Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland

M. MALTAMO1*, J. MALINEN1, A. KANGAS2, S. HÄRKÖNEN1 AND A.-M. PASANEN1

1University of Joensuu, Faculty of Forestry, PO Box 111, FIN-80101 Joensuu, Finland
2University of Helsinki, Department of Forest Resources Management, PO Box 27, FIN-00014 University of Helsinki, Finland
* Corresponding author. E-mail: matti.maltamo@forest.joensuu.fi

Summary
Non-parametric regression was used to predict basal-area diameter distribution and the stand volume of Scots pine (Pinus sylvestris L.) in Finland. The regression is based on weighted averages of most similar neighbours (MSN) of a stand. This is a special case of the $k$-nearest-neighbour method, in which the similarity is measured using canonical correlations. The results were compared with those obtained with percentile-based basal-area diameter distribution models which are currently used in Finland. When constructing the MSN models, stand mean characteristics were used as independent variables, and variables describing the shape of the diameter distribution and stand density as dependent variables. Since the relationships are mainly non-linear, and the weights are based on linear correlations, the use of second powers of independent variables improved the results. The basic MSN model was constructed for the whole study area, but also regional and local models were tested. Regional models were tested for the vegetation zones in Finland. Local models were tested by applying two modified MSN regression methods for similarity-based neighbourhoods. The results obtained indicated that the accuracy of the MSN method is comparable to percentile-based basal-area diameter distribution models in the case of stand volume. However, the description of stand structure, i.e. the number of stems could be improved by using MSN regression. The vegetation zones used seemed to be too small for reference areas. However, it was possible to get more accurate local results if the neighbourhood used was described and selected effectively.

Introduction
In Finland forest inventory is carried out on two levels. National forest inventory is based on systematic cluster sampling and covers the whole country (Tomppo, 1993). These data are used for national level planning. Forest management planning in private forests is usually based on
information collected by forest compartments (stands) (Poso, 1983). According to this method, a few stand characteristics, usually basal area, basal-area median diameter and mean height, are visually assessed with the aid of a few measurements in the field. The variables of primary interest, namely stand volume and timber assortments, are estimated by, first, predicting the diameter distribution in a stand and, then, by applying this information to height and taper curve functions. This procedure includes expensive field work, but the number of assessments in any one stand is so small that the precision of estimated stand variables is low.

Research has aimed recently at improving the results of inventory by compartments by concentrating on developing and testing different alternatives to predict the diameter distribution. The traditional approach is to predict the parameters of Weibull function according to stand variables (Kilkki and Päivinen, 1986). The basal-area diameter distributions used are scaled to the observed stand basal area. This produces accurate estimates for stand total volume, if the basal area is error-free. An estimate of the stem number can be calculated from the distribution for each diameter class (Kilkki and Päivinen, 1986). An alternative to this is to solve the stem number distribution analytically from the basal-area diameter distribution as presented by Gove and Patil (1998).

However, when using basal-area diameter distributions, typically a poor estimate for stem number is obtained. This is due to the fact that the parameter estimates of the model are optimized only with respect to basal-area distribution, not with stem number distribution. According to previous studies, the root mean square error (RMSE) of the stem number estimate can be as high as 30–40 per cent (Siipilehto, 1999; Kangas and Maltamo, 2000b). Obtaining such errors when using basal-area diameter distributions, clearly indicates that (horizontal) stand structure is not described properly and the simulation of future stand development is not based on realistic tree stocks. In addition to diameter distribution, it is also important to describe also the vertical structure, i.e. variation in height distribution.

One possibility to improve the stand structure description is to use the very flexible percentiles method, originally presented by Borders et al. (1987). Maltamo et al. (2000) and Kangas and Maltamo (2000a) used this method to describe diameter distribution of both heterogeneous and managed stands. In the percentile method, the diameters at predefined percentiles of the distribution function are predicted with models (Borders et al., 1987). By interpolating between the predicted diameters, a cumulative basal-area diameter distribution function (CDF) is obtained. This percentiles method has become recently a main alternative to calculate results of inventory by compartments in Finland (Maltamo et al., 2002).

Another possibility is to use the non-parametric methods, which predict the value of the variable, in question, as a weighted average of the values of nearest neighbouring observations. The neighbours are defined with some similarity measure in the predicting variables (e.g. Härdle, 1989; Altman, 1992). The chosen neighbours are selected from a database of previously measured observations, reference data.

Non-parametric methods have been utilized in several forestry applications. These include, the generalization of sample tree information, different growth and yield models and applications of multisource and multivariate forest inventories (Kilkki and Päivinen, 1987; Moeur and Stage, 1995; Korhonen and Kangas, 1997; Moeur and Hershey, 1999; Holmström et al., 2001; Lemay and Temesgen, 2001; Sironen et al., 2003). The prediction of diameter distribution has been presented with different non-parametric methods in the study by Maltamo and Kangas (1998).

The most frequently applied non-parametric method is the k-nearest neighbour method (e.g. Kilkki and Päivinen, 1987). When applying the k-nearest neighbour method, the form of a similarity (or distance) measure must be specified to define the neighbourhood of a given point. Typically, similarity measures based on absolute differences, Euclidean or Mahalanobis distance functions have been used. One important special case is the so-called Most Similar Neighbour (MSN) method, in which the similarity is based on canonical correlations and Mahalanobis distance (Moeur and Stage, 1995). The benefit of the MSN method is that the similarity measure can be solved analytically.
Non-parametric methods can easily describe local variability. In the study by Sironen et al. (2003), local non-parametric regression models for single tree growth were constructed, and promising results were obtained when compared with national parametric growth models. The successful use of non-parametric methods demands knowledge of the correlations of the used variables within the independent variables and dependent variables as well as between these variable groups. For example, assuming linear correlation when the true correlation is nonlinear, might give unnecessarily poor results. Furthermore, different dependent variables are weighted according to inherent correlations, not necessarily according to their importance for the user of the predictions.

At the boundary of the predictor space, the neighbourhood is asymmetric and non-parametric models tend to be highly biased. Bias can be a problem in the interior, as well, if the predictors are non-uniformly distributed or if the regression function has substantial curvature. By increasing the number of neighbours we can enhance the validity of a model, but at the same time the bias of the estimates increases at the boundary. One option to enhance the performance in these situations is to adapt neighbourhood or neighbour selection locally according to predictor space. Malinen (2003) proposed two methods for local adaptation: local MSN method, which re-calculates a weighting matrix for locally selected neighbourhood and a Locally Adaptable Neighbourhood (LAN) MSN method, which selects a locally symmetric combination of neighbours according to predictor space.

The aim of this study was to develop a stand variable prediction model for Scots pine (Pinus sylvestris L.), in Finland, using MSN regression. The main emphasis was on the prediction of basal-area diameter distribution, which also provides a realistic description of stand structure. A second objective was to investigate local variation by using separate models for different vegetation zones in Finland. Finally, two different methods to modify the MSN method, local MSN and LAN MSN, were also tested. The results obtained with the MSN regression were compared with the percentiles method.

Materials and methods

Study material and data preparation

The data set includes the permanent sample plots measured by the Finnish Forest Research Institute (FFRI), originally for growth modelling purposes (Gustavsen et al., 1988). The sample plots were established on mineral soils across Finland. The data include clusters of three circular plots located systematically within a stand, avoiding stand edges. When testing the basal-area diameter distribution prediction methods, the information of these circular plots were combined. Altogether 100–120 trees were measured in each stand for diameter at breast height to the nearest 0.1 cm. Correspondingly, tree height was measured from ~30 sample trees to the nearest 0.1 m. Only the plots with Scots pine were used.

When considering the local variability of the estimates, the data set was further divided into sub-boreal vegetation zones (Kalliola, 1973) (Figure 1). The most southern part of Finland (1) belongs to the hemi-boreal vegetation zone. However, there were only eight sample plots in this area. The south-boreal vegetation zone was further divided into south-western (2), Lake-Finland (3) and south Ostrobothnia (4) parts which include 32, 87 and 7 sample plots, respectively. The middle-boreal vegetation zone was divided into western Ostrobothnia (5) (141 sample plots) and eastern Kainuu (6) (73 sample plots). The north-boreal zone was further divided into southern Lapland (7) (110 sample plots) and forest Lapland (8) (18 sample plots).

Because there were so few observations in some of the vegetation zones they were combined into larger ones consisting of three main groups which are: southern Finland (areas 1–4), Ostrobothnia (area 5) and northern Finland (areas 6–8).

The tree height model of Näslund (1937) was constructed separately for each stand using sample tree measurements. The height of each tally tree was then predicted with these models. A random component was added to the predicted heights from a normal distribution using the estimated standard deviation of each height model. Total, sawlog and pulpwood volumes were calculated for each tree using taper curve functions
Figure 1. The vegetation zones in Finland: (1) hemi-boreal, (2) south-western, (3) Lake-Finland, (4) south Ostrobothnia, (5) Ostrobothnia, (6) Kainuu, (7) southern Lapland and (8) forest Lapland.
presented by Laasasenaho (1982). In young stands (age below 40 years), all trees were taken into account. In older stands, only trees larger than 5 cm in diameter were considered. Basal-area diameter distributions were formed by using basal areas of individual trees. Finally, stand characteristics were calculated as averages and sums of tallied trees (Table 1).

**Methods**

In the MSN method, \( k \)-nearest-neighbours (or \( k \)-most similar) sample plots are used for predicting stand variables. The stand variables obtained include basal-area diameter distribution and also tree heights in certain cases. The most similar neighbours to the target sample plot \( u \) are chosen from the reference sample plots. 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 \( (k) \) was allowed to vary from 1 to 15 in the calculations.

The MSN regression is based on canonical correlation between independent and dependent variables. In the MSN method, the most similar neighbour to the observation \( u \) in the target data is that observation in the reference data, for which Mahalanobis distance function (Mahalanobis, 1936) \( (Y_u - Y_j)^TW(Y_u - Y_j) \) is minimized over all \( j = 1, \ldots, n \) reference trees, where \( Y_u \) is a row vector of the unknown variables in the target data, \( Y_j \) is a row vector of the observed variables in the reference data and \( W \) is a weighting matrix. 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 \( (U_r, V_r) \) are formed from the set of dependent and independent variables, in such a way that the correlation between them is maximized

\[
U_r = \alpha_r Y \quad \text{and} \quad V_r = \gamma_r X, \tag{1}
\]

where \( \alpha_r \) are canonical coefficients of the dependent variables \( (r = 1, \ldots, s) \) and \( \gamma_r \) are canonical coefficients of the independent variables \( (r = 1, \ldots, s) \). There are \( s \) possible pairs of canonical variates \( (U_r, V_r) \) as the result of the analysis, where \( s \) is either the number of dependent or independent variables, depending on which is smaller (Moeur and Stage, 1995).

The final distance function is

\[
D_{ij}^2 = (X_u - X_j)^T \Gamma \Lambda^2 \Gamma^T (X_u - X_j)', \tag{2}
\]

where \( X_u \) is the vector of independent variables of the target sample plot \( u \), \( X_j \) is the vector of independent variables of reference sample plots \( j \), \( \Gamma \) is the matrix of canonical coefficients of the independent variables, and \( \Lambda^2 \) is the diagonal matrix of squared canonical correlations.

The final estimate vector \( (z_u) \) for the basal-area diameter distribution (and height information) of sample plot \( u \) was calculated as the weighted average of the basal-area diameter distribution and height information vectors of reference sample plots \( (z_j) \):

\[
z_u = \frac{\sum_{u=1}^{k} \frac{1}{D_{uj}} z_j}{\sum_{u=1}^{k} \frac{1}{D_{uj}}}, \tag{3}
\]

The independent and dependent stand variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A ) (years)</td>
<td>16</td>
<td>183</td>
<td>66.8</td>
<td>31.3</td>
</tr>
<tr>
<td>( d_{M}(cm) )</td>
<td>5.8</td>
<td>35.0</td>
<td>15.6</td>
<td>6.0</td>
</tr>
<tr>
<td>( h_{M}(cm) )</td>
<td>4.1</td>
<td>32.5</td>
<td>11.8</td>
<td>5.1</td>
</tr>
<tr>
<td>( N (ha^{-1}) )</td>
<td>146</td>
<td>3684</td>
<td>1121.2</td>
<td>679.4</td>
</tr>
<tr>
<td>( G (m^2 ha^{-1}) )</td>
<td>1.5</td>
<td>32.6</td>
<td>12.7</td>
<td>6.4</td>
</tr>
<tr>
<td>( V (m^3 ha^{-1}) )</td>
<td>4.4</td>
<td>313.1</td>
<td>84.5</td>
<td>62.7</td>
</tr>
<tr>
<td>( V_s (cm) )</td>
<td>0</td>
<td>250.0</td>
<td>39.7</td>
<td>50.9</td>
</tr>
</tbody>
</table>

\( A \) = age; \( d_{M} \) = basal area mean diameter; \( h_{M} \) = height of basal area median tree; \( N \) = number of stems; \( G \) = basal area; \( V \) = volume; \( V_s \) = sawlog volume.
used in the calculation of canonical correlations are presented in Table 2. Different combinations and powers of these variables were also tested.

Different ways to generate tree height information were also tested. The basic method included the MSN estimate only for basal-area diameter distribution and the separate parametric height model by Siipilehto (1999). In addition, the MSN method was used to simultaneously predict both basal-area diameter distribution and tree heights. Another alternative was to use the MSN method for prediction of basal-area diameter distribution and both parameters of Näslund’s (1937) height curve. Finally, the MSN method was used for prediction of basal-area diameter distribution and the shape parameter of Näslund’s (1937) height curve. In the last case, the other parameter, which adjusts the level of height predictions, was estimated using the approach presented by Siipilehto (1999).

The MSN method can be localized in many ways. The locality can be defined either according to geographic closeness (regional MSN) or similarity in stand characteristics (local MSN). Also a combination of these two approaches may be used. The regional results by vegetation zones were calculated, first, by using common MSN models. Secondly, separate MSN models for each region were constructed. Then, a common weighting matrix was calculated, but neighbours were selected only from corresponding vegetation zones and, finally, constructed weighting matrices were regional but neighbours were selected from the whole study material.

In the local MSN method, the reference data are reduced to consist of only locally important observations by using the MSN method to select the local neighbourhood (Malinen, 2003). The size of the local neighbourhood can be set according to the number of reference data observations. This local neighbourhood is then used to calculate a new local weighting matrix, which concentrates on local correlations between independent and dependent variables. Similar ideas are employed in the local regression method (e.g. Cleveland, 1979). A new feature, when compared with Malinen (2003), was that the local neighbourhood was selected iteratively to assure that the local neighbourhood would be elongated according to local reference data instead of global reference data. In the first iteration, the global weighting matrix was utilized. In the next steps, the estimated local correlations were utilized in selecting the local neighbourhood. Altogether 20 iterations were used. At the last stage, the MSN analysis was performed by using this local weighting matrix and local reference data.

In the other modification, the LAN MSN method, every possible combination that can be formed from the \( k \)-nearest neighbours is examined, in turn, and averages of predictor variables of neighbour combinations are calculated (Malinen, 2003). Every vector of average predictor variables is compared with the vector of predictor variables of the target stand with MSN metrics and the combination of neighbours whose average is most identical to the target stand is chosen to be used in calculation of estimates. The different independent variables are weighted according to the distance formula used (2).

This method differs from the basic \( k \)-nearest-neighbour methodology in the sense that the number of neighbours may vary. Therefore, it could be interpreted as a variable-\( k \)-nearest-neighbour method. The number of neighbours,
from which the optimal combination is selected, is limited to 15 neighbours due to computer time consumption. When the number of the neighbours \( k \) is expanded, the demand for computer time increases by \( 2^k \).

As a comparison for different MSN methods, the stand characteristics were predicted by using parametric methods. First, the basal-area diameter distribution was predicted using the percentile method and, second, the tree heights, volumes and timber assortments in each diameter class were predicted by using height and taper curve models similarly as in the case of MSN models.

For predicting the basal-area diameter distribution, 12 diameters at different percentiles of the cumulative basal area distribution function (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95 and 100 per cent) were predicted with models. The CDF function was obtained by interpolating between the percentiles. Interpolation was carried out using Späth’s rational spline (see Lether, 1984), in order to obtain a non-decreasing function (i.e. a proper CDF) (Kangas and Maltamo, 2000a; Maltamo et al., 2000). The models used for predicting the percentiles were those presented by Kangas and Maltamo (2000a). The regressors were the stand basal area, stand age and the basal-area median diameter.

The test criteria used in the comparison of different model forms and number of neighbours were the root mean square error (RMSE) and bias of predicted stand characteristics. The considered stand characteristics were stand total and sawlog volume, and number of stems.

For example, the RMSE of predicted stand volumes was:

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

(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. Correspondingly, the bias of the predicted volumes was calculated with the formula:

\[
Bias = \frac{\sum_{i=1}^{n}(V_i - \hat{V}_i)^2}{n}
\]

(5)

The relative RMSE (Bias) of the volume estimate were calculated by dividing the absolute RMSE (Bias) by the true mean volume \( \bar{V} \) of the stands.

**Results**

Different model forms were compared including a varying number of independent and dependent variables and number of neighbours. The main results considering the common MSN models, for the whole study material, are presented in Table 3. When different MSN models are compared, it can be said that if independent variables are used as such (i.e. without non-linear transformations) the results are less accurate (Table 3, MSN model 1). When the second powers of stand variables are added as independent variables, the accuracy is improved both in terms of RMSE and bias (Table 3, MSN model 2). At its lowest, the RMSE of the number of stems is 18.39 per cent. It is also notable, according to the results of this study, that volume characteristics should not be used as independent variables (Table 3, MSN models 1 and 2). It is better to use only variables which describe the form of the diameter distribution, i.e. diameter percentiles and variables indicating stand density. This is contrary to earlier studies (e.g. Malinen et al., 2001). For the above-mentioned reasons, MSN model 2 (Table 3) was chosen to be the basic MSN model to which regional, local and LAN MSN models are based on and with which the percentiles models are compared.

The corresponding results for the whole study material, using the percentile method are in Table 4. The accuracy of MSN (Table 3, MSN model 2) is comparable to the percentile method. Although sawlog volume estimates of percentile models are slightly better in most of the cases, the MSN estimates of the number of stems are considerably better. The biases are also smaller for MSN estimates in the case of all stand characteristics.

Different methods to estimate tree heights were examined using the chosen basic MSN model (Table 3, MSN model 2). The results indicated that the choice of height model had a relatively small effect on the accuracy of volume characteristics (Table 5). The basic model, i.e. the separate parametric height model (Table 3, MSN model 2), however, was the most accurate alternative.
Table 3: The results of the basic MSN models for the whole study material

<table>
<thead>
<tr>
<th>MSN model no.</th>
<th>Independent variables</th>
<th>Dependent variables</th>
<th>k</th>
<th>RMSE of V (%)</th>
<th>RMSE of N (%)</th>
<th>RMSE of Vs (%)</th>
<th>Bias of V (%)</th>
<th>Bias of N (%)</th>
<th>Bias of Vs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$d_\text{gM}, h_{\text{gM}}, T, G, X, Y$</td>
<td>$V, V_s, N/G, N$</td>
<td>4</td>
<td>9.15</td>
<td>23.56</td>
<td>20.72</td>
<td>0.36</td>
<td>-0.47</td>
<td>2.70</td>
</tr>
<tr>
<td>2</td>
<td>$d_{\text{gM}}, d_{\text{gM}}^2, h_{\text{gM}}, h_{\text{gM}}^2, T, T^2, G, G^2, X, Y$</td>
<td>$N/G, (N/G)^2$, diameter percentiles 0, 20, 40, 60, 80, 100%</td>
<td>10</td>
<td>8.88</td>
<td>18.39</td>
<td>17.68</td>
<td>0.25</td>
<td>-0.28</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Used independent and dependent variables and number of neighbours ($k$) are presented as well as accuracy (RMSE and bias) of the predicted stand characteristics (total volume, sawlog volume, number of stems).

$T$ = stand age, $X$ and $Y$ are stand coordinates; for other abbreviations of variables, see Table 1.
The regional variation was examined by using different ways to estimate the weighting matrix and choose the neighbours (Table 6). In the first phase, regional results were calculated by using the basic model (Table 3, MSN model 2). The second phase consisted of constructing separate models for each region (Table 6, models 3–5). In the next alternative, a common weighting matrix was constructed but, when calculating results, neighbours were selected from the corresponding vegetation zone (Table 6, model 6). Finally, separate weighting matrices were produced for each region, but neighbours were selected from the whole study area (Table 6, models 7–9).

When constructing these models, different model forms were examined but, in most cases, the form of the basic model (Table 3, MSN model 2) was the most accurate. Only the optimal number of nearest neighbours used varied (Table 6).

The regional results indicate that local variation could not be described more accurately, on average, using regional models. Neither were there any considerable differences in bias. The results, considering stand volume, were slightly better when choosing the neighbours from the corresponding vegetation zone (Table 6, MSN model 6) than when using the basic model (Table 3, MSN model 2). Correspondingly, the regional calculation of the weighting matrix produced better results when considering the number of stems (Table 6, models 7–9).

In general, most accurate results for volume characteristics were obtained in southern Finland, whereas the number of stems was predicted most accurately in Ostrobothnia and northern Finland using different MSN methods. In contrast, in the case of the percentiles model, the number of stems was predicted most accurately in southern Finland (Table 4).

Table 4: The accuracy (RMSE and bias) of the predicted stand characteristics (total volume, sawlog volume, number of stems) using percentile-based models (both results for the whole study area and for the different vegetation zones are presented).

<table>
<thead>
<tr>
<th>Percentile models</th>
<th>RMSE of V (%)</th>
<th>RMSE of N (%)</th>
<th>RMSE of Vs (%)</th>
<th>Bias of V (%)</th>
<th>Bias of N (%)</th>
<th>Bias of Vs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole study area</td>
<td>8.71</td>
<td>26.77</td>
<td>16.14</td>
<td>0.38</td>
<td>-7.32</td>
<td>1.79</td>
</tr>
<tr>
<td>Southern Finland</td>
<td>7.75</td>
<td>18.18</td>
<td>12.20</td>
<td>0.80</td>
<td>-2.95</td>
<td>2.77</td>
</tr>
<tr>
<td>Ostrobothnia</td>
<td>8.85</td>
<td>32.88</td>
<td>21.41</td>
<td>0.84</td>
<td>-13.27</td>
<td>1.33</td>
</tr>
<tr>
<td>Northern Finland</td>
<td>9.73</td>
<td>23.97</td>
<td>17.64</td>
<td>-1.40</td>
<td>-5.64</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 5: The accuracy (RMSE and bias) of the predicted stand characteristics (total volume, sawlog volume, number of stems) using different tree height prediction methods.

<table>
<thead>
<tr>
<th>MSN model no.</th>
<th>Height model</th>
<th>RMSE of V (%)</th>
<th>RMSE of N (%)</th>
<th>RMSE of Vs (%)</th>
<th>Bias of V (%)</th>
<th>Bias of N (%)</th>
<th>Bias of Vs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Tree heights using MSN</td>
<td>10</td>
<td>8.93</td>
<td>18.39</td>
<td>18.26</td>
<td>0.97</td>
<td>-0.28</td>
</tr>
<tr>
<td>2</td>
<td>Parameters of height curve using MSN</td>
<td>10</td>
<td>9.05</td>
<td>18.39</td>
<td>18.22</td>
<td>1.48</td>
<td>-0.28</td>
</tr>
<tr>
<td>2</td>
<td>Shape parameter of height curve using MSN</td>
<td>10</td>
<td>9.08</td>
<td>18.39</td>
<td>17.39</td>
<td>0.14</td>
<td>-0.28</td>
</tr>
</tbody>
</table>
Table 6: The results of the different constructed MSN models for the vegetation zones used as well as for the whole study area

<table>
<thead>
<tr>
<th>MSN model no.</th>
<th>Vegetation area</th>
<th>Independent variables</th>
<th>Dependent variables</th>
<th>RMSE of N (%)</th>
<th>Bias of N (%)</th>
<th>RMSE of V (%)</th>
<th>Bias of V (%)</th>
<th>RMSE of $V_s$ (%)</th>
<th>Bias of $V_s$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Southern Finland</td>
<td>$d_{GM}, d_{GM^2}, h_{GM}, h_{GM^2}, N/G, (N/G)^2$, diameter percentiles</td>
<td>0, 20, 40, 60, 80, 100%</td>
<td>7.67</td>
<td>19.30</td>
<td>14.83</td>
<td>0.62</td>
<td>-2.98</td>
<td>1.95</td>
</tr>
<tr>
<td>2</td>
<td>Ostrobothnia</td>
<td>As above</td>
<td>As above</td>
<td>10.89</td>
<td>17.66</td>
<td>21.51</td>
<td>0.51</td>
<td>1.07</td>
<td>1.51</td>
</tr>
<tr>
<td>2</td>
<td>Northern Finland</td>
<td>As above</td>
<td>As above</td>
<td>10</td>
<td>-9.86</td>
<td>17.31</td>
<td>17.65</td>
<td>-0.33</td>
<td>0.87</td>
</tr>
<tr>
<td>3, 4, 5</td>
<td>Whole area</td>
<td>As above</td>
<td>As above</td>
<td>7.66</td>
<td>19.82</td>
<td>18.46</td>
<td>0.62</td>
<td>-1.28</td>
<td>2.38</td>
</tr>
<tr>
<td>6</td>
<td>Southern Finland</td>
<td>$d_{GM}, d_{GM^2}, h_{GM}, h_{GM^2}, N/G, (N/G)^2$, diameter percentiles</td>
<td>0, 20, 40, 60, 80, 100%</td>
<td>8.93</td>
<td>19.03</td>
<td>20.91</td>
<td>0.29</td>
<td>0.37</td>
<td>2.24</td>
</tr>
<tr>
<td>6</td>
<td>Ostrobothnia</td>
<td>As above</td>
<td>As above</td>
<td>3</td>
<td>8.87</td>
<td>18.00</td>
<td>23.81</td>
<td>0.59</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>Northern Finland</td>
<td>$d_{GM}, d_{GM^2}, h_{GM}, h_{GM^2}, N/G, (N/G)^2$, diameter percentiles</td>
<td>0, 20, 40, 60, 80, 100%</td>
<td>10.07</td>
<td>18.48</td>
<td>19.81</td>
<td>-0.28</td>
<td>1.33</td>
<td>1.67</td>
</tr>
<tr>
<td>6</td>
<td>Whole area</td>
<td>As above</td>
<td>As above</td>
<td>8.93</td>
<td>19.03</td>
<td>20.91</td>
<td>0.29</td>
<td>0.37</td>
<td>2.24</td>
</tr>
<tr>
<td>7</td>
<td>Southern Finland</td>
<td>As above</td>
<td>As above</td>
<td>7.59</td>
<td>19.78</td>
<td>15.40</td>
<td>0.48</td>
<td>-0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>Ostrobothnia</td>
<td>As above</td>
<td>As above</td>
<td>6</td>
<td>8.71</td>
<td>17.21</td>
<td>21.14</td>
<td>0.69</td>
<td>-0.56</td>
</tr>
<tr>
<td>9</td>
<td>Northern Finland</td>
<td>As above</td>
<td>As above</td>
<td>6</td>
<td>9.84</td>
<td>18.20</td>
<td>19.34</td>
<td>-0.23</td>
<td>1.09</td>
</tr>
<tr>
<td>7, 8, 9</td>
<td>Whole area</td>
<td>As above</td>
<td>As above</td>
<td>8.79</td>
<td>18.67</td>
<td>18.49</td>
<td>0.28</td>
<td>-0.11</td>
<td>1.71</td>
</tr>
</tbody>
</table>

The accuracy (RMSE and bias) of the predicted stand characteristics (total volume, sawlog volume, number of stems) is presented. Explanations of different MSN models are presented in the text.
Figure 2. The results of the local MSN calculations. Part (a) describes the accuracy of stand total volume as a function of the size of the neighbourhood, i.e. the number of the reference sample plots. Correspondingly, parts (b) and (c) describe estimates of number of stems and sawlog volume.
15.68 per cent when the size of the neighbourhood used for calculating the weighting matrix was only 30 observations. However, the results, concerning a neighbourhood under 100 observations, are not unambiguously reliable due to the problems in accuracy when calculating canonical correlations for few data. Biases were quite stable in the case of volume characteristics but there was a corresponding decreasing trend in estimates of the number of stems.

Finally, the results, calculated by using the LAN MSN method, indicated that the accuracy of stand characteristics was not improved compared with the basic model (Table 3, MSN model 2). The relative RMSEs of volume, number of stems and sawlog volume obtained were 8.93, 20.15 and 17.00 per cent, respectively. The corresponding biases were 0.23, –1.08 and 1.23 per cent, respectively.

Discussion
The aim of this study was to estimate non-parametric MSN models to predict stand variables and compare the accuracy of prediction to percentiles models, presented by Kangas and Maltamo (2000a). The results indicated that, especially, the description of stand structure could be improved by using the MSN regression. It is also possible to get more accurate local results if the neighbourhood used for calculating the weighting matrix is described and selected optimally.

In the MSN method, linear correlation between independent and dependent variables is assumed when calculating canonical correlation. However, there may exist complex correlation structures in many forestry applications. In this study, the results were considerably improved when second powers of independent and dependent variables were used. An alternative for canonical correlation is the use of linear regression analysis to search for the weight for different independent variables (Mouer and Stage, 1995). This approach has been used, e.g. in the study by Holmström et al. (2001). When the number of dependent variables is one, or all the canonical correlations are utilized (instead of only the few best) the results of canonical correlation and linear regression are identical. When using regression analysis, the correlation structure may, however, be easier to interpret. Also, the non-linear transformations of independent variables needed to linearize the relationships may be easier to find in a linear regression framework.

Localization of models was considered both using geographical regions and similarity-based neighbourhoods in this study. When the study material was divided according to vegetation zones, the results were not promising in most cases. The results were rather surprising though many stand characteristics, such as tree growth and stem form, are known to vary regionally in Finland (Gustavsen, 1998). However, when the study material was divided, there was rather a low number of observations left in the vegetation categories and it may not have been possible to find an adequate number of suitable neighbours. This suggestion is also supported by the fact that when only the weighting matrix was calculated locally and neighbours were searched from the whole material the regional results were most accurate, at least, in the case of estimates of the number of stems.

It is also evident that diameter distribution is a stand characteristic which has no considerable regional variation (excluding the variation explained by the assessed stand characteristics). Different silvicultural regimes affect it strongly and they are usually not known when sample distributions are measured and used as reference material. When these stands are applied locally, there might be more variation within a region than between different regions. Another reason might be the use of national height and volume equations: these may diminish the between-region variation. However, accurate results were obtained when only local weighting matrices were used to indicate that there is variation according to vegetation zones. This is also worth further research.

In Finland, diameter distribution models are weighted by basal area. When applying these distributions, they are scaled by using the stand basal area which is strongly correlated to the volume of the corresponding stand. In such an application, it is rather difficult to improve the accuracy of prediction concerning stand volume by using regional models or modifications of independent variables. Correspondingly, in this study, the changes in the accuracy of prediction
of stand volume were modest when different MSN models were tested. The only stand characteristic which can be considerably affected is stand structure, i.e. number of stems and shape of the diameter distribution. In this study, it was further observed that, when calculating canonical correlation, it was best to use only variables that describe stand structure. The use of basal-area diameter distribution and basal area as independent variables already define the prediction of volume characteristics to be as accurate as possible.

The localized MSN-based models probably work more effectively in the case of tree characteristics. The size of the reference material is usually more comprehensive when compared with stand level and the variation is more concentrated upon according to local conditions. Non-parametric single tree growth models have already been constructed for the Kuusamo region in north-eastern Finland (Sironen et al., 2003) and the results have been very promising. Correspondingly, when harvester-collected stem data is used as reference material (Malinen, 2003), the effect of stem form can be taken into consideration when calculating volumes and timber assortments.

The results of the local MSN method were quite promising in this study. This implies that if the local neighbourhood is defined effectively, which was not the case when using vegetation zones as a basis for locality, the accuracy of MSN models can be improved. On the other hand, the use of the LAN MSN method, where the locality is also based on similarity-based neighbourhood, did not improve the results. However, the situation might be different, if smaller strata of the data were analysed. In the case of vegetation zones, the division could also have been carried out in different ways. Some calculations were also performed by using different zones, but the accuracy of the results was not improved. One possibility might be to use moving windows as geographical zones.

The results of this study, concerning the modified MSN methods contrast with the study of Malinen (2003). In his study Malinen (2003) found that LAN MSN was able to improve the results considerably but the benefit of the local MSN method was not clear. These differences may be due to the different study materials and applications. In the study by Malinen (2003), all stands were mature and concentrated on a very small geographical area. The use of harvester-collected stem data may also have affected the generalization of stem volumes, since it defines single trees very accurately.

The performance of the MSN method and modified MSN methods depends on the structure of the reference data used. The local MSN method should be more appropriate than the MSN method for the applications, where correlations between and within independent and dependent variables are not constant. If they are, local reduction of reference data would only reduce important information. The LAN MSN method should work most efficiently in the applications, where reference data include grouped observations, i.e. edges inside data. By pursuing a symmetrical neighbourhood, the LAN MSN method identifies these edge situations and prevents the use of a large neighbourhood, if observations are located asymmetrically.

The use of different ways to predict tree height did not give notable results in this study. The effect of the methods compared to generalize tree height was quite modest to stand volume. This was also noticed by Siipilehto (2000). However, the stand characteristics compared in this study do not describe variation in tree height in the most efficient way (see also Siipilehto, 2000). If the accuracy of height distribution is of interest, the predicted tree heights should be compared with original tree heights. This would, however, require study material where heights of all trees have been measured, and such materials considering large areas do not exist.

In most applications the number of stems is not known and the most important task is to find models which describe also the stand structure accurately, although basal-area diameter distribution is used. For such purposes, the use of MSN regression seems to be a very promising alternative. The main reason for this is that the characteristics concerning stand structure were included in the model as dependent variables. Therefore, the accuracy of these variables was also considered in determining the optimal weights, unlike in the percentile method which was used as a reference.

It is also notable that the measurement error of the number of stems is considerably bigger when
compared with the basal area. In the study of Kangas et al. (2002) the standard error of the measurement of the number of stems was 80 per cent, whereas in the case of basal area, it was only ~30 per cent. Therefore, in the field it is better to collect such characteristics which are easy and reliable to measure. The accuracy of other characteristics, such as the number of stems, may then be emphasized in the modelling and calculation phases.

Acknowledgements

The study material was provided by the Finnish Forest Research Institute. The authors would like to thank Dr Jeffrey H. Gove and one anonymous referee for reviewing the article. Thanks are also due to M.Sc. Juha Jumppanen for his assistance in the calculation of the vegetation zones. The study is a part of the project ‘Description of commercial roundwood and its distribution in analysing future timber production possibilities’ funded by TEKES and projects ‘Databasis of timber procurement enterprise and forest mensuration as a basis for operational planning’ (decision number 61973), ‘Statistical modelling in forest management as a basis for operational planning’ (decision number 73393) and ‘The usability of single tree laser scanning in forest planning (decision number 78238)’ funded by the Academy of Finland.

References


Malinen, J. 2003 Locally adaptable non-parametric
methods for estimating stand characteristics for wood procurement planning. *Silva Fenn.* 37, 109–120.


Poso, S. 1983 Basic features of forest inventory by compartments. *Silva Fenn.* 17, 313–349.

Siipilehto, J. 1999 Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. *Silva Fenn.* 33, 281–301.


*Received 19 November 2002*