-NN methods applied in this thesis (I–VI). The relative RMSE of the k-NN estimates diminished markedly, i.e., by about 20%, when the number of the neighbours was increased from one to about ten, after which it remained relatively stable. The impact of neighbourhood size on relative bias of the growth estimates was not so straightforward. Usually the relative bias was largest with few neighbours. Otherwise the bias was either virtually stable, irrespective of the number of neighbours, or varied randomly. The number of the nearest neighbours used in the final imputations was set to be the value of *k* where the decrease in RMSE% was stabilized and the relative bias was at its lowest or as low as possible. The neighbourhood size is substantially larger in the generalized additive models than in the nearest neighbour methods. The span size is a percentage of all the observations in the space of independent variables. While GAMs were applied (II and VI), span sizes from 0.05 to 0.5 were tested. The smaller the span, the smaller the RMSE and the bias; however, the model did not fully converge with the smallest span sizes. Moreover, the accuracy of the GAM estimates was not as considerably affected by the size of the neighbourhood as the accuracy of the nearest neighbour methods.

5.2 Local non-parametric growth estimates (I and II)

The performance of different non-parametric methods in estimating diameter increment under bark for local conditions was tested and compared in papers I and II. These papers differed in that data-splitting was applied in the second paper (II) in order to be able to compare generalized additive models to the nearest neighbour methods under similar conditions. The procedure used in GAM analysis did not allow restrictions to the pool of possible nearest neighbours, thus the GAM would have given overly optimistic results without data-splitting. The distance functions applied in the nearest neighbour methods included Manhattan distance (k-NN Manhattan) and distance function based on canonical correlation (k-MSN). The grid search applied to the k-NN Manhattan restricted the number of independent variables included; however, the results showed that the best accuracy was achieved with just a few independent variables included in the other methods as well, these variables being tree diameter, tree height, stand age and BAL. Furthermore, when the data were divided into separate target and reference data, the best accuracy was achieved without BAL in the k-MSN and GAM, since the relative RMSE increased by 20% if the BAL was included in the k-MSN, for example.

In the first paper, the accuracy of both of the nearest neighbour methods was at the same magnitude at the tree level, the relative RMSEs being about 50% for Scots pine diameter increment estimates and slightly under 70% for Norway spruce diameter increment estimates (Table 3 of paper I). Both methods produced fairly unbiased growth estimates at tree level for Scots pine, the biases being somewhat larger for spruces. In the second paper

(II), which included only Scots pines, the results were similar, although the RMSE% in general was slightly larger. Data-splitting increases the estimated error, since the target observations will be paired with a more remote reference observation, the withheld reference observations could have supplied imputations for nearby target observations without data-splitting (e.g. Stage and Crookston 2007). However, when the stand-level volume growths were compared in paper I, the accuracy of the growth estimates obtained with the k-MSN method was remarkably poorer. The relative RMSE was 67% for the k-MSN method, while it was 39% for the k-NN Manhattan (Table 4 of paper I). Stand age had relatively more weight than, for example, tree diameter in the k-MSN method. The effect of this was that the k-MSN method tended more often to produce growth estimates on the basis of different sized trees situated in the same stand, while k-NN Manhattan produced growth estimates on the basis of same-sized trees situated on different stands. The k-MSN method included stand-level bias, notably it underestimated the five-year stand-level volume growth. Additionally, the reliability of the bark thickness estimates was worse in the k-MSN method; however, the k-MSN method slightly overestimated the bark thickness, thus they did not increasingly cause stand-level underestimations of the method.

Locally weighted regression was found to be the most reliable smoother while fitting GAMs. The same independent variables were chosen for the growth model as in the k-MSN method. The accuracy of the diameter increment estimates obtained with generalized additive models was notably poorer than the accuracy of other applied methods. The relative RMSE was as high as 118.2%. In contrast to the nearest neighbour methods, generalized additive models gave notable underestimations for small trees, of which 40% of the target data consisted.

The parametric models that were constructed for comparison purposes gave better results for Norway spruce than the non-parametric methods, but the accuracy of Scots pine diameter increment estimates was much lower (Table 3 of I). In addition, the method was less accurate at the stand level than the non-parametric methods, as were the stand-level volume growth estimates produced by the Monsu forest planning programme (Pukkala 2000) applying regression models developed by Nyyssönen and Mielikäinen (1978) as well (Table 4 of I). Especially the stand-level volume growth estimates of the k-NN method were more reliable than those of the parametric methods.

5.3 Localization of the non-parametric growth estimates (III)

The results achieved with the basic non-spatial k-NN method for Scots pine were most accurate when DBH, stand age, H_{dom}, BAL, BA and temperature sum were used as independent variables, while those for Norway spruce were most accurate when the inverse of stand age was used instead of stand age as such. All the localized methods had mainly the same independent variables, except that the temperature sum was not included, since the accuracy diminished notably by including too many variables. Separate local k-NN estimates were constructed with three different ways in order to examine whether optimal regional variables would be found and improve the results. However, search for the regionally optimal values and dependent variables did not improve the accuracy of the regional growth estimates. In the localizing method, applying moving circular areas around the target tree, the radius of the circle was determined to be 125 km for Scots pine and 150 km for Norway spruce.

Both basic non-spatial k-NN and all local and localized methods applied produced fairly similar accuracy for the diameter increment estimates at the country level, the relative RMSEs being nearly 60% for Scots pine and 68% for Norway spruce and the relative biases nearly zero for all the methods (Table 1 below, Table 4 of paper III). Furthermore, the accuracies of the stand-level volume growth estimates were similar, the relative RMSE being 20% and the bias 2% for the whole of Finland (Table 1 below, Table 5 of paper III).The results of the different non-parametric methods in terms of relative RMSE of the diameter increment estimates were closely similar when viewed regionally as well; only in southern Ostrobothnia and Forest Lapland did separate local k-NN method perform somewhat better. The relative RMSEs of the different methods varied across the regions from 43% to 68% for Scots pine and from 49% to 83% for Norway spruce (Table 4 of III). The local k-NN method produced less biased estimates at the tree level than the basic and localized k-NN. The local k-NN method was almost unbiased in most of the vegetation zones as well. However, at the stand-level the biases of volume growth estimates produced by all the methods were at the same magnitude in almost every vegetation zone. The relative RMSEs of the five-year stand-level volume growth varied in the range of 10–20% across the regions, except that all methods performed much poorer in the hemi-boreal zone (Table 1, Table 5 of III).

Table 1. Accuracy of the diameter increment estimates obtained by the different methods, by vegetation zones (Paper III).

Scots pine	Bias, %				RMSE, %				
	Local	Basic	Coords	Circular	Local	Basic	Coords	Circular	Obs.
Hemi-boreal	2.8	11.9	9	12.8	65.6	62.8	63.2	63.7	1646
South-western	-0.1	-3.5	-2.7	-1.8	65.3	64.7	64	62.8	5767
Lake District	0	3.7	2.9	3.1	56.1	56.7	56.5	55.8	9904
S. Ostrobothnia	2.2	-13.1	-8	-16.2	42.7	47.2	46.2	49.9	767
Ostrobothnia	0.2	-11.3	-10.4	-7.5	66.2	68.4	67.3	66.7	22056
Kainuu	0.2	7.4	4.7	9.2	54.6	55.9	55.4	54.2	10677
S.Lapland	-0.3	4.6	6.3	1.1	63.6	64	63.1	61.8	21415
Forest Lapland	0	9.4	14.1	0.1	46.6	50.5	50.6	47.2	3556
Whole Finland	0.1	0.5	0.9	0.3	61.3	62.3	61.6	60.6	75788
Norway spruce	Bias, %				RMSE, %				
	Local	Basic	Coords	Circular	Local	Basic	Coords	Circular	Obs.
Hemi-boreal	3.6	3.3	4.3	7.8	71.6	71	71	75.5	1248
South-western	0	-2.7	-1.3	-1.2	67.2	67.7	68.7	68.8	8375
Lake District	-1.7	5.6	4.4	-0.4	60.8	57.8	59.3	59.7	8021
S. Ostrobothnia	-0.8	17.4	19.3	21.1	55.1	71.3	73.5	72	598
Ostrobothnia	0.1	-4.4	-3.2	-1.5	68.2	69.5	70.4	68.6	9501
Kainuu	-0.1	-1.2	-4.2	-0.8	68.3	67.3	70.3	66.5	4589
S.Lapland	0	-0.1	-1.6	2.3	79.4	81.4	83	78.8	4899
Forest Lapland	-0.2	-1.5	-0.4	-2.5	49.1	57	57.3	53.4	168
Whole Finland	-0.4	0.5	0.4	0.1	68.1	67.8	69.2	68.7	37399

All the non-parametric k-NN methods produced less biased stand-level volume growth estimates than the parametric models. The parametric models produced more biased results especially in Forest Lapland, Ostrobothnia, the hemi-boreal zone and in southern Finland, while the bias of the k-NN estimates did not vary notably across the regions (Fig. 4). The relative RMSEs of the parametric model varied from 35.6% to 66.1% across the regions, and were therefore larger than those for the k-NN methods (Table 2 below, Table 5 of paper III). The k-NN methods were also tested without stand age as an independent variable, and even then they produced smaller RMSEs and biases in the various regions than the parametric models.

Table 2. Accuracy of the stand-level volume growth estimates obtained by the different methods, by vegetation zones (Paper III).

Stand-level I_{V5}	Bias, %					
	Local	Basic	Coords	Circular	Parametric	Obs.
Hemi-boreal	-5.4	-4.7	-7	-6.5	-29.6	28
South-western	3	3.1	4.5	3.8	-29.2	134
Lake District	3.2	6.5	6.1	4.3	7.1	188
S. Ostrobothnia	2.4	-2.1	1.6	-1.7	21	16
Ostrobothnia	1.8	-4.1	-2.9	-1.7	14.5	337
Kainuu	2	4.8	3.2	6	11	163
S. Lapland	0.9	3.6	6.1	-0.1	23.5	280
Forest Lapland	-3	3.2	8.2	-4.1	33.2	40
Whole Finland	2	1.9	2.3	1.9	4.4	1186
	RMSE, %					
	Local	Basic	Coords	Circular	Parametric	
Hemi-boreal	40.4	42.3	45	44.6	66.1	
South-western	18.7	18	18.2	18.2	61.5	
Lake District	14.1	13.9	14.3	13.8	36.7	
S. Ostrobothnia	9.8	10.6	12.5	13.4	35.6	
Ostrobothnia	21.2	21.3	20.7	20.4	56.8	
Kainuu	20	21.5	21.6	21.5	38.2	
S. Lapland	19.8	20	21.1	20	48.7	
Forest Lapland	18.4	14.5	16.3	19.1	44.9	
Whole Finland	21.5	21.6	22	21.6	53.8	

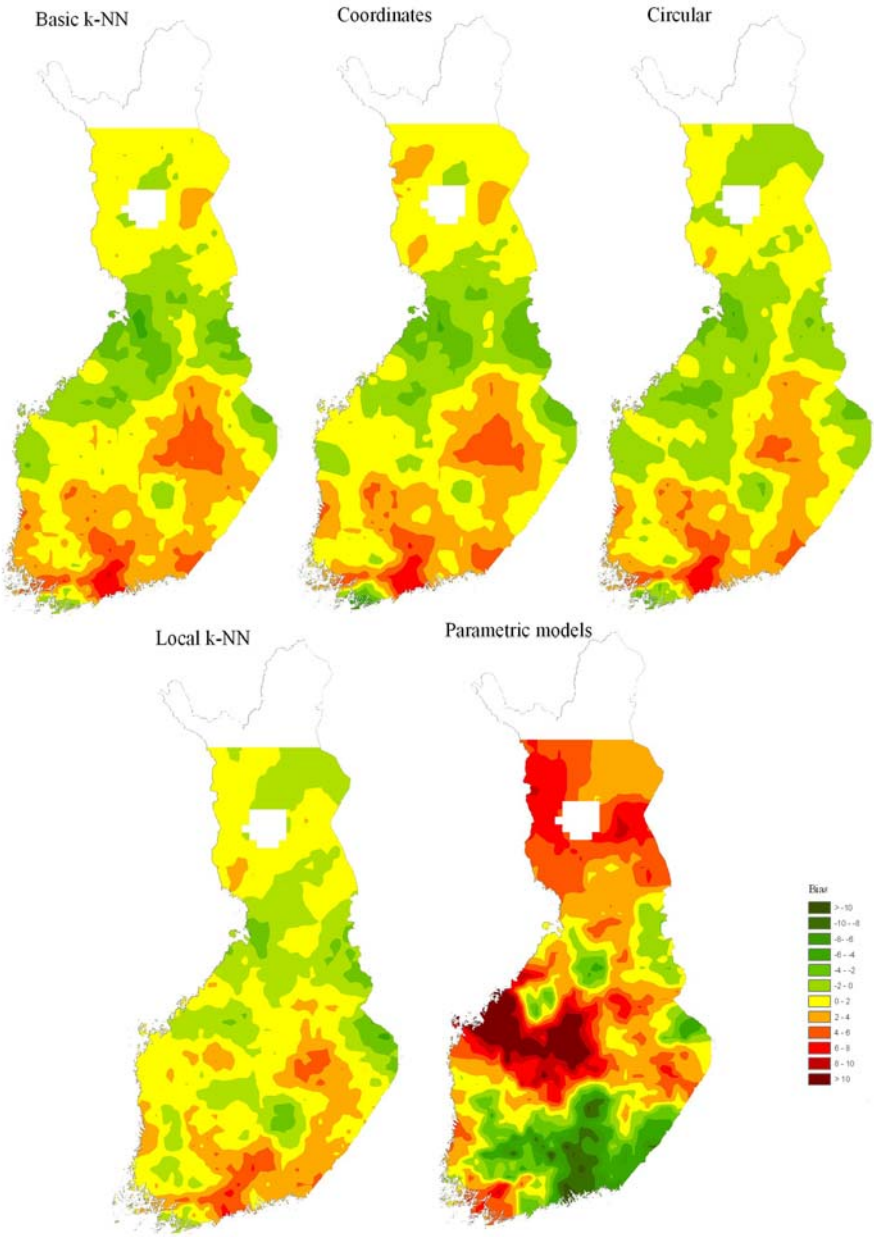

Figure 4. Biases of the stand-level growth estimates (m^3ha^{-1} in 5 yrs) produced by the different k-NN methods (basis non-spatial k-NN, separate local k-NN method, localized k-NN with coordinates and localized k-NN with neighbours selected from circular areas around the target tree). Green colour represents overestimation and red underestimation.

5.4 Simultaneous estimation of individual tree diameter and height increment (IV)

The simultaneous estimation was performed using k-NN method with distance function based on canonical correlation analysis. Best accuracy for both Scots pine and Norway spruce was achieved with the same independent variables used in the imputations, including DBH, tree crown ratio, BAL, H_{dom}, BA, temperature sum, and forest site type. The relative RMSE of the Scots pine diameter increment estimates was 49.77% and height increment estimates 50.69%, the corresponding measures for Norway spruce being 50.69% and 68.39% (Table 3 below, Table 3 of paper IV). Diameter and height increment estimates were almost unbiased for both of the tree species.

The results of the k-NN method were compared with the parametric method both at tree- and stand-level. The parametric models produced larger biases for the diameter and height increment estimates, especially for Scots pine. The relative biases were about 30% for the parametric estimates, while those for the non-parametric method were nearly zero. For Norway spruce, the bias obtained with the parametric models was 6% for the diameter increment estimates and 20% for the height increment estimates. In addition, the RMSEs were somewhat larger for the parametric growth estimates, except that the RMSE of height increment estimates of Norway spruce was somewhat lower than what was achieved with the k-NN method. The distributions of the growth estimates were calculated as well; and the results showed that the distributions of the k-NN estimates were more similar to the observed ones than the distributions of the parametric growth estimates (Fig. 1 of IV). Moreover, the parametric method seemed to average the results more than the k-NN method.

Table 3. Accuracy of the five-year tree-level and stand-level growth estimates obtained by the simultaneous k-NN method and parametric method with the INKA (IV) data and Kuusamo data (V).

	INKA					**Kuusamo**				
	Pine		Spruce		Stand	Pine		Spruce		Stand
	i_{d5},cm	i_{h5},m	i_{d5},cm	i_{h5},m	I_{V5},m^3 ha^{-1}	i_{d5},cm	i_{h5},m	i_{d5},cm	i_{h5},m	I_{V5},m^3 ha^{-1}
k-NN method										
Bias	0	0	0	0	0.3	-0.1	-0.1	-0.3	-0.3	-3.6
Bias,%	0	0	0.1	-0.3	1	-12.5	-12	-57	-76.8	-27.9
RMSE	0.6	0.5	0.6	0.6	11.2	0.7	0.4	0.6	0.6	6.9
RMSE,%	49.8	50.7	56.9	68.4	34.7	70.7	52.2	100.3	125.7	54
Parametric models										
Bias	0.4	0.3	0.1	0.2	8.7	-0.3	-0.2	-0.4	-0.3	-5
Bias,%	31.8	28.3	6.1	20.3	26.8	-31.3	-20	-67	-60.8	-39.1
RMSE	0.8	0.6	0.7	0.6	15.3	0.7	0.5	0.6	0.4	7.5
RMSE,%	60.4	56.2	61.9	62.5	47.3	67.5	55.4	102.1	98.5	58.7

The relative RMSE of the five-year stand-level volume growth estimates obtained with the k-NN method was 34.7% (Table 3 above, Table 3 of IV). The relative RMSE produced by the parametric models was about 12% larger. In addition, the parametric models underestimated the volume growth at the stand level considerably. The relative bias of the non-parametric estimates was 0.95%, while it was nearly 27% for the parametric estimates. Similarly to the tree-level results, the parametric volume growth estimates concentrated notably to the smallest volume growth classes, while the k-NN method produced distribution more similar to the observed one and was capable of producing larger volume growths as well (Fig. 1 of IV).

5.5 Effects of reference data selection on the accuracy of non-parametric growth estimation (V)

The methods developed and results obtained in papers III and IV were combined in the fifth paper (V). The simultaneous k-NN estimation method was evaluated against independent test data. This evaluation was carried out by using Kuusamo data as the target data for which the diameter and height increment estimates were imputed, and the INKA data as the reference data. The accuracy of the k-NN method was noticeably poorer than the results obtained merely with the constructing data from INKA. In particular, the RMSE and bias of the diameter and height increment estimates of Norway spruce were notably larger (Table 3 above, Table 3 of paper V). The diameter and height increments of the spruce trees were highly overestimated and the relative RMSEs of the estimates were over 100%, while they were almost unbiased with the INKA data, and the relative RMSEs were 56.9% and 68.4%, respectively (Table 3 above, Table 3 of IV). The results were not as poor for Scots pine, although the RMSE of the diameter increment estimates was noticeable and the method overestimated both diameter and height increment in Kuusamo. Furthermore, the stand-level volume growth was overestimated by nearly 28% and the relative RMSE was 54% (Table 3 above, Table 3 of V). These results were further compared with parametric estimates constructed within the SIMO system for the same data. Although the difference in accuracy between the non-parametric and parametric methods were generally smaller than with the INKA data, the parametric models still produced somewhat more biased estimates both at tree and stand level, except for the height increment of Norway spruce (Table 3 above, Table 3 of V). In addition, the RMSEs were of the same magnitude for both methods, except that the RMSE of the five-year height increment estimates of Norway spruce were markedly larger for the k-NN method.

Local and localized k-NN methods, as well as the effects of local observations, were further tested in Paper V. The target data in this analysis was the Kuusamo data and the reference data was either INKA, with increasing amount of local observations included, or merely the local observations from the Kuusamo data. First, the basic non-spatial k-NN method developed in paper IV was applied and an increasing amount of local observations were added amongst the INKA data. The accuracy of the growth estimates improved, especially for Norway spruce, by incorporating local observations from just ten stands into INKA data (Table 4 of V). However, the accuracy did not improve notably after including local observations from the first 40 stands. The number of local observations among the selected nearest neighbours did not markedly increase with increasing amount of local observations in the reference data (Table 5 of V). The results were markedly better when only local observations were used as the reference data, even with a low number of

measurements. In particular, the biases of the growth estimates decreased both at tree and stand level (Table 4 of V).

Localizing by including coordinates as auxiliary independent variables and sub-setting the reference data by selecting the neighbours from a circular area around the target tree were tested. Proportions of the local nearest neighbours used in forming the estimates were calculated in order to find out which of the methods selected the most local observations as nearest neighbours. The k-NN method with coordinates produced better results than the basic k-NN method. The accuracy of diameter and height increment estimates for Norway spruce especially were improved (Table 4 of V). In addition, the bias of the stand-level volume growth estimates was notably smaller. The increase in local observations in the reference data did not have any marked effect on the accuracy of this method, but the accuracy increased somewhat more than with the basic k-NN method. The proportion of the cases where no local observations were used in forming the estimates was, on average, lower than in the basic k-NN method. However, the proportion of the local observations used as nearest neighbours seemed not to depend greatly on the amount of local data.

The accuracy of the k-NN method localized by sub-setting the reference data into circular areas around the target tree was in most cases at the same magnitude as the accuracy of the local k-NN method based on local reference data in terms of the RMSE of the growth estimates. Local k-NN method produced smaller relative RMSEs for the diameter increment estimates of Scots pine and height increment estimates of Norway spruce. The k-NN method localized by sub-setting the reference data produced less biased estimates for Scots pine than the local k-NN method, but more biased estimates for Norway spruce. The localized method overestimated the diameter and height increment of spruces markedly at first, and small biases of these increment estimates were not achieved until local observations from 60 stands were included in the reference data. The stand-level volume growth was also overestimated at first. The proportion of local neighbours used in forming the estimates with this localized method increased notably the more local data was available. When the reference data included local observations from 71 measured stands, all of the estimates were constructed using seven local nearest neighbours.

The results of the local k-NN and localized k-NN by sub-setting the reference data showed that the RMSE and bias diminished rapidly at first, but after including measured observations from 40 stands, no marked improvements to the results were achieved. The variances of the variables in the reference data stabilized to the same magnitude as they were in the target data. Furthermore, the composition of the rarest age classes did not change notably after including local observations from 40 stands. However, the order of the local observations included might have effects on the accuracy as well. The latter method was therefore tested by including the stands by starting from the oldest ones instead of randomly including them. The relative RMSE were smaller at first, but the bias somewhat larger. It would have been appropriate to examine the target and reference datasets in more detail, and include observations from the stands that affect the accuracy most. Thus, observations from the sparse age and site type classes should have been included evenly.

5.6 Comparison of different non-parametric growth imputation methods in the presence of correlated observations (VI)

The performance of two different nearest neighbour methods employing either Euclidean distance function (k-EUC) or distance function based on canonical correlation (k-MSN) and generalized additive models in the presence of correlated observations were analysed in the sixth paper (VI). The analysis was carried out by restricting the pool of possible nearest neighbours. Otherwise, the independent variables used in this analysis for all the different non-parametric methods were the same as in the nearest neighbour method in the paper III. Most accurate tree-level growth estimates were obtained without restrictions in all of the non-parametric methods (Table 4 below, Table 4 of paper VI). The results were very similar for both tree species, although the accuracy in general was poorer for Norway spruce. The intraclass correlations at different levels were similar for both tree species. The restrictions had most effect on the nearest neighbour imputation with non-weighted Euclidean distance function, especially when Scots pines were considered. The k-EUC method had equal weights for all of the variables hence the nearest neighbours were mainly selected on the basis of stand-level variables.

Table 4. Accuracy of the five-year tree-level diameter increment and stand-level volume growth estimates obtained by the different non-parametric method and restriction alternatives (Paper VI).

	Scots pine		Norway spruce		Stand-level I_{V5}	
	Bias,%	RMSE,%	Bias,%	RMSE,%	Bias,%	RMSE,%
k-MSN						
NoRestrictions	-0.45	60.45	-0.53	65.73	1.8	16.2
PlotRestriction	0.04	67.77	-0.17	71.96	2.5	18.9
StandRestriction	0.12	69.04	0.1	72.84	2.7	19.8
OnePerPlot	-9.86	69.88	-13.99	79.36	-2.2	20.2
OnePerStand	-8.69	69.52	-17.02	80.64	-2.5	20.4
k-EUC						
NoRestrictions	0.22	53.53	-0.1	61.25	1	2.8
PlotRestriction	0.27	56.88	-0.39	64.36	0.8	4.3
StandRestriction	-2.89	72.57	0.51	76.9	0.7	23
OnePerPlot	-6.82	66.29	-11.09	76.42	-4.3	21.6
OnePerStand	-8.95	65.8	-13.09	76.34	-5.5	21.3
GAM						
NoRestrictions	0	56.11	-0.02	62	1.6	12.1
PlotRestriction	-0.12	58.17	-0.4	64.45	1.4	13.8
StandRestriction	-1.11	65.38	-0.8	72.12	1	21.9
OnePerPlot	-2.42	60.87	-7.47	70.8	-0.6	19.2
OnePerStand	-2.31	61.3	-9.35	71.94	-1.1	19

While applying the k-EUC method without restrictions, only for 1% of the target trees were none of the nearest neighbours selected from the same stand as where the target tree was situated. This was about 15% for the k-MSN method without restrictions. Hence, the k-EUC method tended to produce growth estimates on the basis of different sized trees in the same stand, while k-MSN method more often selected trees of the same size from different stands. This was also the case with the stand restriction alternative, the k-EUC method tended to select the neighbours from one particular stand, which caused the poorest accuracy for the k-EUC method with stand restriction. Although the relative weights of stand-level and tree-level variables were fairly similar in both of these methods, the k-MSN method placed greater weight on tree diameter in the distance function of Scots pines and on BAL in the distance function of Norway spruce. The effected of this was that even without restrictions, the nearest neighbours were more often selected from different stands in the k-MSN method.

The biases of the diameter increment estimates were considerably larger when only one tree from one stand or plot was allowed to be included in the group of nearest neighbours. This was mainly due to large overestimations of diameter increments in stands with zero or very low five-year diameter increments for all trees. The estimates were more likely to be formed on the basis of neighbouring observations having larger five-year diameter increment when this kind of restriction was applied, and thus growth was heavily overestimated. In general, supposing that the least restricted alternatives, i.e., no restrictions or plot restriction, were ignored, the generalized additive models performed most accurately and produced the smallest RMSEs and biases of the tree-level growth estimates.

The k-EUC method performed least accurately at the stand level supposing that the neighbours from the same stand were excluded. The k-EUC method with stand restriction tended to select all the neighbours from one particular stand, and if that stand is situated, for example, in a different forest site type than the target tree stand, the errors of the individual trees might all point in the same direction, thus diminishing the accuracy of the growth predictions at the stand level. The best accuracy at the stand level was achieved with the GAMs and with one per stand restriction, the relative RMSE being then 19%. However, the differences in accuracy between the methods were fairly minor with these restricted options. In the k-MSN method the biases with plot and stand restrictions were larger than the biases of the one per plot and one per stand restrictions, although the former alternatives were almost unbiased at the tree level and the latter overestimated the individual tree growth markedly.

5.7 Performance of the non-parametric methods under different growing conditions (I–VI)

The behaviour of the non-parametric methods was analysed under different conditions including the performance at the data ranges and within stands. In particular, the residuals of the non-parametric five-year diameter increment estimates were plotted as a function of the predicted value, and the relative biases in various diameter classes were calculated. The residuals of the tree-level growth estimates were mainly homogeneous and no obvious dependences between them and the predicted values were found. The relative biases were found to be somewhat notable at the boundaries of the data range, especially for large trees, which were presented by small amount in the data (e.g. Fig. 4 in paper I, Figs 3 and 4 in paper III). In order to reduce the bias at the boundaries of the data, different numbers of

nearest neighbours for small and large trees were tested in papers II and III while performing the k-MSN imputations; however, no improvements were achieved.

When the behaviour of the methods within stands was analysed (IV and VI), the results indicated that the different non-parametric methods produced fairly logical growth estimates for both tree species in relation to the position of a tree in a stand. Diameter increments produced by the k-NN methods were smaller for the suppressed trees than for the dominating trees, and in many cases the growth reached the maximum at some point and diminished after that (Fig. 2 of paper IV and Figs 1 and 2 of paper VI). The GAMs produced more averaged results owing to the substantially larger number of nearest neighbours included in the calculations. The GAMs in many cases seemed to overestimate growth either for the suppressed or dominated trees, while underestimating it at another extreme of the data. Height increment estimates produced by the k-NN method were fairly consistent irrespective of the position of tree in a stand, although mainly smaller for the suppressed spruce trees than for the dominating spruces. However, the true relationship between height growth and relative tree size varied quite considerably. In addition, height growth differentiation among trees was higher in younger stands, which can be seen especially in the spruce-dominated stands.

The behaviour of the non-parametric methods at stands of various ages and densities were also compared in each paper by plotting the residuals of the stand-level volume growth against stand age and stand basal area. The residuals of the non-parametric estimates were mainly evenly scattered and showed no obvious trends (e.g. Fig 5 of paper I and Fig. 5 of paper III). In paper III, the residuals of the stand-level volume growth estimates produced by the k-NN method were homoscedastic in every vegetation zone, while those for the parametric method were heteroscedastic in most of the regions.

The performance of the k-NN method and parametric models were analysed in various site types in paper IV by calculating the averages and standard deviations of the growth estimates within each forest site type group. In general, the standard deviations of the growth estimates were smaller for the k-NN method only in those site type classes which included very small numbers of observations (Fig. 3 of paper IV). Thus, the estimates had to be formed on the basis of nearest neighbours from another site type than the type in which the target tree was situated.

The accuracy of the growth estimates in various site type groups was also calculated. In addition, these were calculated by excluding temperature sum (TS) and forest site type group (FST) from the set of independent variables while performing the imputations. The accuracy of the k-NN method was low in those classes where the data was sparse (Figs 4 and 5 of paper IV). Considering Scots pines, these were mainly the most fertile classes. For Norway spruce, these were very rich and barren sites, as well as drained peatlands. Otherwise, the k-NN produced more accurate results in various site types than the parametric method. The biases of the parametric models especially were larger than those of the k-NN method. At the stand level, the k-NN produced larger RMSE only in the most fertile class; however, the bias was still slightly smaller than the bias of the parametric models. Furthermore, the accuracies were calculated by excluding temperature sum and forest site type from the set of independent variables while performing the imputations. In general, forest site type had more notable effect on the accuracy than temperature sum. Both of them had similar effects on the accuracy of the growth estimates. Excluding either of them diminished the accuracy at some forest site types, but increased it at other site types.

Future forecasts obtained with the k-NN method and parametric models were compared in the fourth paper (IV). The analysis was carried out in four different forest site types. Two stands situated in the damp forest site were selected. One of these stands was dominated by Scots pine, the other by Norway spruce. The rich site was dominated by Norway spruce and the dry sites by Scots pine. The tree-level results showed that reasonable long-term diameter and height increment estimates could be obtained with the k-NN method, although some differences between the k-NN method and parametric models were achieved. Furthermore, the lack of real data for longer growing periods restricted the analysis to merely comparing the methods in question to each other and at different forest site types. In addition, mortality of the trees or self-thinning was not taken into account while performing the imputations. For the first growing periods, there was more variation among the different sites within the k-NN method than within the parametric method (Fig. 6 of IV). After 30 years growing period, the diameter increments became the same for all of the methods. The k-NN and parametric methods differed mostly in that the parametric models produced larger diameter increments for the pines in damp sites than the k-NN. The results indicated also that the k-NN method was capable of taking the effect of thinning into account both implicitly and explicitly. In general, the diameter increments increased immediately after thinning (Fig. 7 of IV). Responses to thinnings were largest in damp sites. The responses were somewhat smaller when an additional thinning dummy was included. The height increment increased owing to thinning; however, not immediately after thinning (Fig. 8 of IV). The increase in height increment was most noticeable in fertile, spruce-dominated stands.

The variation in accumulated volume among the test sites increased markedly, and more in the parametric method than in the k-NN method. The inclusion of self-thinning models in the parametric method influenced the accumulating volume markedly in Scots pine-dominated stands producing noticeably smaller volumes than the k-NN method. In addition, the development of rich sites ended up at a higher level with the parametric models than was achieved with the k-NN method (Fig. 9 of IV). Moreover, the production capacities of the different sites ended up in a different order. The results of the k-NN imputation with simulated thinnings were calculated at the stand level as well (Fig. 10 of IV). The response to one thinning at the stand-level was largest in the rich and damp spruce-dominated stand. The response was larger without the explicit thinning variable, while it was opposite in the Scots pine-dominated stands. The responses to the second thinning were not noticeable. The level of growth was smaller after the second thinning in every stand for the rest of the forecasting period.

6 DISCUSSION

6.1 Different non-parametric methods

The purpose of this thesis was to test and compare different non-parametric methods in estimating individual tree growth. The different methods were compared to each other as well to parametric models both at tree and stand level. Additionally, regional level comparisons were made, since one of the main objectives was to reduce the regional level bias associated in the growth estimates and obtain regionally unbiased estimates. Two different approaches were applied, namely, k-nearest neighbour methods and generalized

additive models. Both of these approaches could produce accurate growth estimates, although no method was superior in every condition, thus the question about the most accurate non-parametric method still depends on the purpose and the data used in the imputations. The generalized additive models require enough variation in the data to perform well, in small datasets with independent variables having low variation, the method performed poorly (paper II). However, the performance of all the non-parametric methods is greatly dependent on the data used, since the idea of the non-parametric methods is to associate the previously measured information to the estimation of the chosen characteristics for the target tree (e.g. Malinen 2003). If the reference data does not contain similar trees, then the estimates for the target trees may be inaccurate as was the situation when the Kuusamo data was used as target data and the INKA data as reference data (paper IV). However, the difference in measurement methods applied to collecting the INKA data and Kuusamo data may have caused differences in the datasets that caused the poor results and require the use of local observations.

In addition, the number of the neighbours used in the non-parametric methods is of critical importance. The choice of the optimal number can be somewhat problematic and the data-driven methods used in the decision may not provide an unambiguous solution for the neighbourhood size. In general, using several neighbours may improve the estimation accuracy, but the results are more averaged, and especially the bias of the extreme observations may increase (e.g. McRoberts 2002). Concerning growth estimation, one nearest neighbour was not enough, and on the other hand, a large number of neighbours did not improve the results. The number of the neighbours in the final k-NN imputations was set to be the number where the decrease in relative RMSE stabilized, and the bias was as low as possible. The neighbourhood size was more critical to the k-nearest neighbour methods, while it did not notably affect the accuracy of generalized additive models. The neighbourhood size is substantially larger while applying GAMs in any case. Thus the generalized additive models gave estimates that were precise in average, while being more likely biased at the extremes of the data range. Nearest neighbour imputations seemed to retain more of the natural variance in the growth estimates. Furthermore, the generalized additive models may be more difficult to understand and to implement than the nearest neighbour methods.

There are several functions to be applied in the search for the nearest neighbours in both approaches. Locally weighted regression was found to be a more suitable smoothing function than splines in GAMs. Other possibilities exist as well, although they are not yet implemented in statistical software packages. Semi-parametric methods could also be applied. An appropriate parametric model could be used for the terms that have a clear relationship with the dependent variable, while modelling the other terms non-parametrically (e.g. Opsomer 2000a). Non-weighted distance functions or functions with weights for the variables obtained using grid search are more robust. The weights are acquired by minimizing the error one is interested in, not by maximizing the overall correlation. The weighting based on linear regression or canonical correlation, on the other hand, is based on assumed linearity between the variables. Compared with grid search, the linear regression and canonical correlation are computationally fast. Linear regression and canonical correlation give equivalent weighting, when there is only one dependent variable, as was in the most cases of the papers. However, when simultaneous estimation of several variables, for example, diameter and height increment and thickness of the bark, is considered, the distance measure based on canonical correlation might be a more appropriate alternative. The use of Mahalanobis distance with weights derived from linear

regression for more than one dependent variable, when these are correlated, is probably not a good choice. There are other possibilities to obtain weights, and especially non-linear or genetic optimization algorithms, as applied by Haara et al. (2002) and Tomppo and Halme (2004), would be worth testing in the context of growth estimation as well. In addition to barely implementing the optimization algorithm, other non-parametric methods could be tested. Artificial neural networks or random forests could be implemented in growth estimation for Finnish conditions as well. The results of Liao et al. (1998) indicated that applying Neurogenetic Algorithm System it is possible to simulate individual tree growth effectively and to improve markedly the quality of growth predictions. The method was also expected to perform better for future data.

6.2 Dependent and independent variables

The non-parametric methods, in most cases, were constructed by using tree diameter increment as a dependent variable. However, the simultaneous estimation of diameter and height increment with the k-NN method proved to be suitable growth estimation method as well. The simultaneous estimation did not produce more accurate results than a separate estimation of diameter and height increment, however, it did not diminish the accuracy either. The same independent variables were required in both diameter and height increment estimation, therefore no marked difference was achieved. The estimation of height increment of spruces seemed to be somewhat more problematic, thus producing the most inaccurate results. The correlations between height increment of Norway spruce and possible independent variables were quite low. It correlated moderately well only with crown ratio and stand age, the latter of which was not included in the imputations. The variation in height increment was large, since it was calculated as difference between two successive height measurements. Measurement error in re-measured heights on standing trees may be so large that the underlying height increment signal is nearly hidden (e.g. Hasenauer 2006). The results would have been better if the stand age was included; however, it was decided to apply only those variables that are allowed to be used in practical situations. In other papers, the stand age was included as an independent variable, and therefore the methods developed have certain limitations in practical situations where the methods should also be applicable in uneven-aged stands. Furthermore, the measured age includes large error, and growth models are sensitive to erroneous age (e.g. Haara and Korhonen 2004). However, the measured stand age was determined to be more suitable as an independent variable – especially in papers III and VI, in which the methods were applied to the whole tree tally instead of sample trees –, than crown ratio, for which a predicted value would have been needed. Moreover, the reference data included trees only from even-aged stands. Otherwise, the crown ratio might be more suitable as an independent variable, also giving more relative weight for the tree-level variables in the distance function.

Furthermore, forest site types were tested as independent variables in every paper, but were employed only in papers IV and V, since it diminished the accuracy of the non-parametric method in most cases. The variation in these kinds of stand-level variables might be too low in small local or regional data. Correlation between diameter increment and forest site type is low, hence it may not contain much weight in the distance function, if the weights are obtained through linear regression or canonical correlation analysis. Furthermore, unrelated variables included in the subset of covariates used to calculate

distances may not only fail to improve the objective criterion, but actually may have adverse effects (McRoberts et al. 2002). Thus inclusion of additional independent variables does not necessarily improve the results of non-parametric methods, and may cause that nearest neighbours are even more difficult to find especially in small datasets (e.g. McRoberts et al. 2002, LeMay and Temesgen 2005).

However, even though the site factors account for a small percentage of the variation, they are important, and serve to localize a particular prediction (Monserud and Sterba 1995). It is necessary to include variables describing the fertility in order to guarantee that the non-parametric methods do not produce similar growths for trees under different growing conditions and average the estimates over different fertility classes, although the overall accuracy may diminish as well as the accuracy at some extremes of the data. Site index might be a more appropriate independent variable than forest site type and should be tested. However, site index measurement errors have also created some of the largest variations in predicted values (e.g. Gertner and Dzialowy 1984, Mowrer and Frayer 1986, Gertner 2002).

More accurate temperature sum information and variables describing slope and moistness, for example, could be tested in order to describe the local growing conditions more accurately. As remote sensing techniques are nowadays playing a more important role in forest inventories, these variables could be obtained from remotely sensed data. Remote sensing techniques might introduce new variables to be used in non-parametric growth estimation as well. Moistness could be derived from radar, and slope from a digital terrain model (DTM) produced by airborne laser scanning (ALS) data. Furthermore, a variable describing crown condition derived from the vertical point cloud of ALS data could yield some new information.

6.3 Effects of correlated observations

The effects of correlated observations should be taken into account and carefully analysed while applying non-parametric methods. The results showed that the dependency of observations did not have any marked effect on the selection of the best possible neighbourhood size in any method. Otherwise, the papers I and VI revealed similar effects of correlated observations, although the study data, the dependent variables and the distance functions used were somewhat different. The correlated observations from one particular stand, excluding the stand where the target tree is situated, may have the result that the tree-level errors are also correlated, and thus diminish the accuracy at the stand level. Stand-level bias may result from selecting the neighbours from a stand that has different kinds of stand-level factors, for example, is situation in a different forest site type than the target tree stand.

Including many reference trees from the same stand was an inefficient strategy. It did not improve the stand-level and regional results if it was not the same stand in which the target tree was situated. Furthermore, the k-MSN method with the optimal allocation of weight based on canonical correlations seemed to be safer from the effects of dependency, while the k-NN method of applying non-weighted Euclidean distance seemed to be more sensitive to the tree-level dependency. Additionally, the generalized additive models were not affected by the restrictions as much as the other methods. The generalized additive models, like linear models in general, perform better if there is enough variation in the data, and the restrictions ensure the variation.

The results of paper VI indicated that allowing only one tree per plot or stand to be included in the set of nearest neighbours would be appropriate when considering the accuracy of stand-level or regional growth estimates. However, the data used with these kinds of restrictions should be extensive enough, so that the restrictions do not make the search for the nearest neighbours even more difficult. This result is similar to the recommendations of McRoberts et al. (2007). They suggested that only one nearest neighbour should be permitted within the range of spatial correlation of other neighbours when obtaining areal estimates of forest attributes using k-NN approach and the ranges of spatial correlation are small and the reference set is relatively large. Another option, concerning mainly distance functions that do not guarantee optimal solutions for weights, would be to include more tree-level variables in the distance functions or to give more weight to the tree-level variables. The inclusion of variables describing fertility and local growing conditions to the search for the nearest neighbours may also diminish the amount and effect of correlated tree-level errors on the accuracy of stand-level volume growth, at least in those situations where the neighbours are selected from one stand in a different site. Furthermore, the dependency of the observations could have been taken account by giving less weight to the observations that are situated in the same stand (Altman 1990). Two neighbours from one stand would attract less weight per observation than two observations from different stands. Obviously, this method only works if there are neighbours from more than one stand, and the number of neighbours per stand varies.

6.4 Localization of the non-parametric growth estimates

The difference between the basic k-NN method and the localized methods was very small in paper III, and they were in general quite similar in terms of their performance in different vegetation zones. Localization by including physical space in the k-NN method did not notably reduce the regional biases relative to the basic k-NN method, which was in any case able to find nearest neighbours similar enough. The basic k-NN method had the temperature sum as an independent variable, which may be seen as a form of localization and may reduce the difference. Furthermore, it utilizes the whole data, and better matches may usually be found with increasing sample size. It might be hard to find neighbours, at least for the exceptional observations, if the number of possible candidates is reduced, as could happen in localized methods owing to a larger number of variables involved in the search for neighbouring observations or a reduction in the search area. Both of the applied localized k-NN methods produced somewhat larger biases than the basic k-NN in dense stands in most of the regions, and the biases for the exceptional observations in the hemi-boreal zone, Kainuu and Lapland were also larger for the localized methods.

The most promising alternative to the means of localization was the sub-setting of the reference data by selecting the neighbours from a circle around the target tree, as the results obtained with this method were better than those achieved with the basic k-NN method in most cases, even though there were no major differences. Both Tokola (2000) and Katila and Tomppo (2001) applied similar geographical reference areas. The former studied the maximum geographical distances of training areas to be used in obtaining accurate point and small-area estimates, the latter studied a similar method to be used in Finnish National Forest Inventories. According to both studies, the bias of the estimates could be reduced by restricting the neighbourhood. The inclusion of data beyond the optimal area introduced bias to the estimates. The results of paper III showed that this method produced the largest

biases of the diameter increment estimates in the coastal zones, where the data in a circle around the target tree was usually locally distorted and possessed gaps, in addition to which the number of possible neighbours was smaller, because in most cases only half of the circle included possible neighbours. Neighbouring observations had to be selected from further inland and the difference in diameter increment among the trees of the same size in coastal and inland areas could be considerable. This is a problem affecting the method at every border, but the difference in growth is not so large in eastern Finland or Lapland. More accurate results might be obtained if geographical areas of different size and shape were used in inland sites and at borders rather than average-sized circles everywhere.

In addition to using plot centre coordinates and sub-setting reference data in circular neighbourhoods, separate local k-NN imputations for each region were performed in three ways from a regional database. Searching for the regionally optimal values and dependent variables did not improve the accuracy of the regional growth estimates and was therefore unnecessary. All the critical variables were already included in the basic k-NN and the effects of the other variables were minor. Furthermore, the lack of any marked differences between the local estimates with a regional or whole data weighting matrix indicates that the correlations between the variables used are quite similar over the whole of Finland and in the individual regions. At the tree level all the separate local k-NN methods produced almost unbiased estimates for the diameter increment in the various regions, while the bias of the estimates of the basic and localized k-NN based on the whole data varied quite considerably across the regions. However, when comparing the accuracy of the estimates of stand-level volume growth regionally, the differences in the biases among different methods were small. The variation within stands might be larger than between stands when viewed regionally. Nevertheless, it seemed to be unnecessary to construct local k-NN estimates from regional data. This result is similar to the observation of Maltamo et al. (2003), who predicted diameter distribution with regional MSN models and found that the local variation could not be described any more accurately with regional models. In any case, running the k-NN estimates separately from local data might be too laborious a process. Moreover, separate local estimation may produce unnaturally large differences among predictions for nearby regions, i.e., the growth estimates for stands on two sides of a border might be too dissimilar, given that the stands are physically located close to one another. There could also be gaps in the coverage of the models, as in the case of southern Ostrobothnia.

All the k-NN methods produced promising results in terms of reducing regional biases compared with parametric models. At least, the regional biases in northern Finland and south-western Finland were reduced substantially with the k-NN methods. The biases of the k-NN estimates in all the regions were close to each other, while the differences in bias across the regions with respect to the growth estimates obtained with the parametric model were over 25%. However, the regional biases associated with the parametric method could have been reduced by calibrating the models, for example, by constructing models for the bias and then calculating the calibrated prediction by adding the predicted bias to the initial prediction (Hynynen et al. 2002, Talvitie 2005). One option might also be to take the hierarchy of the data into account more properly in the modelling. The model predictions might be calibrated by producing random parameters for Forestry Centres or provinces, for example. So far the models include random parameters for stands, which usually cannot be used for calibration, since the models are not applied in the stands of the modelling data.

Local and localized k-NN methods were further tested in the fifth paper (V), since the basic non-spatial k-NN method performed poorly when tested against independent

Kuusamo data. All the localized methods performed better than the basic k-NN method. When conditions are rather exceptional as was in this case, the need for local observations seemed to be obvious, although the different measurement methods concerning the reference and target data may have created the need for local observations. Firstly, the basic k-NN method was tested, but measured local data was added amongst the INKA data, which was used as reference. This improved the results of the basic k-NN especially with respect to Norway spruce. However, the results were not substantially improved by increasing the amount of local observations among the reference data. The variables that had most weight in the distance function did not contain any information about the location or low growing rate. Although the temperature sum contained quite notable weight in the distance function and the site types were also included, the local observations were not selected as nearest neighbours. Therefore some auxiliary variables, like geographical location, might be useful in this kind of situation. Localization by including coordinates as auxiliary variables was thus tested, and this method produced better results than the basic k-NN method. However, the increase in local observations in the reference data did not have any marked effect on the accuracy of this method either, but the accuracy increased somewhat more than the accuracy of the estimates constructed with the basic k-NN method.

When the estimates were formed by applying only the local observations as the reference data, the results were substantially improved even with a small amount of local observations. The bias of the tree-level diameter and height increment estimates of Norway spruce, especially, diminished markedly. The results of the localized k-NN by sub-setting the reference data were mainly as accurate as those obtained with the local data only. The method emphasized the local observations in constructing the growth estimates. The more local data were available, the more local observations were selected as nearest neighbours, and therefore the accuracy of the growth estimates increased. This method was able to take advantage of the local observations without losing information from the observations in the INKA data. Thus this kind of localization method seems to be a suitable alternative when constructing growth estimates for a local area, where comprehensive reference data is already available but improved accuracy is can be obtained through local observations.

The vegetation zones might not have been the optimal areas to study the localization methods or to construct local k-NN estimates, since the results were improved in the Kuusamo area, but not when the performance was analysed by vegetation zones. The areas should be formed in such a way that the within-group variation is as low as possible and the between-group sub-area variation is as large as possible (e.g. Tomppo and Halme 2004). Some ancillary regional information could help in selecting the localization areas. Tomppo and Halme (2004) used large-scale variation of forest variables as ancillary data that were added to the variables of the multi-source k-NN estimation. Räty and Kangas (2007), on the other hand, tested the local indicators of spatial association in the selection of localization areas. Methods of these kinds could be used in selecting more optimal areas to study the localization than the vegetation zones used here.

6.5 Concluding remarks and need for future research

This thesis focused on nearest neighbour methods and generalized additive models, and different issues concerning growth estimation when these kinds of non-parametric methods are applied. The thesis responded to the need to reduce the regional biases associated with the growth estimates, and showed that non-parametric methods are suitable for estimation

of growth. The most accurate results were achieved when the imputations were carried out by including sufficiently weighted tree-level variables in the distance functions or by taking account of the correlated observations by restrictions. Moreover, geographical location and variables describing the site improved the results, especially in extreme conditions. In addition to being able to reduce the regional biases, the non-parametric k-NN methods did not average the results as much as the parametric methods. The k-NN methods were capable of producing distributions of increment estimates rather similar to the observed distributions, while the parametric increment estimates both at tree and stand level were mainly concentrated on the smallest increment classes. However, sparseness of the data caused some averaging. The non-parametric k-NN methods also produced more accurate estimates for different forest site types and retained more of the variation in the growth estimates, except in those site types were the data was sparse.

Thus, the non-parametric methods fulfilled the requirement that growth models used in practical forest management planning should produce unbiased predictions of the development of forest resources (e.g. Hynynen et al. 2002). The methods applied here are also compatible with forest inventory data. The main applications of the growth models include inventory updating, evaluation of silvicultural alternatives, management planning and harvest scheduling (e.g. Burkhart 1992). As such, the methods applied here are mostly applicable to inventory updating and management planning at smaller scale. They could provide locally accurate estimates for determining locally correct silvicultural treatments. Forest owners could improve their management plans by measuring data from their forests. However, inventory updating and especially the evaluation of alternative management schedules requires that the growth models are capable of predicting the responses to various silvicultural treatments (e.g. Hynynen et al. 2002). The k-NN method applied here was able to take the response to thinning into account both implicitly and explicitly. The non-parametric methods were also capable of predicting the responses to other silvicultural treatments; however, these issues require further testing with adequate data as well as more consideration of how to incorporate them into non-parametric methods. Growth models used for practical applications should be carefully tested in stands with various management and thinning conditions. The non-parametric methods performed well in stands of various ages and densities, without producing more biased estimates at the extremes of the data. In addition, the non-parametric methods produced mainly logical diameter and height increment estimates in relation to the position of a tree in stands of various ages.

Models used for management planning and forest policy analysis should behave reliably and logically when applied in long-term simulations (e.g. Hynynen et al. 2002). Non-parametric methods were also a suitable method for forecasting growth over longer periods, although testing revealed some differences compared with parametric models. The largest differences occurred because self-thinning was not included in the k-NN method. The self-thinning models had a large effect on the predicted development of pine-dominated stands by diminishing the volume growth in these stands. The production capacities of the different sites ended up being different as well. There were quite a few stands in the reference data representing the very rich site, and fewer older stands with large stand basal area. Therefore the neighbours had to be selected from more infertile stands, and thus the growth was underestimated. However, the self-thinning models applied may underestimate especially the development of dense stands (Välimäki and Kangas 2009). Nevertheless, self-thinning, mortality, as well as the effects of thinning in detail were not taken into account in this thesis. Long-term planning requires that the models are able to predict the

effects of silvicultural practices that may be applied in the future (e.g. Hynynen et al. 2002). Thus, while applied in long-term forecasts, it is particularly important that the models extrapolate well beyond the calibration range (Hasenauer 2006). Non-parametric methods do not extrapolate outside the reference data. Therefore, the use of non-parametric methods in this kind of applications might be more limited. Moreover, the error propagation over time might prevent the use of non-parametric methods for long-term growth predictions, thus the degree of error propagation should be studied with adequate data.

In addition to testing further the incorporation of the effects of silvicultural treatments, some issues still remain to be considered in the context of non-parametric methods. The capability of the non-parametric methods in producing growth estimates for trees in sapling stands is yet to be studied, since the study data used in this thesis did not include sapling stands. Employing non-parametric methods requires reference data that should be extensive enough to represent all the possible conditions, and therefore growth measurements are needed. However, Mehtätalo (2004) suggested that it might be appropriate to measure growth in order to predict the growth more accurately than it is currently predicted while applying the parametric methods as well. The optimal number of sample trees used for reference data has not yet been studied in detail. In the different papers comprising this thesis, all the trees in one plot were used as reference data; however, when collecting data for these purposes, another selection method for the sample trees might be more efficient. Testing the effect of the size of reference dataset on accuracy showed that non-parametric methods do not necessarily require remarkably larger datasets or large amounts of local observations to ensure accurate estimates of growth. Usually larger reference data should result in better imputation results, because the reference data would better represent the variability in the population. Small datasets may also be usable, if the properties of the target and reference data are uniform enough (e.g. Malinen et al. 2001). Localized and local non-parametric imputation produced sufficiently accurate estimates with a fairly low number of measured local observations. However, this concerned only this situation and results in other applications might be different. Furthermore, the databases may be updated with repeated measures. Related to this, temporal correlation and how repeated measurements should be taken into account while applying non-parametric methods is an issue that also needs to be studied. The non-parametric methods constructed here might also be useful for imputing growth estimates for the tallied trees for which growth is not measured. Additionally, the method could be used simultaneously for many variables in such a situation.

REFERENCES

Adams, D.M. & Ek, A.R. 1974. Optimizing the management of uneven-aged forest stands. Canadian Journal of Forest Research 4: 274-287.

Altman, N.S. 1990. Kernel smoothing of data with correlated errors. Journal of the American Statistical Association 85: 749-759.

— 1992. An Introduction to kernel and nearest-neighbor nonparametric regression. American Statistician 46: 175-184.

Andreassen, K. & Tomter, S.M. 2003. Basal area growth models for individual trees of Norway spruce, Scots pine, birch and other broadleaves in Norway. Forest Ecology and Management 180: 11-24.

Atkeson, C.G., Moore, A.W. & Schaal, S. 1997. Locally weighted learning. Artificial Intelligence Review 11: 11-73.

Barth, A. & Ståhl, G. 2007. Determining sampling size in a national forest inventory by cost-plus-loss analysis. In: Barth, A. Spatially comprehensive data for forestry scenario analysis – Consequences of errors and methods to enhance usability. Doctoral thesis No. 2007:101. SLU, Faculty of Forest Sciences.

Batchelor, B.G. 1978. Pattern recognition: Ideas in practice. Plenum Press: New York. 485 p.

Bellman, R.E. 1961. Adaptive Control Processes. Princeton University Press: Princeton, NJ. 255 p.

Borders, B.E. & Bailey, R.L. 1986. A compatible system of growth and yield equations for slash pine fitted with restricted three-stage least squares. Forest Science 32: 185-201.

Breiman, L. 2001. Random Forests. Machine Learning 45: 5-32.

Buongiorno, J. & Michie, B.R. 1980. A matrix model for uneven-aged forest management. Forest Science 26: 609-625.

Burkhart, H.E. 1992. Tree and stand models in forest inventory. In: Nyyssönen, A., Poso, P. & Rautala, J. (Eds.). Proceedings of the Ilvessalo Symposium on National Forest Inventories. Finland 17–21 August 1992. The Finnish Forest Research Institute. Research Papers 444: 164-170.

Cajander, A.K. 1909. Über waldtypen. Fennia 28:1-175.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74: 829-836.

— & Devlin, S.J. 1988. Locally Weighted Regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83: 596-610.

— & Loader, C. 1994. Smoothing by local regression: Principles and methods. Technical report 95.3. AT&T Bell Laboratories, Statistics Department, Murray Hill, NJ. [Online article]. Available at: http://cm.bell-labs.com/cm/ms/departments/sia/doc/smoothing.springer.pdf. [Accessed 5 December 2008].

Clutter, J.L. 1963. Compatible growth and yield models for loblolly pine. Forest Science 9: 354-371.

Crookston, N.L., Moeur, M. & Renner, D. 2002. Users' Guide to the Most Similar Neighbour imputation Version 2. RMRS-GTR-96. US Department of Agriculture, Forest Service, Rocky Mountain Research Station: Ogden, UT. 35 p.

Daniels, R.F. 1993. Twenty years of stand modelling. Paper presented at IUFRO S4.01 Conference, Blacksburg, Virginia, USA. 27 September– 1 October 1993. 7 p.

— & Burkhart, H.E. 1975. Simulations of individual tree growth and stand development in managed loblolly pine plantations. In: FWS-5-75, Division of Forestry and Wildlife Research, Virginia Polytechnic Institute and State University: Blacksburg VA.

— & Burkhart, H.E. 1988. An integrated system of forest stand models. Forest Ecology and Management 23: 159-177.

Diday, E. 1974. Recent progress in distance and similarity measures in pattern recognition. Second International Joint Conference on Pattern Recognition: 534-559.

Eerikäinen, K. 2002. A site dependent simultaneous growth projection model for *Pinus kesiya* plantations in Zambia and Zimbabwe. Forest Science 48: 518-529.

— & Maltamo, M. 2003. A percentile based basal area diameter distribution model for predicting the stand development of *Pinus kesiya* plantations in Zambia and Zimbabwe. Forest Ecology and Management 172: 109-124.

Ek, A.R. 1974. Nonlinear models for stand table projection in northern hardwood stands. Canadian Journal of Forest Research 4: 23-27.

— & Monserud, R.A. 1974. Trials with program FOREST: Growth and reproduction simulation for mixed species even- or uneven-aged forest stands. Julkaisussa: Fries, J. (Ed.) Growth models for tree and stand simulation. Royal College of Forestry, Research Notes 30: 56–73.

— , Robinson, A. P., Radtke, P. J. & Walters, D. K. 1997. Development and testing of regeneration imputation models for forests in Minnesota. Forest Ecology and Management 94: 129-140.

Fan, J. 2000. Prospects of nonparametric modeling. Journal of the American Statistical Association 95: 1296-1300.

Fang, Z., Bailey, R.L. & Shiver, B.D. 2001. A multivariate simultaneous prediction system for stand growth and yield with fixed and random effects. Forest Science 47: 550-562.

Frescino, T., Edwards, T.C. Jr. & Moisen, G. 2001. Modelling spatially explicit structural attributes using generalized additive models. Journal of Vegetation Science 12: 15-26.

Friedman, J.H. 1994. On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery 1: 55-77.

Furnival, G.M. & Wilson, R.W. 1971. Systems of equations for predicting forest growth and yield. In: Patil, G.P., Pielou, E.C. & Walters, W.E. (Eds.). Statistical Ecology 3: 43-55.

Gartner, D. 2002. The effect of data quality on short-term growth model projections. Proceedings of the 4th Annual Forest Inventory and Analysis Symposium. November 19-21 2002. New Orleans, LA. Gen. Tech. Rep. NC-252: 41-43.

Gertner, G. Z. & Dzialowy, P. J. 1984. Effects of measurement errors on an individual tree-based growth projection system. Canadian Journal of Forest Research 14:311-316.

Guisan, A., Edwards Jr., T.C. & Hastie, T.J. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89-100.

Gustavsen, H.G., Roiko-Jokela, P. & Varmola, M. 1998. Kivennäismaiden talousmetsien pysyvät (INKA ja TINKA) kokeet. Suunnitelmat, mittausmenetelmät ja aineistojen rakenteet. The Finnish Forest Research Institute. Research Papers 292. 212 p. (In Finnish).

— 1998. Volymtillväxten och övre höjdens utveckling i tall-dominerade bestånd i Finland – en utvärdering av några modellers validitet i nuvarande skogar. The Finnish Forest Research Institute. Research Papers 707: 190 p. (In Swedish).

Haara, A. 2002. Kasvuennusteiden luotettavuuden selvittäminen knn-menetelmällä ja monitavoiteoptimoinnilla. Metsätieteen aikakauskirja 3/2002: 391-406. (In Finnish).

— , Maltamo, M. & Tokola, T. 1997. The k-nearest-neighbour method for estimating basal area diameter distribution. Scandinavian Journal of Forest Research 12: 200-208.

— & Korhonen, K.T. 2004. Kuvioittaisen arvioinnin luotettavuus. Metsätieteen aikakauskirja 4/2004: 489-508. (In Finnish).

Haight, R.G. & Getz, W.M. 1987. A comparison of stage-structured and single-tree models for projecting forest stands. Natural Resource Modelling 2: 279-298.
Härdle, W. 1989. Applied nonparametric regression. Cambridge University Press. 323 p.
— 1990. Smoothing techniques with implementation in S. Springer Series in Statistics. Springer-Verlag: Heidelberg, New York.261 p.
Hasenauer, H. 1997. Dimensional relationships of open-grown trees in Austria. Forest Ecology and Management 96: 197-206.
— (Ed.). 2006. Sustainable forest management. Growth models for Europe. Springer Verlag: Berlin and Heidelberg. 398 p.
— , Monserud, R.A. & Gregoire, T.G. 1998. Using simultaneous regression techniques with individual-tree growth models. Forest Science 44: 87-95.
Hastie, T. & Tibshirani, R. 1986. Generalized Additive Models. Statistical Science 1: 297-318.
— & Tibshirani, R. 1987. Generalized Additive Models: Some Applications. Journal of the American Statistical Association 87: 371-386.
— & Tibshirani, R. 1990. Generalized Additive Models. Chapman and Hall: London. 356 p.
Hool, J.N. 1966. A dynamic programming-Markov chain approach to forest yield control. Forest Science Monograph 12: 1-26.
Huang, S. & Titus, J.T. 1999. Estimating a system of nonlinear simultaneous individual tree models for white spruce in boreal mixed-species stands. Canadian Journal of Forest Research 29: 1805-1811.
Huuskonen, S. & Miina, J. 2007. Stand-level growth models for young Scots pine stands in Finland. Forest Ecology and Management 241: 49-61.
Hynynen, J. 1995. Modelling tree growth for managed stands. Academic dissertation. The Finnish Forest Research Institute. Research Papers 576. 59 p. + appendices.
— , Ojansuu, R., Hökkä, H., Siipilehto, J., Salminen, H. & Haapala, P. 2002. Models for predicting stand development in MELA system. The Finnish Forest Research Institute. Research Papers 835. 116 p.
Kangas, A. & Korhonen, K.T. 1995. Generalizing sample tree information with semiparametric and parametric models. Silva Fennica 29: 151-158.
Kalliola, R. 1973. Suomen kasvimaantiede. WSOY: Porvoo-Helsinki. 308 p. (In Finnish).
Katila, M. & Tomppo, E. 2001. Selecting estimation parameters for the Finnish multisource National Forest Inventory. Remote Sensing of Environment 76: 16-32.
— & Tomppo, E. 2002. Stratification by ancillary data in multisource forest inventories employing k-nearest-neighbour estimation. Canadian Journal of Forest Research 32: 1548-1561.
Koistinen, P., Holmström, L. & Tomppo, E. 2008. Smoothing methodology for predicting regional averages in multi-source forest inventory. Remote Sensing of Environment 112: 862-871.
Korhonen, K.T. & Kangas, A. 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scandinavian Journal of Forest Research 12: 97-101.
Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae 108. 72 p.
Lappi, J. 1986. Mixed linear models for analysing and predicting stem form variation of Scots pine. Communicationes Instituti Forestalis Fenniae 134. 69 p.
— 1991. Calibration of height and volume equations with random parameters. Forest Science 37: 781-801.
— 1993. Metsäbiometrian menetelmiä. Silva Carelica 24. 182 s. (In Finnish).

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

— , Maedel, J. & Coops, N. C. 2008. Estimating stand structural details using variable-space nearest neighbour analyses to link ground data, forest cover maps, and Landsat imagery. Remote Sensing of Environment 118: 2578-2591.

Liao, W., Nogami, K. & Imanaga, M. 1998. An application of neurogenetic algorithm system to individual tree growth model. Journal of Forest Research 3: 79-83.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India 12: 49-55.

Mäkinen, A., Kangas, A., Välimäki, E., Kalliovirta, J., & Rasinmäki, J. 2008. Comparison of treewise and standwise forest simulators with the aid of quantile regression. Forest Ecology and Management 255: 2709-2717.

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

— 2003. Prediction of characteristics of marked stand and metrics for similarity of log distribution for wood procurement management. Academic dissertation. University of Joensuu. Faculty of Forestry. 46 p. + appendices.

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

— & Eerikäinen, K. 2001. The Most Similar Neighbour reference in the yield prediction of *Pinus kesiya* stands in Zambia. Silva Fennica 35: 437-451.

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

Martínes Pastur, G.J., Cellini, J.M., Lencinas, M.V. & Peri, P.L. 2008. Stand growth model using volume increment/basal area ratios. Journal of Forest Science 54: 102-108.

Matala, J., Ojansuu, R., Peltola, H., Raitio, H. & Kellomäki, S. 2006. Modelling the response of tree growth to temperature and CO2 elevation as related to the fertility and current temperature sum of a site. Ecological Modelling 199: 39-52.

McCullagh, P. & Nelder, J. 1989. Generalized Linear Models. 2nd edn. Chapman and Hall: London. 532 p.

McRoberts, R.E., Nelson, M.D. & Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. Remote Sensing of Environment 82: 457-468.

McRoberts, R.E., Tomppo, E.O., Finley, A.O. and Heikkinen, J. 2007. Estimating areal means and variances of forest attributes using k-nearest neighbors technique and satellite imagery. Remote Sensing of Environment 111: 466-480.

Mehtätalo, L. 2004. Predicting stand characteristics using limited measurements. Finnish Forest Research Institute. Research Papers 929. 39 p. + appendices.

Moeur, M. & Stage, A.R. 1995. Most Similar Neighbor. An improved sampling inference procedure for natural resource planning. Forest Science 41: 337-359.

— & Hershey, R.R. 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. In: Lowell, K., Jaton, A. (eds). Spatial Accuracy Assessment: Land information uncertainty in natural resources. Papers presented at the Third International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences in Quebec City, Canada. May 20-22 1998. Ann Arbor Press: Chelsea, MI. p. 419-430.

Moisen, G.G. & Frescino, T.S. 2002. Comparing five modeling techniques for predicting

forest characteristics. Ecological Modelling 157: 209-225.
Monserud, R.A. & Sterba, H. 1996. A basal area increment model for individual trees growing in even- and uneven-aged forest stands in Austria. Forest Ecology and Management 80: 57–80.
Mowrer, H.T. & Frayer, W.E. 1986. Variance propagagation in growth and yield projections. Canadian Journal of Forest Research 16: 1196-1200.
Munro, D.D. 1974. Forest growth models: a prognosis. In: Fries, J. (Ed.). Growth models for tree and stand simulation. Research Note 30, Department of Forest Yield Research, Royal College of Forestry, Stockholm. p. 7-21.
Nuutinen, T., Matala, J., Hirvelä, H., Härkönen, K., Peltola, H., Väisänen, H. & Kellomäki, S. 2006. Regionally optimized forest management under changing climate. Climatic Change 79: 315-333.
Nyyssönen, A. & Mielikäinen, K. 1978. Metsikön kasvun arviointi. (Estimation of stand increment.) Acta Forestalia Fennica 60: 1-17. (In Finnish)
Ochi, N. & Cao, Q.V. 2003. A comparison of compatible and annual growth models. Forest Science 49: 285-290.
Ohmann, J.L. & Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbour imputation in Coastal Oregon, USA. Canadian Journal of Forest Research 32: 725-741.
Opsomer, J.D. 2000a. Asymptotic properties of backfitting estimators. Journal of Multivariate Analysis 73:166-179.
— 2000b. Nonparametric regression in environmental statistics. Iowa State University. [Online article]. Available at: http://www.public.iastate.edu/~jopsomer/papers/Env. [accessed 5 June 2008].
— 2002. A Note on Local Scoring and Weighted Local Polynomial Regression in Generalized Additive Models. Department of Statistics, Iowa State University. [Online article]. Available at: http://www.stat.colostate.edu/~jopsomer/papers/Local_scoring.pdf. [accessed 5 June 2008].
— , Wang, Y. & Yang, Y. 2001. Nonparametric regression with correlated errors. Statistical Science 16: 134-153.
— , Breidht, F.J., Moisen, G.G. & Kauerman, G. 2007. Model-assisted estimation of forest resources with generalized additive models. Journal of the American Statistical Association 102: 400-416.
Packalén, P. & Maltamo, M. 2007. The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment 109: 328-341.
— & Maltamo, M. 2008. Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. Canadian Journal of Forest Research 38: 1750-1760.
Pienaar, L.V. & Harrison, W.M. 1989. Simultaneous growth and yield prediction equations for *Pinus elliottii* plantations in Zululand. South African Forestry Journal 149: 48-53.
Pretzsch, H., Biber, P. & Ďurský, J. 2002. The single tree-based stand simulator SILVA: Construction, application and evaluation. Forest Ecology and Management 162: 3-21.
Rao, C.R., Miller, J.P. & Rao, D.C. (Eds.). 2008. Epidemiology and medical statistics. Handbook of Statistics 27. Elsevier Ltd., Oxford. 852 p.
Rasinmäki, J., Kalliovirta, J. & Mäkinen, A. 2009. An adaptable simulation framework for multiscale forest resource data. Computers and Electronics in Agriculture (in press).
Räty, M. & Kangas, A. 2007. Localizing general models based on local indices of spatial association. European Journal of Forest Research 126: 269-289.

— & Kangas, A. 2008. Localizing general models with the classification and regression trees (CART). Scandinavian Journal of Forest Research (in press).
Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27: 832-837.
Sallnäs, O. 1990. A matrix growth model of the Swedish forest. Studia Forestalia Suecica 183. 23 p.
SAS. 1992. Technical report P-229. SAS/STAT software: Changes and enhancements, release 6.07. SAS Institute Inc.: Cary, NC. 620 p.
Silverman, B.W. 1984. Spline smoothing: The equivalent variable kernel method. The Annals of Statistics 12: 898-916.
Solomon, D.S., Herman, D.A. & Leak, W.B. 1995. FIBER 3.0: An ecological growth model for northeastern forest types. USDA Forest Service. Gen. Tech. Rep. NE-204. 24 p.
Stage, A.R. & Crookston, N.L. 2007. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. Forest Science 53: 62-72.
Stone, C.J. 1977. Consistent nonparametric regression. The Annals of Statistics 5: 595-645.
Sullivan, A.D. & Clutter, J.L. 1972. A simultaneous growth and yield model for loblolly pine. Forest Science 18: 76-86.
Talvitie, M. 2005. Männyn pohjapinta-alan kasvumallien kalibrointi Pohjois-Pohjanmaan metsäkeskuksen alueelle. Metsätieteen aikakauskirja 1/2005: 19–32. (In Finnish).
Temesgen, H., LeMay, V.M., Marshall, P.L. & Froese, K. 2003. Imputing tree-lists from aerial attributes for complex stands of British Columbia. Forest Ecology and Management 177: 277-285.
— , Barrett, T. M. & Latta, G. 2008. Estimating cavity tree abundance using nearest neighbour imputation methods for western Oregon and Washington forests. Silva Fennica 42: 337-354.
Theodoridis, S. & Koutroumbas, K. 2006. Pattern recognition. 3rd edn. Academic Press. Orlando, FL, USA. 837 p.
Tokola, T. 2000. The influence of field sample data location on growing stock volume estimation in Landsat TM-based forest inventory in eastern Finland. Remote Sensing of Environment 74: 421-430.
— , Kangas, A., Kalliovirta, J. Mäkinen, A. & Rasinmäki, J. 2006. SIMO-SIMulointi ja Optimointi uuteen metsäsuunnitteluun. Metsätieteen aikakauskirja 1/2006: 60-65. (In Finnish).
Tomppo, E. & Halme, M. 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: A genetic algorithm approach. Remote Sensing of Environment 92: 1-20.
Usher, M.B. 1969. A matrix model for forest management. Biometrics 25: 309-315.
Välimäki, E. & Kangas, A. 2009. Kasvumallien toiminnan validointi ylitiheissä metsiköissä. Metsätieteen aikakauskirja 2/2009: 97–112. (In Finnish).
Vancley, J.K. 1994. Modelling forest growth and yield: Applications to mixed tropical forests. CAB International. 312 p.
— & Skovsgaard, J.P. 1997. Evaluating forest growth models. Ecological Modelling 98: 1-12.
Wykoff, W.R. 1990. A basal area increment model for individual conifers in the northern Rocky Mountains. Forest Science 36: 1077-1104.
— , Crookston, N.L. & Stage, A.R. 1982. User's Guide to the Stand Prognosis Model. INT-133. US Department of Agriculture, Forest Service: Ogden, UT.
Zhang, L., Ma, Z. & Guo, L. 2008. Spatially assessing model errors of four regression techniques for three types of forest stands. Forestry 81: 209-225.

Zhao, D., Borders B.E. & Wilson M.D. 2003. Individual-tree diameter growth and mortality models for bottomland mixed-species hardwood stands in the lower Mississippi alluvial valley. Forest Ecology and Management 199: 307-322.