The Most Similar Neighbour Reference in the Yield Prediction of *Pinus kesiya* Stands in Zambia

Matti Maltamo and Kalle Eerikäinen

Maltamo, M. & Eerikäinen, K. 2001. The Most Similar Neighbour reference in the yield prediction of *Pinus kesiya* stands in Zambia. Silva Fennica 35(4): 437–451.

The aim of the study was to develop a yield prediction model using the non-parametric Most Similar Neighbour (MSN) reference method. The model is constructed on stand level but it contains information also on tree level. A 10-year projection period was used for the analysis of stand growth. First, the canonical correlation matrix was calculated for the whole study material using stand volumes at the beginning and at the end of the growth period as independent variables and stand characteristics as dependent variable. Secondly, similar neighbour estimates were searched from the data categories reclassified according to thinnings. Due to this, it was possible to search for growth and yield series which is as accurate as possible both at the beginning and at the end of the growth period. The reliability of the MSN volume predictions was compared to the volumes predicted with the simultaneous yield model. The MSN approach was observed to be more reliable volume predictor than the traditional stand level yield prediction model both in thinned and unthinned stands.

Keywords difference equations, non-parametric regression, plantation forests, stand development

Authors' address Faculty of Forestry, University of Joensuu, P.O. Box 111, FIN-80101 Joensuu, Finland **Tel.** + 358 13 251 3615 **E-mail** matti.maltamo@forest.joensuu.fi

Received 5 May 2000 **Accepted** 6 November 2001

1 Introduction

Stand level growth and yield predictions are traditionally based on stand yield tables, static regression models and growth simulators derived from re-measurement data. If the only stand characteristic of interest is the total stem volume of the stand, a straightforward stand yield prediction model with separate but simultaneously estimated static difference equations for different stand characteristics is a very good choice (e.g. Borders and Bailey 1986, Borders 1989, Pienaar and Harrison 1989). Simultaneous estimation of separate prediction models guarantees that their relations remain logical even in the extremes of observations and in the case of extrapolations. The disadvantage is the loss of information on structural stand characteristics, even if the mean stand characteristics are predicted accurately. A traditional alternative for the difference equation approach is a separate prediction of static relation models for the development of stand mean characteristics over age (see e.g. Clutter et al. 1983).

More information about the size distribution of trees is obtained when the growth model is structured on the development of diameter or basal area distribution (e.g. Borders and Patterson 1990, Maltamo and Kangas 1998). When changes in stand structure are analysed on the tree level, one alternative is to use single tree growth or diameter transition models (e.g. Pukkala 1989, Kolström 1992, Hynynen 1995).

Non-parametric methods are an alternative for the traditional approaches based on regression models for stand characteristics. Non-parametric methods predict the value of the variable in question as a weighted average of the values of neighbouring observations, the neighbours being defined with the predicting variables (e.g. Härdle 1989, Altman 1992). The chosen neighbours are selected from a database of previously measured observations.

Non-parametric methods such as the k-nearest neighbour, Most Similar Neighbour (MSN), kernel and grid have been utilised in several forestry applications (Holm et al. 1979, Kilkki and Päivinen 1987, Kangas and Korhonen 1995, Moeur and Stage 1995, Korhonen and Kangas 1997, Tommola et al. 1999). These studies include, for example, the generalisations of sample tree information, estimation of characteristics of a marked stand and applications of multisource and multivariate forest inventories. Applications considering the smoothing and prediction of diameter distributions have also been presented with different non-parametric methods (Droessler and Burk 1989, Haara et al. 1997, Maltamo and Kangas 1998, Maltamo and Uuttera 1998, Niggemeyer and Schmidt 1999).

The advantages of non-parametric methods are that they retain the full range of variation of the data as well as the covariance structure of the population (Moeur and Stage 1995). Because estimates are chosen directly among actual samples, no unrealistic predictions can occur. Furthermore, the estimates for the characteristics to be predicted are obtained in all situations where at least some measurements are available (e.g. Haara et al. 1997).

A disadvantage of non-parametric methods is the requirement for reference material also in the application phase. Moreover, it is not guaranteed that non-parametric regression estimates are unbiased (Altman 1992). However, if a priori information such as forest type classification is available, the bias of the sub-area estimation can be reduced and more representative neighbours can be chosen (Tokola 2000).

The most frequently applied non-parametric method is the k-nearest neighbour method. In this method, the estimator uses a neighbourhood consisting of a constant (*k*) number of observations, but the width of the neighbourhood may vary. When applying the k-nearest neighbour method, the form of a distance measure must be specified to define the neighbourhood of a given point (e.g. Korhonen and Kangas 1997). The distance function used can, for example, be based on the differences of mean stand characteristics. A closely related method to the basic k-nearest neighbour regression is the most similar neighbour (MSN) reference (Moeur and Stage 1995, Moeur and Riemann Hershey 1999). In this case the coefficients of the variables in the distance function are searched using canonical correlation. The benefit of the MSN method is that all possible independent (stand mean characteristics etc.) and dependent (stand variables of interest) information can be used in the calculation of canonical correlation. The enormous number of iterations in the search for nearest neighbours can also be avoided. However, as an disadvantage of the MSN reference, a linear correlation between dependent and independent variables is assumed (Moeur and Stage 1995).

In the k-nearest (similar) neighbour methods the number of nearest neighbours must also be defined. The larger the number of chosen neighbours is, the more average the results are (Altman 1992). The bias of the k-nearest neighbour estimator can be reduced using weighted averages, which are defined using weights, expressed as a function of distance (Altman 1992).

The aim of this study is to develop a yield prediction model for *Pinus kesiya* (Royle ex Gordon) using the MSN regression. The principle of the model is that all information on tree tally during a growth period is directly obtained from nearest neighbour estimates. The results obtained with the MSN reference are compared to the system of simultaneously estimated difference equations.

2 Material and Methods

2.1 Data

P. kesiya is the most important commercial plantation species of Zambia (e.g. Armitage and Burley 1980). Most of Zambian *P. kesiya* plantations are located on the Copperbelt region (Saramäki 1992). In 1999, there were approximately 26 000 hectares of planted *P. kesiya* that is nearly half of the total area under commercial plantations in Zambia (Sekeli and Phiri 1999). In terms of the planted net area on *P. kesiya* the most important country in the world is Madagascar, while Zambia is the second.

In the African forest plantations, *P. kesiya* has been grown for sawn timber, poles, pulpwood but also for the production of firewood and resin (Mbuya et al. 1994). In southeastern Africa , the rotation period for the sawn timber production of *P. kesiya* varies between 30 and 50 years.

The study data were collected from Permanent Sample Plots (PSP) from the forest plantations of Chati and Ndola Hill in the Copperbelt region (about 13°S, 28°E, 1200–1300 m a.s.l.). All stands of the data were managed. The data consisted of trials measured from stands with two different initial planting densities. A total of 140 circular sample plots were measured from stands with a density of 1328 trees per hectare (spacing: 2.74 m). Only two trials were measured from stands with a planting density of 1076 trees per hectare (spacing: 3.05 m). For the planting densities of 1328 and 1076 trees per hectare the number of planting spots per sample plot was 52 and 44, respectively. Therefore, the computational plot size was about 400 m^2 for the sample plots of both planting densities.

According to the measuring instructions for the sample plot assessments (Systems of measurements … 1969), all diameters (mm) at the breast height of living trees and heights (dm) of sample trees were measured. At least one tree per one centimetre diameter class was selected as a height sample tree. For the calculations of tree and stand volumes, heights were needed for all trees. The height development trends of sample trees were generalised for the rest of the trees using a mixed height prediction model with a treewise calibration component (Eerikäinen 1999). Parameters of the height model were determined for all plots separately.

2.2 Data Preparation

The stand dominant height (dm) was defined in each of the assessments as the mean height of the 100 largest trees per hectare, according to diameter at breast height. Total stand volumes were calculated by predicting stem volumes for all trees on the plot list. Stem volumes over bark were obtained using the volume function of Eerikäinen (2001) for *P. kesiya* in Zambia, Tanzania and Zimbabwe.

To get as comparable as possible growth periods for the application of the MSN reference, the PSP data were restricted so that a growth period of only about 10 years was available for each sample plot. The age of the tree stock varied from 7.4 to 9.9 years at the beginning of the growth period (Table 1). The chosen 142 sample plots were measured 3 to 5 times during the growth period resulting a total of 474 growth periods.

Although the PSP data included thinned sample plots, the rate and timing of thinnings were not registered. Therefore the PSP data were reclassified into three categories according to the timing and intensity of thinnings. If the decrease in the number of stems between the successive measurements was greater than 15 percent the reduction was treated as a thinning. Otherwise the decrease was understood to be a sign of natural mortality, i.e. self-thinning. The timing of the thinnings was also taken into account in the grouping of the data. The sample plots, which were thinned during the first half of the 10 year period, were separated from trials thinned during the latter 5 year period. The first thinning category was for the data with no thinnings. The second category was for the PSPs that were thinned during the first 5 year growth period. If the trial was thinned during the

Table 1. Mean stand characteristics of the Permanent Sample Plot data ($n = 142$) at the beginning and at the end of the growth period.

	$\hat{SI}$	T_1	T_2	H_{dom_1}	H_{dom_2}	N_1	N_2	G_1	G_2	V_1	V_2
Combined data											
$n = 142$											
Median	24.3	8.7	18.7	13.2	24.8	1073.0	613.0	23.6	33.3	133.6	324.8
s.d.	2.424	0.692	0.616	1.607	1.940	203.040	207.069	4.203	9.613	33.703	99.029
Minimum	18.1	7.4	17.7	9.8	18.5	511.0	178.0	13.8	13.1	70.2	122.7
Maximum	31.1	9.9	19.8	17.8	29.9	1328.0	1124.0	34.2	58.7	232.7	562.7
Thinning group 0											
$n = 25$											
Median	26.6	8.8	17.9	15.1	25.0	996	868	25.4	47.1	172.0	470.7
s.d.	2.449	0.622	0.544	1.810	2.038	204.769	187.789	6.144	5.882	48.637	58.072
Minimum	21.8	7.7	17.8	11.4	21.7	511	434	13.8	33.2	70.2	324.1
Maximum	31.1	9.8	19.8	17.8	29.9	1277	1124	34.2	58.7	232.7	562.7
Thinning group 1											
$n = 34$											
Median	24.1	7.8	18.7	12.4	25.5	1201	728	24.5	39.6	129.7	396.9
s.d.	2.300	0.609	0.875	1.636	1.582	164.074	145.986	3.712	6.387	29.561	71.468
Minimum	19.7	7.4	17.7	9.8	21.4	587	357	17.4	24.9	77.8	230.6
Maximum	29.1	9.6	19.8	17.2	29.4	1328	945	32.3	53.5	194.6	548.3
Thinning group 2											
$n = 83$											
Median	24.0	8.7	18.8	13.1	24.3	1022	485	22.8	27.7	130.7	258.5
s.d.	2.147	0.696	0.469	1.215	1.942	194.576	139.019	3.483	6.135	25.460	62.304
Minimum	18.1	7.4	17.7	9.9	18.5	587	178	15.0	13.1	81.2	122.7
Maximum	29.8	9.9	19.8	15.3	28.3	1328	766	32.2	43.7	199.5	427.4

Explanation of the variable codes (the index i refers to the beginning ($i = 1$) and end ($i = 2$) of the growth period): $\hat{SI}$ = Site Index, i.e. predicted stand dominant height (Equation 6) at the index age of 18 years, m; T_i = stand age, years; H_{dom_i} = stand dominant height, m; N_i = number of stems, ha^{-1}; G_i = stand basal area, m^2; V_i = total stand volume over bark, m^3; n = number of sample plots; s.d. = standard deviation.

second 5 year growth period, or in both periods, it was classified into the third thinning category.

2.3 Growth and Yield Modeling

2.3.1 The MSN Reference Method

The growth and yield model is based on the MSN reference, where the k-most similar sample plots are used for predicting the growth series for a ten year period. The growth series includes full description of tree stock, i.e. diameter distribution, tree heights, mortality and thinnings. The same principles that were used in the study by Moeur and Stage (1995) were also applied in this study. The MSN reference is based on canonical correlation between independent and dependent variables. However, the number of neighbours is also included in the calculations.

The independent stand variables used in the calculation of canonical correlations include number of stems per hectare, stand age and dominant height at the beginning of the growth period and site index, i.e. dominant height predicted at the index age of 18 years. In addition, the effect of stand basal area at the beginning of the growth period as independent variable was also examined. The stand volume both at the beginning and at the end of the growth period were chosen as the dependent variables. The usage of both volumes was based on the requirement that the constructed growth and yield series should be as accurate as possible for the volume both at the beginning and at the end of the growth period. The independent and dependent variables were standardised, i.e. by subtracting the mean of the variable and dividing it by the standard deviation of the variable, to prevent the unit of measurement from influencing the distance.

When calculating the MSN estimates, the target sample plot is the one for which the nearest neighbour estimates are calculated. Each sample plot, in turn, is used as the target sample plot and the target sample plot is temporarily excluded from the reference sample plots. The number of reference stands (k) varied from 1 to 15 in the calculations. The estimates were calculated especially for thinning categories, which were reclassified using a priori information, but also for all of the study material.

In the MSN reference, the most similar neighbours to the target sample plot u are chosen from the reference sample plots, for which $(\hat{Y}_u - Y_j)W$ $(\hat{Y}_u - Y_j)$ is minimised over all $j = 1,\ldots,n$ reference sample plots, where $\hat{Y}_u$ is a row vector of the unknown variables of the target sample plot, Y_j is a row vector of the observed variables of the reference sample plots and W is a weighting matrix. In applications the relation of unknown and observed variables is not known. Therefore, the corresponding relation of independent variables, which are known both in the target and reference sample plots, is used. According to Moeur and Stage (1995), the canonical correlation analysis is used in the calculation of the weighting matrix in the distance function by summarising the relationships between dependent (Y) and independent (X) variables simultaneously.

When using the canonical correlation analysis, linear transformations U_r and V_r are formed from the set of dependent and independent variables such that the correlation between them is maximised:

$$U_r = \alpha_r Y \text{ and } V_r = \gamma_r X \tag{1}$$

where α_r and γ_r are canonical coefficients of the dependent and independent variables ($r = 1,\ldots,s$). There are s possible pairs of canonical variates (U_r and V_r) as the results of the analysis, where s is either the number of dependent or independent variables, depending on which is smaller (Moeur and Stage 1995). In our study, the number of dependent variables is most often two, whereas the number of independent variables is usually at least four, i.e. $s = 2$. Canonical variates are ordered in such a way that canonical correlation between them is the largest for the first variate (U_1,V_1). Thus, the predictive relationship between original variables is concentrated in the first few canonical variates (Moeur and Stage 1995). The calculation of canonical correlation was done using IMSL library.

In our study, the distance was measured using the Mahalanobis distance formula:

$$D_{uj}^2 = (X_u - X_j)\Gamma\Lambda^2\Gamma'(X_u - X_j)' \tag{2}$$

where X_u are independent variables of the target sample plot, X_j are independent variables of reference sample plots, Γ is the matrix of canonical coefficients of the independent variables, and Λ^2 is the diagonal matrix of squared canonical correlations.

The final estimate ($\hat{z}_u$) for the growth and yield series of sample plot u was calculated as the weighted average of the growth and yield series of reference sample plots (z_j):

$$\hat{z}_u = \frac{\sum_{u=1}^{k} \frac{1}{D_{uj}} z_j}{\sum_{u=1}^{k} \frac{1}{D_{uj}}} \tag{3}$$

The test criteria used in the comparison of different model forms and number of neighbours were Root Mean Square Error (RMSE) and bias of predicted stand volumes. The RMSE of predicted stand volumes was:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(V_i - \hat{V}_i\right)^2}{n}} \tag{4}$$

where n is the number of sample plots, V_i is the true volume of sample plot i and $\hat{V}_i$ is the volume of sample plot i estimated from the predicted distribution. The bias of the predicted volumes was calculated with the formula:

$$\text{bias} = \frac{\sum_{i=1}^{n}\left(V_i - \hat{V}_i\right)}{n} \tag{5}$$

The relative RMSE and bias of the volume estimate were calculated by dividing the absolute RMSE by the true mean volume $\overline{V}$ of the stands.

Table 2. Stand characteristics of 322 growth periods with no thinnings at 142 permanent sample plots used in the modelling of the simultaneous yield model (Equations 6–10).

	T_1	T_2	H_{dom_1}	H_{dom_2}	N_1	N_2	G_1	G_2	V_1	V_2
Median	11.7	14.7	17.6	21.1	817.0	792.0	27.8	34.1	200.3	285.1
s.d.	2.876	3.174	4.159	3.818	236.681	230.960	7.319	7.498	87.756	93.429
Minimum	7.4	9.1	9.8	12.7	306.0	306.0	13.8	19.9	70.2	122.7
Maximum	16.8	19.8	26.8	29.9	1328.0	1303.0	50.9	58.7	475.4	562.7

2.3.2 Projection Models

The reliability of the MSN method as a total stand volume predictor was analysed by comparing it to a simultaneous stand level yield model. Since both of these methods are based on projections of stand characteristics at the end of the growth period, the system of simultaneous difference equations by Pienaar and Harrison (1989) was chosen as a reference method for the MSN (see also Miina et al. 1999). If an estimated parameter of the simultaneous yield model by Pienaar and Harrison (1989) was not significant, then it was removed from the final model, which is as follows:

$$H_{\mathrm{dom}2} = H_{\mathrm{dom}1} \cdot \left(\frac{1 - e^{(\mathrm{c}_{11} \cdot T_2)}}{1 - e^{(\mathrm{c}_{11} \cdot T_1)}} \right)^{\mathrm{c}_{12}} + \varepsilon_1 \tag{6}$$

$$\begin{aligned} \ln(G_1) = \mathrm{c}_{20} + \mathrm{c}_{21} \cdot \frac{1}{T_1} + \mathrm{c}_{22} \cdot \ln(N_1) \\ + \mathrm{c}_{23} \cdot \ln(H_{\mathrm{dom}1}) + \mathrm{c}_{24} \cdot \frac{\ln(N_1)}{T_1} + \varepsilon_2 \end{aligned} \tag{7}$$

$$\begin{aligned} \ln(G_2) = \ln(G_1) + \mathrm{c}_{21} \cdot \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \\ + \mathrm{c}_{22} \cdot \left(\ln(N_2) - \ln(N_1) \right) \\ + \mathrm{c}_{23} \cdot \left(\ln(H_{\mathrm{dom}2}) - \ln(H_{\mathrm{dom}1}) \right) \\ + \mathrm{c}_{24} \cdot \left(\frac{\ln(N_2)}{T_2} - \frac{\ln(N_1)}{T_1} \right) + \varepsilon_3 \end{aligned} \tag{8}$$

$$\begin{aligned} \ln(V_1) = \mathrm{c}_{30} + \mathrm{c}_{31} \cdot \ln(N_1) + \mathrm{c}_{32} \cdot \ln(H_{\mathrm{dom}1}) \\ + \mathrm{c}_{33} \cdot \ln(G_1) + \varepsilon_4 \end{aligned} \tag{9}$$

$$\begin{aligned} \ln(V_2) = \ln(V_1) + \mathrm{c}_{31} \cdot \left(\ln(N_2) - \ln(N_1) \right) \\ + \mathrm{c}_{32} \cdot \left(\ln(H_{\mathrm{dom}2}) - \ln(H_{\mathrm{dom}1}) \right) \\ + \mathrm{c}_{33} \cdot \left(\ln(G_2) - \ln(G_1) \right) + \varepsilon_5 \end{aligned} \tag{10}$$

where $\mathrm{c}_{11},\ldots,\mathrm{c}_{33}$ are model parameters to be estimated, and $\varepsilon_1,\ldots,\varepsilon_5$ are random error terms of the models. The other variables are explained in Table 1.

In this study, site indexes, i.e. dominant heights at the index age of 18 years, were obtained with Equation 6 (see Table 1). The stand age, dominant height and number of stems per hectare at the beginning of the growth period were assumed to be measured stand characteristics. Parameters of the above system of equations were estimated with the data set that included only the growth intervals with no thinnings. Altogether 322 unthinned growth periods of 142 sample plots were used, i.e. 152 growth periods were removed from the modelling data (Table 2).

The simultaneous yield model of this study (Equations 6–10) is a system of chained equations where predicted variables of equations are used as independent variables of subsequent equations. Thus, the independent and individually distributed error terms of individual equations were assumed to be contemporaneously correlated and the Three-Stage Least Squares (3SLS) estimator was used in the estimation of parameters (Zellner and Theil 1962, Borders and Bailey 1986). Parameters of Equations 6–10 were estimated simultaneously with the Nonlinear Three Stages Least Squares (N3SLS) regression of the Procedure Model in SAS (SAS Institute Inc. 1993).

3 Results

3.1 Effect of Independent Variables and Number of Neighbours in the MSN Method

The stand volume at the end of the growth period correlates strongest with basal area and projected site index (Table 3). The reason for good correlation for site index is that it is only stand variable on which thinnings do not have direct effects. Number of stems per hectare is the only stand characteristics where the correlation with the stand volume is stronger at the end of the growth period than at the beginning of the growth period.

The accuracy of predicted stand volume varies according to the chosen number of neighbours (Fig. 1). However, if more than four neighbours are used the change in the accuracy of predictions is minor. On average, the accuracy of stand volume prediction produced by the MSN reference after the growth period is over 18% and at the beginning of growth period about 12%. When the biases are considered all figures are minor and increasing number of chosen neighbour does not decrease the residual means of predicted stand volumes (Fig. 2). Furthermore, the biases of predicted stand volumes both at the beginning and at the end of growth period have relatively similar trends (Fig. 2).

Table 3. Correlation matrix between independent and dependent variables.

	T_1	N_1	$\hat{SI}$	H_{dom_1}	G_1	V_1	V_2
T_1	1						
N_1	–0.529	1					
$\hat{SI}$	–0.191	0.043	1				
H_{dom_1}	0.592	–0.362	0.675	1			
G_1	0.212	0.432	0.351	0.461	1		
V_1	0.435	0.118	0.503	0.752	0.919	1	
V_2	–0.147	0.199	0.482	0.300	0.497	0.498	1

In the final model validation, different numbers of neighbours were used for different thinning categories. The selection of number of neighbours was based on the accuracy of predicted stand volumes after the growth period. In this study, the optimal numbers of neighbours for thinning categories 0, 1 and 2 were 10, 8 and 14, respectively (Table 5). The most accurate results were obtained for unthinned category, whereas the poorest results for the third thinning category, where the sample plot was thinned during the second 5 year growth period, or in both periods.

When stand basal area was included as independent variable in the MSN reference the RMSE of predicted stand volumes decreased slightly at the end of the growth period being in minimum

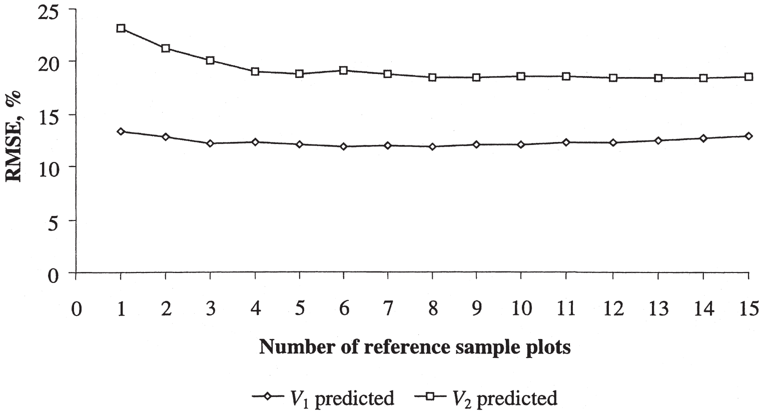

Fig. 1. Relative RMSEs of stand volume predictions at the beginning (V_1) and at the end (V_2) of the growth period in relation to the number of chosen neighbours.

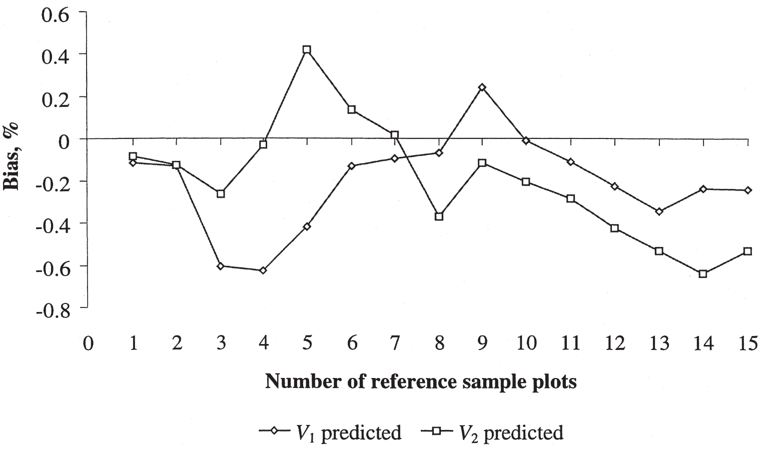

Fig. 2. Relative biases of stand volume predictions at the beginning (V_1) and at the end (V_2) of the growth period in relation to the number of chosen neighbours.

about 16% and notably at the beginning of the growth period being in minimum less than 5%. Some calculations were also carried out where no reclassification according to thinning were done in the search of the nearest neighbours. In this case the accuracy was considerably lower, the RMSE being over 25% at the end of the growth period. Therefore, it seems that the use of the reclassification of data based on thinning information is very important in the construction of the MSN-based yield prediction.

3.2 Growth Projections with the MSN Reference Method

Examples of the MSN based growth and yield series are presented in Figs. 3 and 4. Both examples are determined for the according to thinnings reclassified material using 5 neighbours. The development of the thinned reference sample plots indicates that the effect of thinning can be described by the MSN estimate (Fig. 3). However, the timing of the thinning of two of the reference sample plots is not quite correct. The number of stems in these two reference sample plots are higher than in the other plots during most of the growth periods. In the case of unthinned sample plots, the MSN values are slight underestimates (Fig. 4). There seems to be quite a high variation, especially between the number of stems in the reference sample plots.

3.3 Parameter Estimates and the Reliability of the Simultaneous Yield Model

All the estimated parameters of the simultaneous yield model (Equations 6–10) were statistically significant (Table 4). The adjusted degrees of determination for Equations 6–10 were also very high, i.e. 0.9036, 0.8127, 0.8962, 0.9965, and 0.9941, respectively. Estimates for the residual standard deviations of Equations 6–10 were 1.1853, 0.1100, 0.0695, 0.0241 and 0.0244, respectively. For the bias correction of Equation (10), half of the estimated error variance of the model ($0.0244^2/2 \cong 0.0003$) was added to it before its back-transformation, as suggested by Baskerville (1972).

Aiming at the comparability between the predicted stand volumes of MSN reference and the simultaneous yield model, MSN reference based estimates of the number of stems at the end of the growth period were used as independent variables in Equations 8 and 10. When the yield projection model was tested as the predictor of the total stand volume at the end of the growth

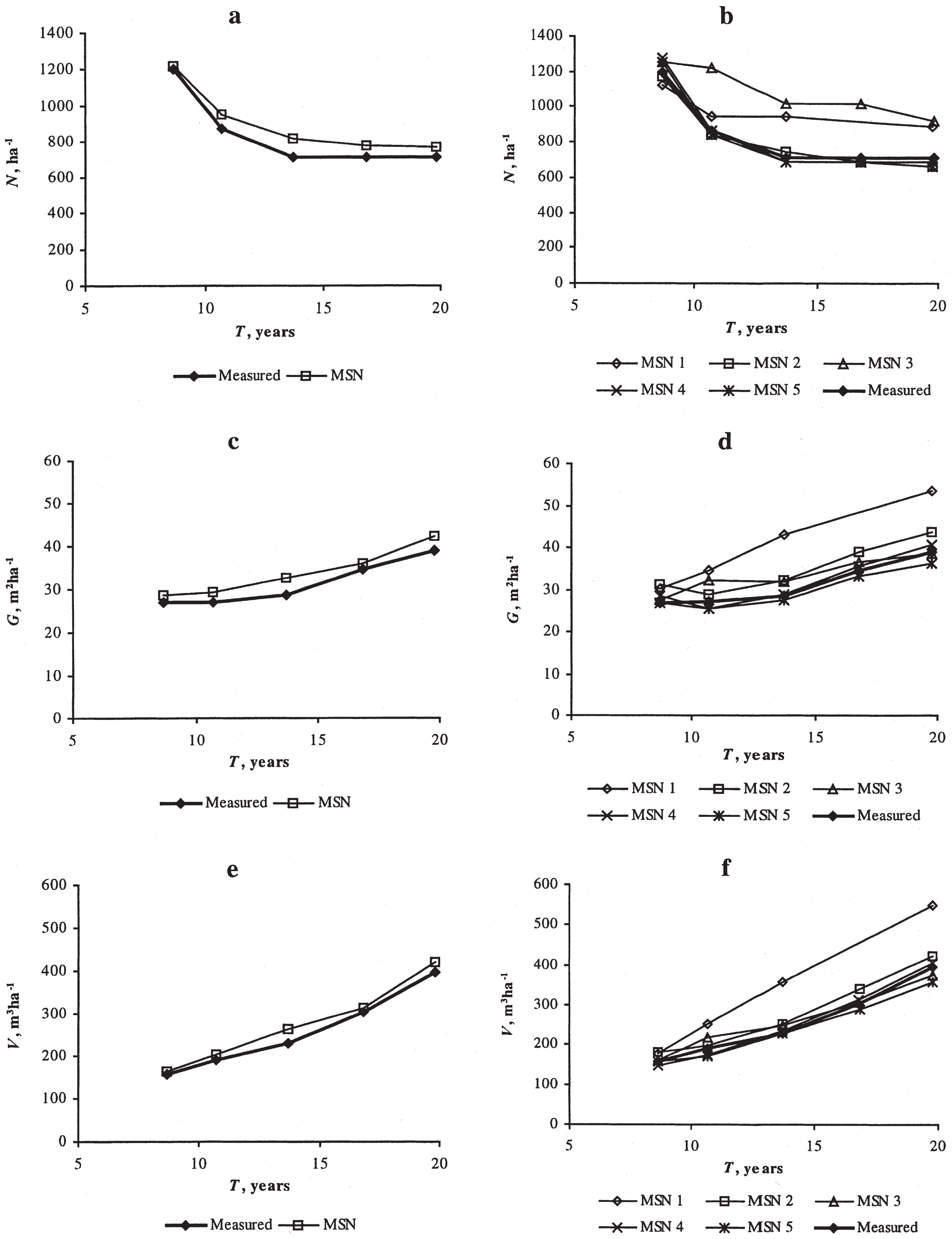

Fig. 3. Observed and predicted development of number of stems (a and b), stand basal area (c and d) and total stand volume (e and f) for a thinned sample plot. MSN 1–MSN 5 refer to the five different Most Similar Neighbour reference sample plots. MSN is the weighted average of the five reference sample plots chosen.

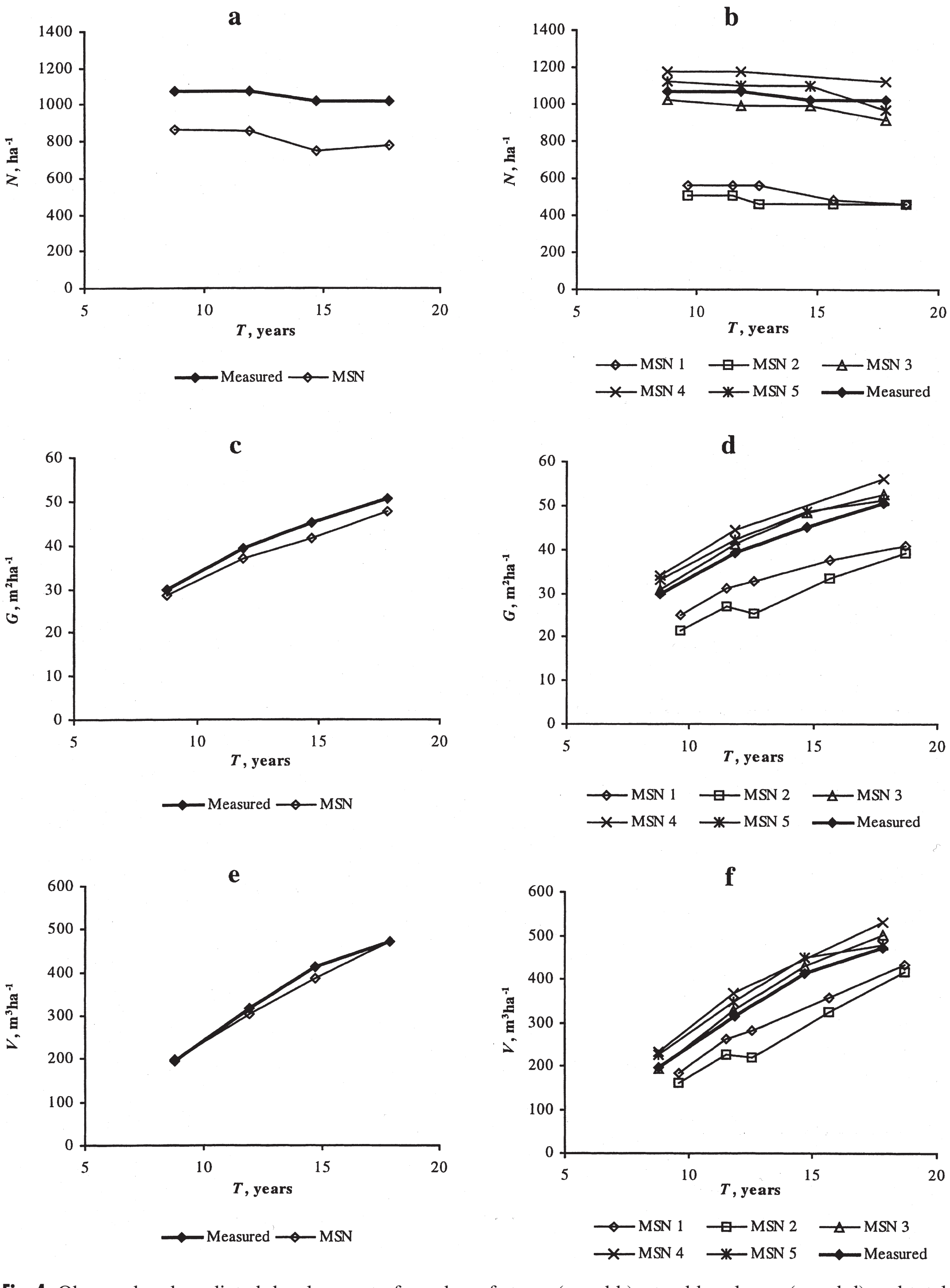

Fig. 4. Observed and predicted development of number of stems (a and b), stand basal area (c and d) and total stand volume (e and f) for an unthinned sample plot. MSN 1–MSN 5 refer to the five different Most Similar Neighbour reference sample plots. MSN is the weighted average of the five reference sample plots chosen.

Table 4. Three-Stage Least Square (3SLS) parameter estimates for Equations 6–10 of the simultaneous yield model.

Equation	Parameter	Estimate	Standard error	*t*-value
6	c_{11}	–0.12974	0.00691	–18.78
	c_{12}	2.08022	0.11320	18.38
7 and 8	c_{20}	–2.72277	0.39660	–6.87
	c_{21}	–25.46057	3.65330	–6.97
	c_{22}	0.39757	0.03980	9.99
	c_{23}	1.19967	0.06510	18.43
	c_{24}	3.73650	0.50870	7.35
9 and 10	c_{30}	0.09489	0.15070	0.63
	c_{31}	–0.06189	0.02260	–2.74
	c_{32}	0.73132	0.04030	18.15
	c_{33}	1.06106	0.03490	30.40

Table 5. Reliability figures of predicted stand volumes at the end of the growth period obtained for the simultaneous yield model (Equations 6–10). Biases and RMSEs are determined for the combined data and separately for each of the three thinning categories.

Prediction method	Combined data	Thinning group		
		0	1	2
n	142	25	34	83
MSN				
k		10	8	14
Bias, m^3	–1.8	5.5	–7.0	–1.8
Bias%	–0.5	1.2	–1.8	–0.7
RMSE, m^3	59.6	54.9	62.9	59.7
RMSE%	18.0	12.0	16.2	22.1
Simultaneous yield model				
Bias, m^3	–29.3	–40.9	–8.6	–34.3
Bias%	–8.8	–8.9	–2.2	–12.7
RMSE, m^3	82.1	100.5	75.4	78.5
RMSE$_{\%}$	24.8	21.9	19.5	29.0

period, it seemed to perform quite well in all of the thinning categories (Table 5). However, the RMSEs of predicted volumes were higher when obtained for the simultaneous yield model than for the MSN reference. The relative RMSE was 24.8% when determined for the whole sample plot data, and varied between 19.5–29.0% when calculated separately for the three thinning categories (Table 5). The highest absolute RMSE estimates were obtained for thinning group 0, and the highest relative values of bias for the sample plots of thinning group 2. However, the relative biases were less than the half of the corresponding RMSEs when they were analysed against all of the thinning groups and for the combined sample plot data.

4 Discussion

The aim of this study was to build a yield prediction model using the non-parametric MSN reference. The use of MSN reference offers some improvements when compared to k-nearest neighbour regression. The effect and amount of varying independent and dependent variables can be examined easily and quickly. More weight can be given to stand volume or to other stand characteristics, which may then improve the description of stand structure. Using the canonical correlation matrix of chosen variables in the calculations of final estimates guarantees that optimal solutions are found.

The simultaneous yield model (Equations 6–10, Table 4) for the prediction of stand dominant height, basal area and total stand volume was the same as presented by Pienaar and Harrison (1989). All the stands of the study were thinned at least once before the first measurement. On the other hand, the rather low initial planting densities made it possible to assume that the self-thinning effect only occurred in unthinned and mature stands. Due to this, no prediction models for the mortality were determined and the estimates of the number of stems at the end of growth period were obtained from MSN reference in the application of the yield model. This may affect the accuracy of volume predictions of the simultaneous yield model. However, estimates of the number of stems at the end of growth period were needed and no earlier developed suitable self-thinning models for *P. kesiya* were available.

The simultaneous yield model was estimated with the data that included only the growth intervals with no thinnings. Therefore, the volume predictions of the simultaneous yield model and the MSN reference are not fully comparable when obtained for stands where thinnings have occurred during considered growth periods. However, when using the same amount of stand characteristics information, MSN reference seemed to be more reliable growth predictor in all thinning categories. Simultaneous yield model produced overestimates in all thinning categories, whereas biases of MSN estimates where relatively smaller. Surprisingly, the differencies between these two models were largest in unthinned stands. This may be due to the fact that the variation of stand structure in unthinned stands is highest.

In more general, it is very simple to determine the total stem volumes with the system of Equations 6–10. However, if more information than the stand volume, basal area and dominant height is needed, more models must be constructed. The MSN method provides and contains all the measured information of neighbour stands until the end of analysis. In this study, the MSN method as a growth predictor was only analysed at the stand level. However, using a database of the diameter and height distributions of the Most Similar Neighbours it is possible to determine tree characteristics for different proportions of diameter distribution, size classes of trees or even single trees.

The presented non-parametric method offers good opportunities to build growth and yield series. The best accuracy obtained was about 18% in the RMSE of stand volume after a 10 year growth period. This result can be considered quite good. The presented method includes a realistic description of the development of tree stock both at tree and stand levels including also thinnings and mortality. The restrictions of the study material, i.e. the lack of the timing and rate of thinnings, had a strong effect on the application constructed. With more detailed and precisely collected material these restrictions could have been avoided and more realistic simulations obtained.

The reclassification of sample plots according to thinnings improved the results at the end of the growth period considerably. Classification helps to choose reference sample plots more logically. Because the study material did not include information about the timing and rate of thinnings, the timing of thinnings was unambiguously determined only in a few classes. If the study material had included more detailed information about thinnings and other possible treatments the problem of how to consider thinnings in simulations could have been taken into consideration more effectively.

The growth period used was about 10 years. The sample plots were measured rather irregularly and therefore it was not possible to achieve an exact temporal determination, i.e. the length, start and end, of the growth period. If there had been material available with the accurate dating of measurements and varying periods the data could have been reclassified according to the length of the growth period. Then it would also have been possible to construct growth and yield series for different durations of the growth periods. In conditions, where long and regularly measured time series exist, the presented MSN method has very promising and potential applicability.

When applying MSN reference the required minimum number of reference observations available should also be taken into consideration. No exact general figures can be presented, because the amount of reference observations is dependent e.g. on variation and locality of the data. In the study by Muinonen et al. (In press) about 60 compartments seemed to be enough in the case of interpretation of digitised aerial photographs. In our study, the amount of sample plots was 142 and it included different thinning categories. In minimum, the number of sample plots was only 25 in unthinned category. If there had been more observations, it would have been possible to construct completely separate models for each category, i.e. calculate canonical correlation and optimise the number of dependent and independent variables in each category.

In this study, the achieved improvement of yield predictions was two-fold when additional measurement of stand basal area at the beginning of growth period was examined. The stand basal area improved the prediction of stand volume considerably at the beginning of the growth period and slightly at the end of the growth period. These results are in line with the study of Maltamo and Mabvurira (1999) where static diameter distributions were predicted using varying information. However, no other calculations using stand basal area were done, because it is normally not included in measurements of forest inventory in Zambian conditions.

One feature of the applied method is that estimates can consist of exceptional observations. This can be seen in Figs. 3 and 4, where the variation in the number of stems was quite high. If the correspondent variable has not been taken into account as a dependent variable in the calculation of canonical correlation, there is no guarantee that all chosen reference sample plots are as accurate as possible in relation to this variable. However, the effect of different neighbours is mostly revealed in the final weighted average estimate. If there had been more characteristics available for use in the description of the stand structure, the calculation of canonical correlation and choice of reference sample plots would have been more reliable. However, then the stand volume predictions achieved would probably have been more inaccurate.

When the forest manager has access to a database of previously measured permanent sample plots from the area in question including different thinning regimes, it is recommendable to apply the MSN method presented for the construction of the local yield prediction model. In addition, the current forest inventory has normally been carried out, and some mean stand characteristics have been measured from stands with varying ages and growing conditions. Using this information, the forest manager can analyse different rotation ages and the effects of different thinning regimes, aiming at the development of stand management practices. The target of all forest planning is to find as optimal a management programme for the given stands of a certain forest area as possible. With the method presented it is possible to base the planning and management of forest plantations effectively on the observed growth figures and existing yield records.

Acknowledgements

The authors wish to thank Professor Timo Pukkala and Dr Sauli Valkonen for their valuable comments on the manuscript. The data collection in Africa was financed by the Society of Forestry in Finland, and the University of Joensuu, Faculty of Forestry. Dr Jussi Saramäki from the FORST-project and Mr Progress M. Sekeli from the Forestry Research of Zambia made it possible to carry out the field assessments and data collection. The study is a part of the INCO DC co-operation project *Tree Seedling Production and Management of Plantation Forest* between the forest researchers of Southeastern African countries and the European Union member countries. We also wish to thank Dr Lisa Lena Opas-Hänninen for revising the English.

References

Altman, N.S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46: 175−185.

Armitage, F.B. & Burley, J. 1980. Pinus kesiya Royle ex Gordon. Unit of Tropical Silviculture, Commonwealth Forestry Institute, University of Oxford, Tropical Forestry Papers 9. 199 p.

Baskerville, G.L. 1972. Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forest Research 2: 49–53.

Borders, B.E. 1989. Systems of equations in forest stand modeling. Forest Science 35: 548–556.

— & 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.

— & Patterson, W.D. 1990. Projecting stand tables: a comparison of the Weibull diameter distribution method, a percentile-based projection method, and a basal area growth projection method. Forest Science 36: 413–424.

Clutter, J.L., Fortson, J.C., Pienaar, L.V., Brister, G.H. & Bailey, R.L. 1983. Timber management: a quantitative approach. John Wiley & Sons, New York. 333 p.

Droessler, T.D. & Burk, T.E.. 1989. A test of nonparametric smoothing of diameter distributions. Scandinavian Journal of Forest Research 4: 407–415.

Eerikäinen, K. 1999. Random parameter model for the relationship between stand age and tree height in Zambia. 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: 153–165.

— 2001. Stem volume models with random coefficients for Pinus kesiya in Tanzania, Zambia, and Zimbabwe. Canadian Journal of Forest Research 31: 879–888.

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, Cambridge. 323 p.

Holm, S., Hägglund, B. & Mårtensson, A. 1979. A method for generalization of sample tree data from the Swedish National Forest Survey. Swedish University of Agricultural Sciences, Department of Forest Survey, Report No. 25. 94 p. (In Swedish with English summary).

Hynynen, J. 1995. Predicting the growth response to thinning for Scots pine stands using individual-tree growth models. Silva Fennica 29: 225–246.

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

Kilkki, P. & Päivinen, R. 1987. Reference sample plots to compare field measurements and satellite data in forest inventory. Proceedings of Seminar on Remote Sensing-aided Forest Inventory, organised by SNS, Hyytiälä, Finland, 10–12 December 1986. University of Helsinki, Department of Forest Mensuration and Management. p. 209–215.

Kolström, T. 1992. Dynamics of uneven-aged stands of Norway spruce – a model approach. Ph.D. thesis summary. Metsäntutkimuslaitoksen tiedonantoja – Finnish Forest Research Institute, Research Papers 411. 29 p.

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

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

— & Mabvurira, D. 1999. Prediction of diameter distribution of Eucalyptus grandis in Zimbabwe using varying information. In: Pukkala, T. & Eerikäinen, K. (eds.). Growth and yield modelling of tree plantations in South and East Africa. University of Joensuu, Faculty of Forestry, Research Notes 97: 97–111.

— & Uuttera, J. 1998. The angle-count sampling in description of forest stand structure. Forest and Landscape Research 1: 448–471.

Mbuya, L.P., Msanga, H.P., Ruffo, C.K., Birnie, A. & Tengnäs, B. 1994. Useful trees and shrubs for Tanzania. Identification, propagation and management for agricultural and pastoral communities. Regional Soil Conservation Unit (RSCU), Swedish International Development Authority (SIDA), Embassy of Sweden. English Press, Nairobi, Kenya. 539 p.

Miina, J., Maltamo, M. & Eerikäinen, K. 1999. Modelling the growth and yield of Pinus kesiya in Chati, Zambia. 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: 167–181.

Moeur, M. & Riemann Hershey, 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. p. 419–430.

— & Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science 41: 337–359.

Muinonen, E., Maltamo, M., Hyppänen, H. & Vainikainen, V. In press. Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure. To appear in Remote Sensing of Environment.

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.

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 diameter growth in even-aged Scots pine stands with a spatial and non-spatial model. Silva Fennica 23: 101–116.

Saramäki, J. 1992. A growth and yield prediction model of Pinus kesiya (Royle ex Gordon) in

Zambia. Acta Forestalia Fennica 230. 68 p.

SAS Institute Inc. 1993. SAS/ETS user's guide, version 6, second edition. SAS Institute Inc., Cary, NC. 1022 p.

Sekeli, P.M. & Phiri, M. 1999. Comparative growth rates of Philippine and Vietnam provenances of Pinus kesiya in the Copperbelt province of Zambia. In: Pukkala, T. & Eerikäinen, K. (eds.). Growth and yield modelling of tree plantations in South and East Africa. Proceedings of the meeting in Mombasa, Kenya 12–15 October, 1999. The University of Joensuu, Faculty of Forestry, Research Notes 97: 43–53.

Systems of measurements in commercial compartments. A system of continuous inventory for Pinus species established by Industrial Plantations. 1969. Zambia Forest Department, Division of Forest Research. Control plan 5/3/1. 5 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.

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

Zellner, A. & Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30: 54–78.

Total of 36 references