Estimating Individual Tree Growth with the k-Nearest Neighbour and k-Most Similar Neighbour Methods

Susanna Sironen, Annika Kangas, Matti Maltamo and Jyrki Kangas

Sironen, S., Kangas, A., Maltamo, M. & Kangas, J. 2001. Estimating individual tree growth with the k-nearest neighbour and k-Most Similar Neighbour methods. Silva Fennica 35(4): 453–467.

The purpose of this study was to examine the use of non-parametric methods in estimating tree level growth models. In non-parametric methods the growth of a tree is predicted as a weighted average of the values of neighbouring observations. The selection of the nearest neighbours is based on the differences between tree and stand level characteristics of the target tree and the neighbours. The data for the models were collected from the areas owned by Kuusamo Common Forest in Northeast Finland. The whole data consisted of 4051 tally trees and 1308 Scots pines (*Pinus sylvestris* L.) and 367 Norway spruces (*Picea abies* Karst.). Models for 5-year diameter growth and bark thickness at the end of the growing period were constructed with two different non-parametric methods: the k-nearest neighbour regression and k-Most Similar Neighbour method. Diameter at breast height, tree height, mean age of the stand and basal area of the trees larger than the subject tree were found to predict the diameter growth most accurately. The non-parametric methods were compared to traditional regression growth models and were found to be quite competitive and reliable growth estimators.

Keywords pine, spruce, single tree growth models, non-parametric models, local estimates
Authors' addresses *Sironen* and *Maltamo*, University of Joensuu, Faculty of Forestry, P.O. Box 111, FIN-80101 Joensuu, Finland; *Kangas* and *Kangas*, Finnish Forest Research Institute, Kannus Research Station, P.O. Box 44, FIN-69101 Kannus, Finland
E-mail susanna.sironen@forest.joensuu.fi
Received 7 March 2001 **Accepted** 21 November 2001

1 Introduction

In forest management, information on both current forest resources and future yields is needed. The future development of forest resources can be predicted with growth and yield models. The main uses of growth and yield predictions are updating forest inventories, comparing silvicultural treatments by simulating them and predicting their outcomes, harvest scheduling, stand and forest level decision support and management planning (e.g. Burkhart 1992, Hynynen 1995). The growth and yield models have been developed for many different purposes. The models can be simple growth and yield tables derived from appropriate data or sophisticated computer models (e.g. Mielikäinen and Gustavsen 1992).

Growth models may be classified in different groups according to the data collected and the information needed. The models which only require stand level information are called stand models. The stand level growth models have earlier been very common in Finland (e.g. Vuokila 1965, Gustavsen 1977, Nyyssönen and Mielikäinen 1978). In these models the relative volume increment of a stand can depend on variables like stand age, basal area and site type. Projection models are a system of simultaneously estimated static difference equations for stand volume and yield prediction in different time points. Predicting future volume yields with the projection models requires, for example, projections of the number of surviving trees per hectare, basal area per hectare and average height (e.g. Pienaar and Harrison 1989). The stand models are easy to use and inventory costs are low. However, stand level may not be reliable in heterogeneous stands, and the allocation of growth to different dimensions cannot be directly evaluated (Gustavsen 1998).

Models which require individual tree information and use individual trees as the basic unit to produce yield estimates are called individual tree models. Usual individual tree growth models separately predict the increment of tree diameter or basal area and height (e.g. Nyyssönen and Mielikäinen 1978, Ojansuu et al. 1991, Hynynen 1995). The individual tree models can be further divided into distance-independent and distance-dependent or spatial growth models. The distance-dependent growth models require information about individual tree locations (e.g. Vuokila 1965, Pukkala 1989, Hynynen 1995, Miina et al. 1991). The models based on individual tree growth provide detailed information about stand dynamics and structure, including the distribution of volume in size classes (Burkhart 1992).

In regression models, growth is predicted as a function of different tree and stand variables correlating with growth (e.g. Mielikäinen 1992). Non-parametric methods are an alternative to these traditional parametric methods. In the non-parametric methods the growth is predicted as a weighted average of the growth of the neighbouring observations. The selection of the nearest neighbours can be based on the differences between tree and stand level characteristics of the target tree and the neighbours.

The nearest neighbours are chosen from a database of previously measured tree and stand level observations. Thus, unrealistic growth estimates cannot occur, because estimates are chosen from actual, measured samples (e.g. Moeur and Stage 1995). Gustavsen (1998) found notable biases in Northern Finland's growth estimates predicted with models which comprise the whole of Finland, e.g. 8–9 m^3/ha in five years. With non-parametric models, the bias may be reduced, as the reference trees can be chosen from nearby areas. In the regression models, localization can be made for instance by calibrating the models or using coordinates as regressors (e.g. Gertner 1984, Korhonen 1993), but not as easily as with non-parametric methods. In addition to localization, advantages of the non-parametric methods include that they retain more of the variation of the data and preserve the correlations of dependent variables (e.g. Moeur and Stage 1995). The non-parametric models do also have parameters like bandwidth in kernel and the number of nearest neighbours (k) in k-nn method, but they do not require predefined functional form. Unlike the regression models, the non-parametric methods need reference data also at the application phase (Maltamo and Eerikäinen 2000). The non-parametric models, however, update themselves when data is added or removed from the database.

The k-nearest neighbour method has been used in many forestry applications, including gener-

alization of sample tree information, estimation of the diameter distribution and estimation of the characteristics of marked stand (e.g. Korhonen and Kangas 1997, Haara et al. 1997, Maltamo and Kangas 1998, Tommola et al. 1999). The Most Similar Neighbour method has been used in multivariate forest inventory applications (Moeur and Stage 1995, Moeur and Hershey 1999). The non-parametric methods also include spline smoothing, kernel and grid, but these methods are more difficult to apply in multi-dimensional cases, i.e. when several independent variables are used, than k-nn and k-MSN methods (e.g. Härdle1989).

The purpose of this study was to test and compare non-parametric k-nearest neighbour and k-Most Similar Neighbour methods in growth prediction. The aim of the prediction was to build single tree diameter growth models for Scots pine and Norway spruce for local conditions in Northern Finland. The non-parametric growth models were further compared to a traditional regression growth model constructed using mixed model technique.

2 Material and Methods

2.1 Study Data

The study data were collected during the summer of 1999 from the areas owned by Kuusamo Common Forest in Kuusamo. Sampling of the study data included seven main strata: pine and spruce dominated moist heaths, pine dominated dryish and dry heaths, pine and spruce swamps and pine forests with low productivity. In the non-parametric methods it is important that the data is evenly distributed to different growing sites and age classes. All the main strata were further divided into six 30-year age classes. Two stands were supposed to be measured from each of these strata i.e. 84 stands. The stands were objectively located to different parts of Kuusamo. The stands with notable damage or dominant height lower than 3 meters were not included in the data.

Two circular sample plots were placed systematically in each stand. The distance between sample plots was 40 meters. The size of the plot varied from 100 m^2 to 700 m^2 according to the density of the stand. Diameter at breast height was recorded for all trees in these plots. From every sample plot an average of 9 sample trees were selected by establishing a circular subplot of a quarter of the plotsize at the centre of each plot. The characteristics of the sample trees measured within the inner circles included height, length of the live crown, bark thickness and 5-year diameter increment. Several variables describing the site and the growing stock were also registered for each stand. These variables included location, altitude, effective sum of temperature, soil type, site class and dominant tree species. The mean stand age was determined by measuring age from one-third of the sample trees.

A total of 71 stands were measured. 53 stands were dominated by Scots pine (*Pinus sylvestris* L.) and 18 stands by Norway spruce (*Picea abies* Karst.). The whole study material consisted of 4051 tally trees and 1308 sample trees, of which 941 were pines and 367 were spruces. Most of the pines were located in moist and dryish (*Myrtillus* and *Vaccinium-Myrtillus)* forest site types and the proportion of pines located in dry (*Vaccinium*) forest site type was small. The spruces were mainly located in moist sites. Most spruces in

Table 1. Description of the mean tree and stand characteristics in the study data according to the tree species (SD = standard deviation).

Character	Scots pine		Norway spruce	
	Mean	SD	Mean	SD
Altitude, m	264	30	275	25
Effective sum of temperature, dd	803	16	805	17
Mean stand age, years	65	41	109	50
Basal area of the stand, m^2/ha	15	7	22	8
Mean diameter, cm	19.6	5.2	22.9	3.5
Diameter at breast height, cm	14.7	8.5	14.2	7.5
Height, m	10.8	5.4	9.9	4.9
5-year diameter growth, cm	0.99	1.09	0.58	0.49

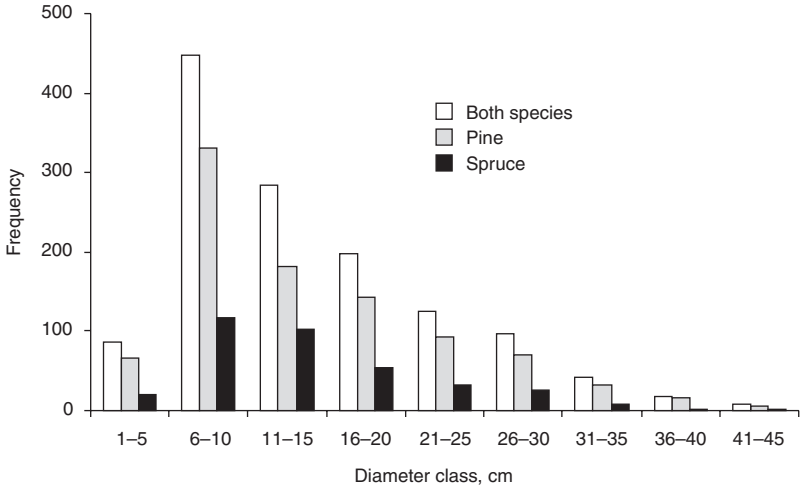

Fig. 1. Diameter distribution of the sample trees in the study data.

the study data belonged to mature forests and the proportions of other stages of stand development were small. The pines were distributed more evenly to different age classes than the spruces. Mean age of the spruce stands was 109 years and pine stands 65 years (Table 1). The greatest frequency was observed in class with dbh smaller than 10 cm (Fig. 1). The proportion of trees with large diameter was small for both tree species.

Data preparation included back-calculations of tree and stand characteristics because the data were collected from temporary sample plots. Tree diameters under bark for the sample trees at the beginning of the growth period were calculated by subtracting the 5-year diameter growth and thickness of the bark from the measured tree diameters. Bark thickness and tree height at the beginning of the growth period were estimated with models. Bark and height models were estimated using mixed models, because the observations were correlated due to the hierarchical structure of the data (e.g. Lappi 1993). 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 characters at the beginning of the growth period were calculated by means of these estimated tree diameters and heights. The data preparation included also calculating characteristics describing the position of the tree in the stand, such as the basal area of trees larger than the subject tree and relative tree size.

2.2 Modelling the Diameter Growth

Two kinds of non-parametric methods were utilized: the k-nearest neighbour regression and the k-Most Similar Neighbour (k-MSN) method (e.g. Härdle 1989, Altman 1992, Moeur and Stage 1995). In the estimation of the non-parametric model a distance function must be determined in order to compare different trees and their characteristics. The distance function can be based e.g. on differences between tree and stand level variables of target and reference trees. In the estimation of the growth for a given target tree the differences across all reference trees are calculated and the growth estimate is formed using the chosen nearest neighbours (e.g. Korhonen and Kangas 1997). In addition to deciding the shape of the distance function, the number of nearest neighbours must be defined. When the number of nearest neighbours is small, the estimate is very close to the original data. The estimate is almost unbiased, but over-fitting. If the number of nearest neighbours is large, the estimate will be very smooth and may be highly biased (Altman 1992). The manner the weights of the reference trees depend on the distance must also be defined.

Weighted averages are used to reduce the bias of the nearest neighbour estimator (Altman 1992).

2.2.1 The k-Nearest Neighbour Method

In the k-nearest neighbour regression, the similarity of the trees was measured by using dimensionless distance function, which is based on absolute differences between stand and tree characteristics. This kind of distance function is not as sensitive to exceptional observations as the squared deviation method (Maltamo and Kangas 1998). The distance function was defined as

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

where

x_{il} = the value of the considered variable l for reference tree i

x_{jl} = the value of the considered variable l for target tree j

p = the number of variables

c_l = the coefficient for variable x_l

In order to avoid the influence of different units of measurements, the variables were standardized by subtracting the mean of the variable and dividing it by the standard deviation of the variable. The weights of the reference trees were based on the inverse of the distance. The weight w_{ij} of reference tree i for target tree j was

$$w_{ij} = \frac{\left(\frac{1}{d_{ij}}\right)^{pm}}{\sum_{i=1}^{k}\left(\frac{1}{d_{ij}}\right)^{pm}} \qquad (2)$$

where k is the number of the nearest reference trees used and pm is the die-off parameter and $i \neq j$ (Haara ym. 1997). The die-off parameter determines how quickly the weights of the nearest reference trees decrease when the distance d_{ij} increases. The effect of the similarity distance function and die-off parameter to the estimates of diameter growth was examined by using cross-validation method (Härdle 1989). In this method, each observation is predicted with the reference data excluding the observation itself. The values of parameters pm, c and k were searched heuristically, using iteration. The nearest neighbours were never taken from the same stand or sample plot as the target tree. This restriction was used because otherwise results would be too optimistic, because observations are strongly correlated in same stands and neighbouring observations from the same stands are usually absent in practical applications.

In the k-nearest neighbour method the final estimate for the 5-year diameter growth of the target tree ($\hat{y}_j$) was calculated as the weighted average of the growth of the k nearest reference trees (y_i)

$$\hat{y}_j = \sum_{i=1}^{k} w_{ij} y_i \qquad (3)$$

in which k is the number of nearest reference trees used and w_{ij} is the weight of the reference tree i to target tree j. Bark thickness of the target tree at the end of the growing period was calculated as a weighted average of the same trees as the growth.

2.2.2 The k-Most Similar Neighbour Method

The Most Similar Neighbour (MSN) method is based on canonical correlation analysis between independent and dependent variables (Moeur and Stage 1995). The benefit of the MSN method compared to basic k-nearest neighbour regression is that the enormous number of iterations in the search of nearest neighbours can be avoided because the coefficients for the variables are obtained directly from the canonical correlation analysis and all the possible independent and dependent variables can be used in the calculation of the weighting matrix (e.g. Maltamo and Eerikäinen 2000). In the MSN method, the most similar neighbour to the observation j in the target data is that observation in the reference data, for which $(\hat{Y}_j - Y_i)W(\hat{Y}_j - Y_i)$ is minimized over all $i = 1,\ldots,n$ reference trees, where $\hat{Y}_j$ is a row vector of the unknown variables in the target data, Y_i is a row vector of the observed variables in the reference data and W is a weighting matrix. In the MSN method, the relation of unknown and

observed variables is replaced by the relation of independent variables which are known both in the target data and reference data. The weighting matrix in the distance function is calculated on canonical correlation analysis by summarizing the relationships between dependent (Y) and independent (X) 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

$$U_r = \alpha_r Y \text{ and } V_r = \gamma_r X \quad (4)$$

where α_r are canonical coefficients of the dependent variables ($r = 1 \ldots s$) and γ_r are 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).

The distance function derived from canonical correlation analysis is

$$d_{ij}^2 = \underset{1\times p}{(X_i - X_j)}\ \underset{p\times p}{\Gamma\Lambda^2\Gamma'}\ \underset{p\times 1}{(X_i - X_j)'} \quad (5)$$

where

X_j = independent variables of the target tree
X_i = independent variables of reference tree
Γ = matrix of the canonical coefficients of the independent variables, $\underset{p\times s}{\gamma_r}$
Λ^2 = diagonal matrix of squared canonical correlations, $\underset{s\times s}{\lambda_r^2}$
s = number of the canonical correlations used
p = number of the independent variables

The distance function calculates the squared distance between the target tree and reference tree. Each sample tree, in turn, is used as a target tree and the target tree is temporarily excluded from the reference trees. The variables were standardized for being able to avoid the influence of different units of the variables. The Most Similar Neighbour method was applied by testing different number of nearest neighbours in the calculations of the final estimate (k-MSN). The standardization of the variables, the weighting of the reference trees w_{ij} (2) and the final growth estimate $\hat{y}_j$ (3) were similar to the basic k-nearest neighbour method except that the die-off parameter (pm) was 1 for the k-MSN method.

2.2.3 Criteria of Evaluation

In both methods, the optimal combination of variables and parameters was achieved when the relative root mean square error (RMSE%) and bias (b_e%) of the growth estimates were the lowest. The RMSE is a widely used criteria to evaluate the estimations given by the k-nearest neighbour methods. The relative RMSE was calculated by using

$$\text{RMSE\%} = \text{RMSE} \cdot 100 / \bar{\hat{y}} \quad (6)$$

where RMSE is the root mean square error and $\bar{\hat{y}}$ the mean of the growth estimates. The root mean square error was calculated by using

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_j - \hat{y}_j)^2}{n}} \quad (7)$$

where n is the number of trees, y the observed growth of tree j and $\hat{y}$ the growth estimate of tree j. The relative bias was

$$b_e\% = b_e \cdot 100 / \bar{\hat{y}} \quad (8)$$

where b_e is the mean of the residuals.

2.2.4 Regression Model with Mixed Model Technique

The non-parametric k-nearest neighbour and k-Most Similar neighbour methods were compared to 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 hierarchical structure of the study data. The Ordinary Least Squares (OLS) method assumes that all observations used in modelling are independent. The observations are often spatially or temporarily correlated in forestry applications, if there are several trees measured in the same stands in the study data or trees are measured more than once (e.g. Lappi 1993). The correlation between the observations can be taken into account in random parameter models. The data used in this study were measured from stands including two sample plots. Thus, three random variables were included in the model: random stand variable, random plot variable and error variable. The mixed model including the fixed part and the random variables can be described using the following function

$$y_{ijk} = b_1 x_{1ijk} + b_2 x_{2ijk} + ... + b_n x n_{ijk} + s_i + p_{ij} + e_{ijk} \quad (9)$$

where y_{ijk} is the 5-year diameter growth of tree k in plot j in stand i, $x_{1ijk},\ldots,x_{nijk}$ are independent variables for the kth tree in the jth plot in the ith stand, $b_1,\ldots,b_n$ are fixed parameters and s_i is the random stand variable with $E(s_i)=0$ and $var(s_i)=\sigma_s^2$, p_{ij} random plot variable with $E(p_{ij})=0$ and $var(p_{ij})=\sigma_p^2$ and e_{ijk} random error with $E(e_{ijk})=0$ and $var(e_{ijk})=\sigma_e^2$.

3 Results

3.1 Diameter Growth Models

3.1.1 The k-Nearest Neighbour Method

In this study, the optimal variables for the distance function, coefficients of the variables, the number of nearest neighbours and the weighting parameter were determined heuristically when applying k-nearest neighbour method. Due to the estimation method, enormous number of different combinations of the parameters were tested. The variables used in modelling were chosen among easily measured or traced tree and stand characteristics, including e.g. tree diameter, height, tree basal area and relative size of the tree. Stand age, basal area of the stand, basal area mean diameter, altitude and temperature sum were tested as stand level variables.

Tree diameter, tree height, stand age at breast height and basal area of trees larger than the subject tree were found to predict the diameter growth most accurately. When searching for the optimal coefficients of the variables, all possible combinations of values from 1 to 10 were tested. The chosen coefficients of the variables are presented in Table 2. The coefficient of the basal area larger than the subject tree (G_{large}) affected strongly the accuracy of the pine growth estimates. The relative root mean square error

Table 2. Number of the nearest neighbours (k) in the k-nn and k-MSN method, coefficients of the independent variables and values of the die-off parameters (pm) in the k-nn method, canonical coefficients of the independent variables (Γ) and squared canonical correlations (Λ^2) in the k-MSN method and parameter estimates of the mixed models. Independent variables include tree diameter at breast height (dbh), tree height, stand age at breast height and basal area larger than the subject tree (G_{large}).

	k-nn method			k-MSN method			Mixed model	
	Pine	Spruce		Pine	Spruce		Pine	Spruce
k	15	15		15	14	Intercept	2.289738	2.532122
dbh	9	3	dbh	–0.4205	–0.0083	ln(dbh)	–0.134446	–0.858552
height	2	2	height	0.6514	–0.2762	ln(height)	0.613229	1.252947
age	8	6	age	0.6368	0.9080	ln(age)	–0.985427	–0.875844
G_{large}	1	0.5	G_{large}	0.2578	0.4207	G_{large}	–0.012210	–0.00184
pm	3	1	Λ^2	0.5260	0.4061	σ^2_{stand}	0.067368	0.062343
						σ^2_{plot}	0.071383	0.062339
						σ^2_e	0.197005	0.283910

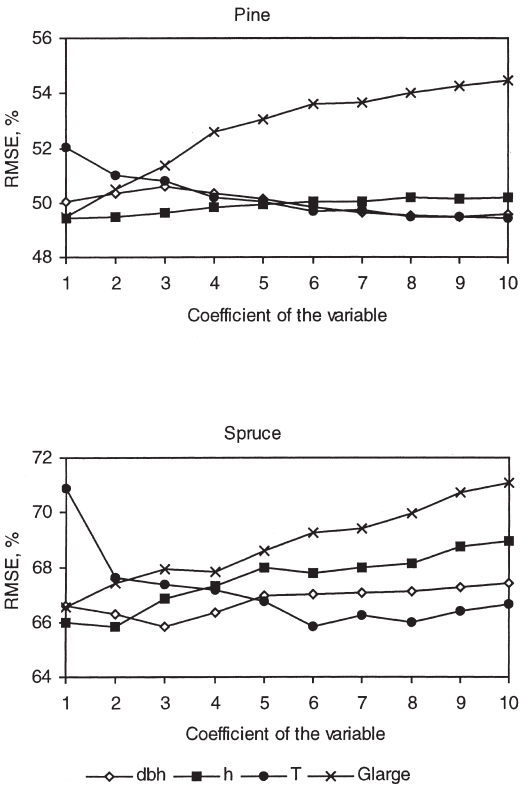

Fig. 2. Influence of the coefficients of the variables on the diameter growth of Scots pine and Norway spruce in the k-nn method. Other parameters are held constant while changing the value of the coefficient of the variable in question from 1 to 10. Variables include tree diameter (dbh), tree height (h), stand age (T) and basal area larger than the subject tree (G_{large}).

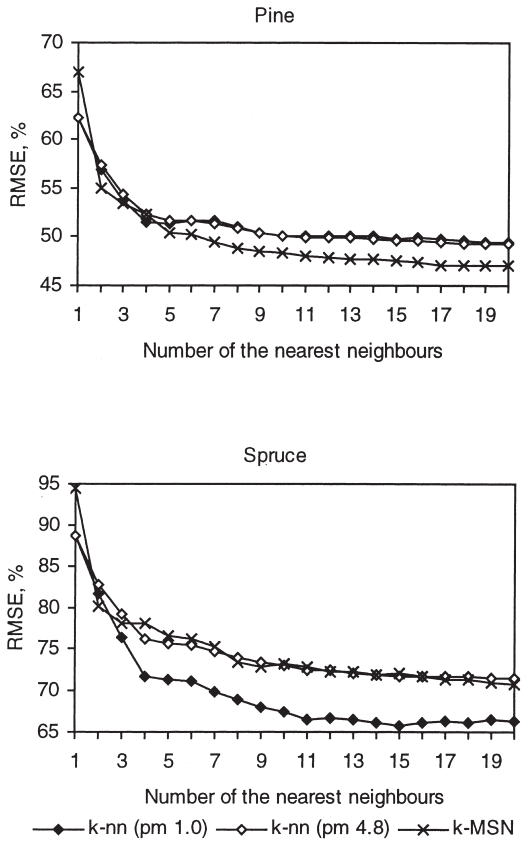

Fig. 3. Influence of the reference trees (k) with two different die-off parameter (pm) values on the relative root mean square error (RMSE%) of the growth estimates of Scots pine and Norway spruce in the k-nn method and influence of the reference trees (k) in the k-MSN method.

(RMSE) of the growth estimates increased 10% when the weight of this variable increased from 1 to 10 (Fig. 2). The value of the coefficient of G_{large} had to be small also in the distance function of spruces. Changing the value of the coefficient of stand age had also marked effect on the relative RMSE of the growth estimates for both tree species. The relative RMSE decreased 3% for pines and 5% for spruces when the weight of the variable increased from 1 to 10 for pines and from 1 to 6 for spruces.

The number of nearest neighbours (k) had the largest effect on the accuracy of growth estimates. The relative RMSE of the growth estimates of pine varied from 65% to 50% when the number of nearest neighbours varied from 1 to 20. The difference was larger for spruce. The relative RMSE of the growth estimates was 90% with 1 nearest neighbour and 65% with 15 nearest neighbours. Determination of the optimal number of nearest neighbours was not simple. The appropriate number of nearest neighbours were found to be over 10. When the number of nearest neighbours increased over 10, the errors decreased slightly (Fig. 3). On the other hand the standard error increased rapidly when the number of near-

Table 3. Reliability of the diameter growth and thickness of the bark predictions of the k-nearest neighbour and k-MSN methods and mixed models.

	Growth model		Bark model	
	Scots pine	Norway spruce	Scots pine	Norway spruce
k-nn method				
Number of the neighbours (*k*)	15	15	15	15
RMSE, mm	4.98	3.66	4.38	2.76
RMSE, %	49.5	65.8	40.9	25.3
Bias, mm	–0.14	0.21	0.03	0.12
Bias, %	–1.5	3.7	0.4	1.1
k-MSN method				
Number of the neighbours (*k*)	15	14	15	14
RMSE, mm	4.71	3.80	6.19	3.92
RMSE, %	47.6	69.7	55.6	35.6
Bias, mm	0.03	0.32	–0.39	–0.02
Bias, %	0.3	5.8	–3.5	–0.2
Mixed model				
RMSE, mm	8.23	3.41	4.17	2.66
RMSE, %	75.3	58.3	38.7	23.5

est trees decreased. The relative bias of the growth estimates of both tree species did not vary much with different number of nearest neighbours.

The die-off parameter (*pm*) which determines how quickly the weights of the nearest trees decrease when distance d_{ij} increases, did not have much effect on the reliability of the growth estimates of Scots pine (Fig. 3). With 15 nearest neighbours, the variation of relative RMSE and bias was only 0.5% when the values of the die-off parameter varied from 1 to 5. The die-off parameter affected the accuracy of growth estimates of Norway spruce when more than 3 nearest neighbours were used (Fig. 3). Small values of *pm* gave smaller standard errors and biases. With 15 nearest neighbours the relative standard error increased almost 6% when the value of the die-off parameter increased from 1 to 5.

The absolute and relative standard errors and biases were minimized when the value of the die-off parameter in the distance function were $pm=3$ for pine and $pm=1$ for spruce and the number of nearest neighbours was 15 for both tree species (Table 2). The root mean square error of the growth estimates was 4.98 mm for pine and 3.66 mm for spruce and the corresponding relative RMSE was 49.5% and 65.8%, respectively (Table 3). Bark thickness at the end of the growth period was calculated as the mean of the same reference trees as the growth. The absolute RMSE value of the bark estimates was 4.38 mm for pine and 2.76 mm for spruce and corresponding relative RMSE was 40.9% for pine and 25.3% for spruce (Table 3).

The results were slightly biased for both species in the k-nn method. The growth model of pine slightly overestimated the average diameter growth. The average growth of spruce was a slight underestimate (Table 3). The relative biases of the growth estimates versus diameter classes are presented in Fig. 4. The estimates are most accurate for the diameter classes with high frequency. The relative standard error increases especially for Norway spruce when the diameter increases. The greater variation in mean residuals in the largest diameter classes probably are due to low number of observations. The k-nn models did not, however, result in systematic over- or underestimates for large trees.

3.1.2 The k-MSN Method

In the k-nearest neighbour method, the maximum number of variables in the distance function can not be very high because of the enormous number

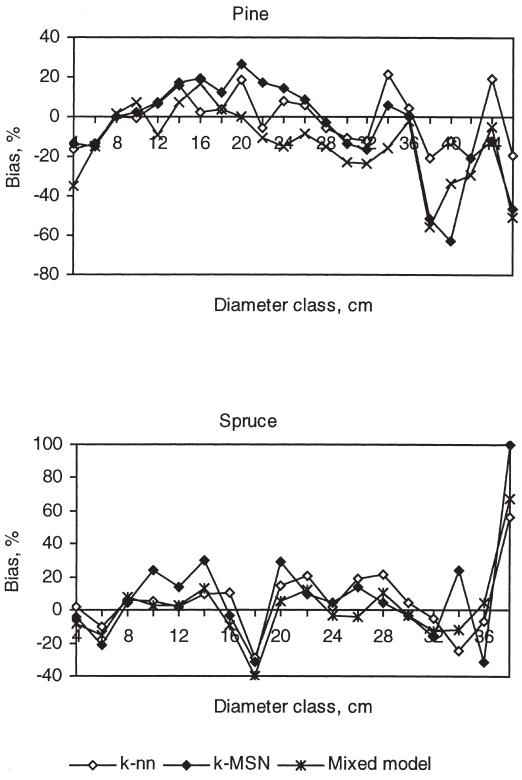

Fig. 4. Relative biases (b_e%) of the growth estimates of Scots pine and Norway spruce in relation to diameter classes.

of iterations required for heuristical searching for optimal parameters. In the k-MSN method all possible independent and dependent tree and stand level variables can be used in the calculations of canonical correlations. The variables used in modelling included e.g. tree diameter, height, tree basal area and relative tree size. Stand age, basal area of the stand, basal area mean diameter, altitude and temperature sum were tested as stand level variables. Site types were tested as dummy variables. All the possible combinations of the chosen variables were tested, but the RMSE and bias of the growth estimates were clearly better when only tree diameter, tree height, stand age at breast height and basal area of trees larger than the subject tree were used as independent variables. Correspondingly, diameter growth was chosen to be the only dependent variable. Canonical coefficients of the chosen independent variables (Γ) and squared canonical correlation (Λ^2) are presented in Table 2.

The value of the k most similar neighbours (k) varying from 1 to 20 were also considered in the calculations of the k-MSN growth estimates. The influence of the k-value was similar with the k-nearest neighbour method. The RMSE of the growth estimates decreased when the number of nearest neighbours increased (Fig. 3). Satisfactory results were obtained when the number of the neighbours was 15 for pines and 14 for spruces (Table 2).

The accuracy of the k-MSN estimates with 15 nearest neighbours was slightly better than k-nn estimates with 15 nearest neighbours for Scots pine, but worse for Norway spruce with 15 neighbours in the k-nn method and 14 in the k-MSN method (Table 3). The reliability of the bark thickness estimates was worse in the k-MSN method. Relative biases were in general higher in the k-MSN method in relation to diameter classes. The k-Most Similar Neighbour method produced clear overestimates for large pines (Fig. 4). Different transformations were tested in order to reduce the bias of the estimates, including diameter squared and inverse of the stand age as an independent variable. In both methods, the accuracy of the growth estimates decreased when the transformed variable was used as independent variable.

We attempted to reduce the prediction bias also by using different numbers of nearest neighbours for small and large trees in the k-MSN method. Smaller numbers of nearest neighbours were tested for small and large trees than for middle-sized trees. The standard error and bias of the estimates were remarkably larger with less than 5 nearest neighbours at the extremes of the data for both tree species. Residuals of the estimates were larger for small and large trees if the number of neighbours was too small. In the case of pines, the estimates were most reliable when the number of nearest neighbours was 5 for trees with diameter greater than 20 cm and 15 otherwise. In the case of spruces, the most accurate results were obtained when large (d>20 cm) and small trees (d<5 cm) had 7 neighbours and middle-sized trees 15 neighbours. The results where similar as with equal number of nearest neighbours, in both the tree and stand levels.

At the stand level, fewer nearest neighbours increased the variation of residuals in stands with large basal areas.

3.1.3 Comparison of the Non-Parametric Methods and Mixed Model

The non-parametric diameter growth models were compared to the regression models constructed from the same study data using mixed model technique. The same tree and stand characters as in the non-parametric methods were found to be the most reliable growth predictors and were used as the independent variables in the regression model. Logarithmic diameter growth was used as an independent variable and logarithmic transformations were also used for tree diameter, tree height and stand age. The coefficients of the independent variables and the values of the random parameters are presented in Table 2. The mixed model gave better results for spruces than non-parametric methods, but the accuracy of Scots pine growth estimates was much lower. The standard errors of the regression estimates were 52% for spruces and 73% for pines (Table 3). The regression model produced more accurate growth estimates for Norway spruce than the non-parametric methods, but the relative RMSE of the growth estimates of pine was 20% lower. The regression model overestimated the growth of the pines with diameter larger than 20 cm and produced large overestimates for the largest pines (Fig. 4).

3.2 Stand Level Growth Estimates

The k-nearest neighbour regression was found to be more reliable than the k Most Similar Neighbour method at the stand level. The measured mean stand volume at the end of the growth period was 125 m^3/ha and mean volume growth 13.8 m^3/ha. The k-nn method gave volume and growth estimates almost equal to true values, while both were 3 m^3/ha smaller for the k-MSN method. The relative RMSE of the stand growth was 39.8% for the k-nn method and 67.1% for the k-MSN method (Table 4). The relative biases of the k-nn and MSN methods were 1.5% and 29.3%, respectively. The relative errors and biases were calculated by dividing the absolute values by the predicted growth. The relative RMSE of the k-MSN volume growth was almost 20% lower when the absolute error was divided by the true mean volume growth of the stands, which is also often used as a test criterion.

Table 4. Accuracy of the 5-year stand growth (IV_5) estimates of the non-parametric methods and regression models.

	k-nn method	k-MSN method	Mixed model	Monsu
Mean IV_5, m^3/ha	13.7	10.7	15.7	13.7
RMSE, m^3/ha	5.4	7.2	11.5	9.8
RMSE, %	39.8	67.1	73.1	71.5
Bias, m^3/ha	0.2	3.2	–1.9	0.2
Bias, %	1.5	29.3	–11.8	1.3

The k-Most Similar neighbour underestimated the stand volume growth more than the k-nearest neighbour method (Fig. 5). The k-nearest neighbour method also seemed to predict the volume growth better at both extremes where the edge effect usually influences results. Both methods underestimated the volume growth in the stands with the largest basal areas. However, the results for the stands with the smallest basal areas were not systematically over- or underestimated. The accuracy of the k-MSN method improved 10%, when the quite evident outlier stand were removed (see Fig. 5). The relative RMSE of the stand growth estimates were then 57% for the k-MSN method and 36% for the k-nn method.

Comparison of the stand growth estimates showed that the regression model was less accurate at the stand level than the non-parametric growth models. Especially the estimates of the k-nn method were more reliable. The relative standard error of the stand growth estimates of the mixed model was 73.1%, while it was 39.8% for the k-nn method and 67.1% for the k-MSN method. The regression model overestimated the 5-year stand growth by 1.9 m^3/ha. The stand level growth estimates were also compared to the volume growths produced with the Monsu-forest planning program (Pukkala 2000). The program uses single-tree regression growth models devel-

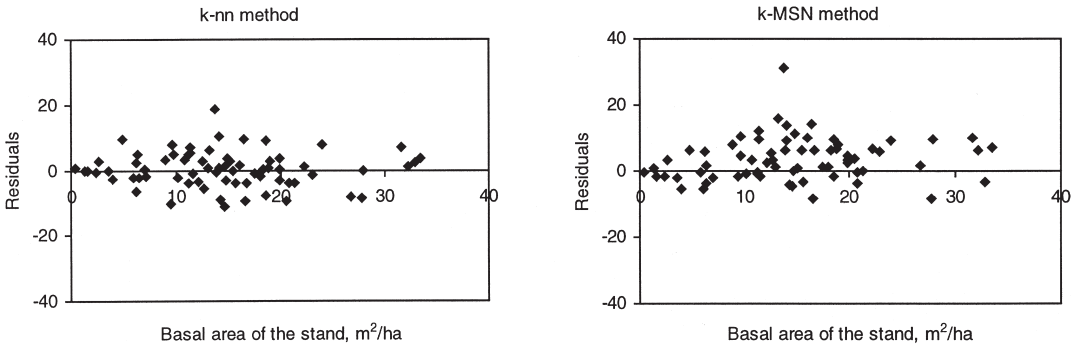

Fig. 5. Residuals of the stand volume growths in the k-nearest neighbour and k-MSN methods in relation to the basal area of the stand.

oped by Nyyssönen and Mielikäinen (1978). The absolute standard error of the growth estimates of Monsu was 9.8 m^3/ha and corresponding relative RMSE was 71.5%. When compared to non-parametric methods, the models used in Monsu underestimated the volume growth of stands with small basal area and produced larger overestimates especially in the stands of average densities.

4 Discussion

The aim of this study was to construct individual diameter growth models with non-parametric k-nearest neighbour and k-Most Similar Neighbour methods. The growth models were built for 5-year growth period. In addition to the growth models, bark thickness at the end of the growth period was predicted. The nearest trees were selected using tree diameter, tree height, stand age at breast height and basal area of the trees larger than the subject tree. Tree diameter had relatively more weight in the k-nn method than in the k-MSN method in which the coefficients of the variables are obtained by means of the canonical correlation analysis. This may have caused bias to the k-MSN estimates. Correspondingly, basal area larger than the subject (G_{large}) had relatively much larger weight in the k-MSN method especially in the distance function of Norway spruces. G_{large} had to have small weight in the distance function of the k-nn method, because the RMSE of the growth estimates increased notably if the weight of the variable increased. Other variables describing the position of a tree in the stand were tested in the k-MSN method, but without G_{large} the relative RMSE was at minimum 6% higher. Stand age had much weight in the distance function relative to other variables in both methods for Norway spruces. In this case, the nearest neighbours were selected among neighbouring stands with as similar age as possible.

If the study data had been larger, also other important variables would definitely have been found to describe locality and improved the results. Especially stand level variables do not have enough variation in small data sets. Increasing the number of independent variables also reduces the number of potential neighbours. The number of the nearest neighbours had a greater effect on standard errors of the estimates than the values of the coefficients of the variables. The influence of k-value was similar in both methods. Increasing the k-value from 1 to 10 improves greatly the accuracy of the growth estimates and increasing the number of the neighbours beyond $k=10$ improves slightly the accuracy. The appropriate number of nearest neighbours was found to be 14–15. The relative biases of the growth estimates were largest with 1 nearest neighbour in the k-MSN method and with 3 nearest neighbours in the k-nn method. The bias of the pine growth estimates reduced slightly when the k-value was

increased from 1 to 16 and beyond that the bias slightly increased. The bias of the spruce growth estimate varied more with different *k*-value and therefore the exact number of nearest neighbours was not simple to decide.

Both methods gave slightly biased results for diameter growth. The bias of the estimates increased when the tree diameter increased. There were few large trees in the data, only 12% of the sample trees had diameter larger than 25 cm. For that reason, in most cases, the nearest neighbours of large trees were middle sized trees. This could be partly avoided by increasing the weight of the diameter in the distance function for large trees. One possibility to try to reduce the trend in bias is to use transformations for the independent variables. This could reduce the bias, if the correlation between transformed variable and diameter growth is more linear than the correlation of diameter growth and original variable. However, in the study data the effect of transformations was small and did not improve the results.

The structure of the study data affected the reliability of the applied methods. The restrictions of the study material had a strong influence on the results. Especially the scarcity of trees with diameter over 20 centimetres probably caused biased predictions. The data applied in non-parametric methods should be evenly distributed, but it should also include exceptional observations, e.g. exceptionally large trees. Variability of the characters, such as stand basal area, mean diameter and dominant height, would increase if the study data consisted of more stands. The amount of possible neighbouring observations would be higher and more realistic estimates could be obtained.

The results of this study indicate that especially the k-nearest neighbour regression can be a competitive growth prediction method. The k-Most Similar Neighbour and k-nearest neighbour methods seemed to be almost equally reliable, when the accuracy of individual tree growth estimates was analysed. However, the stand level growth estimates were much more reliable for the k-nn method. The k-MSN method underestimated the volume growth especially in the stands with large volume growth and in old stands more than the k-nn method. The k-MSN method also produced more biased growth estimates for large trees. The k-MSN method expects linearity between dependent and independent variables, because the coefficients of the variables are obtained by using canonical correlation analysis. The k-nearest neighbour method is more robust, but the heuristical search of the values of the coefficients is very time consuming. In the k-MSN method the values of the coefficients are found easily and fast and there can be many independent and dependent variables. However, the heuristic searching method is not the only alternative when using the k-nearest neighbour regression, but the parameters in the distance function could also be searched using non-linear regression (Niggemeyer and Schmidt 1999) or numerical optimization (see e.g. Miina and Pukkala 2000).

In this study the simulation of stand development was done only for one 5 year growth period. However, in many applications predictions of longer growth periods are needed. In these situations, the non-parametric methods can be applied in principle like traditional individual tree growth models, i.e. growth is simulated separately in 5 years periods. Another possibility is to utilise long growth series, if such exists. Then the simulation of stand development could be done for the whole growth period. Instead of predicting treewise diameter or height growth all stand characteristics of interest could be obtained simultaneously (see Maltamo and Eerikäinen 2000).

Although the growth models of this study were constructed only for regional use in Finland, the non-parametric methods have wide application possibilities. The use of the non-parametric methods is efficient especially for tree characteristics which vary locally or in time. Such characteristics are for example tree growth, stem form and log reduce. The problems which occur when applying common parameter models can be reduced if local data are available. Correspondingly, in conditions quite different than in Finland, e.g in Africa, the non-parametric growth and yield models could be constructed for plantations, different growing densities or seed origins.

The nearest neighbour methods can be further applied in semiparametric models, which are combinations of ordinary regression models and non-parametric models. In semiparametric models, variables with clear relations are estimated with linear models and the remaining part

of the model is fitted with non-parametric methods. Non-parametric and semiparametric models can also be constructed by applying non-parametric generalized additive models. The usefulness of such models is one potential direction for future work.

References

Altman, N.S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46(3): 175–184.

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

Gertner, G.Z. 1984. Localizing a diameter increment model with a sequential Bayesian procedure. Forest Science 30: 851–864.

Gustavsen, H.G. 1977. Valtakunnalliset kuutiokasvuyhtälöt. Summary: Finnish volume increment functions. Folia Forestalia 331. 37 p.

— 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.

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

Härdle, W. 1989. Applied nonparametric regression. Cambridge University Press. 323 p.

Hynynen, J. 1995. Modelling tree growth for managed stands. 1995. The Finnish Forest Research Institute, Research Papers 576. 59 p.

Kangas, A. & Korhonen, K.T. 1995. Generalizing sample tree information with semiparametric and parametric models. Silva Fennica 29(2): 151–158.

Korhonen, K.T. 1993. Mixed estimation in calibration of volume functions of Scots pine. Silva Fennica 27(4): 269–276.

— 1994. Calculation system for large scale forest inventory. The Finnish Forest Research Institute, Research Papers 505. 36 p.

— & Kangas, A. 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scandinavian Journal of Forest Research 12: 97–101.

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

— 1993. Metsäbiometrian menetelmiä. Study book. University of Joensuu. Silva Carelica 24. 182 p.

Maltamo, M. & Eerikäinen K. 2000. Non-parametric growth and yield model for Pinus kesiya in Zambia. In: Pukkala, T. & Eerikäinen, K. (eds.). Establishment and management of tree plantations Southern and Eastern Africa. University of Joensuu, Faculty of Forestry, Research Notes 120: 81–99.

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

Mielikäinen, K. 1992. Growth models for predicting stand development. In: Salminen, H. & Katermaa, T. (eds.). Simulation of Forest Development. The Finnish Forest Research Institute, Research Papers 407: 10–14.

— & Gustavsen, H.G. 1992. The empirical basis for tree and stand modelling in Finland. In: Nyyssönen, A., Poso, P. & Rautala, J. (eds.). Proceedings of Ilvessalo Symposium on National Forest Inventories, Finland, 17–21 August 1992. The Finnish Forest Research Institute, Research Papers 444: 179–184.

Miina, J. & Pukkala, T. 2000. Using numerical optimization for specifying individual-tree competition models. Forest Science 46: 277–283.

— , Kolström, T. & Pukkala, T. 1991. An application of a spatial growth model of Scots pine on drained peatland. Forest Ecology and Management 41: 265–277.

Moeur, M. & Stage, A.R. 1995. Most similar neighbor. an improved sampling inference procedure for natural resource planning. Forest Science 41(2): 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, Michigan. p. 419–430.

Niggemeyer, P. & Schmidt, M. 1999. Estimation of the diameter distributions using the k-nearest neighbour method. In: Pukkala, T. & Eerikäinen, K. (eds.). Growth and yield modelling of tree plantations in South and East Africa. University of Joensuu, Faculty of Forestry. Research Notes 97: 195-209.

Nyyssönen, A. & Mielikäinen K. 1978. Metsikön kasvun arviointi. Summary: Estimation of stand increment. Acta Forestalia Fennica 60. 17 p.

Ojansuu, R., Hynynen, J., Koivunen, J. & Luoma, P. 1991. Luonnonprosessit metsälaskelmassa (MELA) – Metsä 2000-versio. The Finnish Forest Research Institute, Research Papers 385. 42 p.

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.

Pukkala, T. 1989. Predicting of diameter growth in even-aged Scots pine stands with a spatial and non-spatial model. Silva Fennica 23(2): 101–116.

— 2000. Monsu-metsäsuunnitteluohjelma. Ohjelmiston toiminta ja käyttö. User's manual. Joensuu.

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

Vuokila, Y. 1965. Functions for variable density yield tables of pine based on temporary sample plots. Communicationes Instituti Forestalis Fenniae 63(2). 86 p.

Total of 30 references