Comparing parametric and non-parametric modelling of diameter distributions on independent data using airborne laser scanning in a boreal conifer forest

Ole Martin Bollandsás*, Matti Maltamo, Terje Gobakken and Erik Næsset

Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Box 5003, N-1430 Ås, Norway

*Corresponding author. Tel: +47 6496 5753; E-mail: olebo@umb.no

Received 29 January 2013

This study used two different approaches to model diameter distributions on data from 201 field plots in a boreal conifer forest in south eastern Norway using airborne laser scanning. These two methods were a non-parametric most similar neighbour (MSN) approach and a parametric seemingly unrelated regression (SUR) approach to predict diameter percentiles, and their accuracies were compared by validation with an independent dataset. Based on calculated differences between predicted and observed number of stems on the entire validation dataset, we found that SUR gave unbiased results and that MSN slightly underestimated total number of stems. However, both methods overpredicted the number of stems per hectare in the range of 15.6–61.5 stems in the smallest diameter classes (between 4 and 12 cm). If the predicted diameter distributions were converted into basal area per hectare \(G\), both methods gave unbiased results. The average difference for \(G\) was 1.9 per cent of the observed value for the MSN approach. The corresponding number for the SUR model was 12.4 per cent. Neither of these differences were statistically significant \((P > 0.05)\). We concluded that the even though both methods overall yielded accurate results, the MSN approach was more reliable in terms of predicting the number of large trees.

Introduction

The breast height diameter distribution of trees in forest stands is useful information in many respects. Diameter distributions enable the calculation of timber volume and biomass in size classes, which is useful in order to distinguish between total and merchantable volume. It is also an indicator of the biodiversity as large tree-size diversity correlates with large diversity of species (Fries et al., 1997). Furthermore, if diameter distribution information is available, future growth and yield can be predicted by means of single-tree models (Bollandsás et al., 2008), which allows inter-tree competition to be taken into account. Such models are especially useful for stands with great variability in tree sizes. However, from an inventory point of view, the traditional mean stand characteristics needed as input in stand-level models have been easier and cheaper to acquire than higher resolution information such as the diameter distributions. Recent development in remote sensing techniques, however, enables effective retrieval of information also on this level of detail, which allows for higher resolution information that also in many cases is more accurate than the output from the traditional inventories (Næsset et al., 2004).

Data from airborne laser scanning (ALS) have successfully been used to model many biophysical properties of forest stands (Næsset et al., 2004). For example, in recent years, diameter distributions and other structural characteristics of stands have been modelled. Uni-modal diameter distributions, typically found in managed forest stands, have been modelled in several studies utilizing ALS (Gobakken and Næsset, 2004, 2005; Maltamo et al., 2007; Mehtätalo et al., 2007; Breidenbach et al., 2008; Holopainen et al., 2010; Peuhkurinen et al., 2008; Thomas et al., 2008). Diameter distributions for unmanaged forests that tend to be more multi-modal have also been modelled (Maltamo et al., 2006; Bollandsás and Næsset, 2007; Thomas et al., 2008; Peuhkurinen et al., 2011). Furthermore, Maltamo et al. (2005) predicted different structural characteristics in a forest with diverse tree sizes. The results of all these studies show that the technique can accurately reproduce the observed diameter distributions and stand structure. The potential of ALS data for diameter distribution modelling rests on the strong correlation between the three dimensional structures of the canopy as depicted by the ALS point cloud and the diameter distribution. All studies that have been carried out by reproducing the diameter distribution per unit area have utilized low-pulse density data with \(~1\) pulse m\(^{-2}\), except for the studies...
In order to model the diameter distribution from ALS data, the different studies have utilized different methods. The first study dealing with these problems (Gobakken and Næsset, 2004) represented the observed distribution by a two-parameter Weibull distribution. The Weibull was modelled by direct prediction of the scale and shape parameters (parameter prediction method – PPM) and by the parameter recovery method (PRM) (Dubey, 1967; Bailey et al., 1981). PPM and/or PRM have also been utilized in other studies (Maltamo et al., 2006, 2007; Mehtätalo et al., 2007; Breidenbach et al., 2008; Packalén and Maltamo, 2008; Thomas et al., 2008). In another study, Gobakken and Næsset (2005) compared the PRM to a percentile-based approach (Borders et al., 1987) where diameters at 10 percentiles of the empirical basal area distribution and stand basal area were modelled to reproduce the diameter distribution. Later, two studies have utilized this percentile method to model the empirical diameter distribution (Maltamo et al., 2006; Bollandsås and Næsset, 2007). The most recent studies (Packalén and Maltamo, 2008; Peuhkurinen et al., 2008; Holopainen et al., 2010), where diameter distributions are modelled from ALS data, have utilized the k nearest neighbour approach (k-NN) or the most similar neighbour (MSN) approach to impute diameter distributions from stands that are similar according to some measure of distance based on ALS data.

All these methods have been used to accurately estimate diameter distributions. Most of the studies report the standard deviation for the differences between observed total volume per area unit and the total volume computed from the predicted diameter distribution. These precision figures vary between 10 and 20 per cent of total observed volume, which in most cases are acceptable levels. However, the stem frequency in each diameter class can be important. In such cases, for example when the distribution is going to be used as input in growth projections or timber value estimation, evaluation on volume prediction accuracy per area unit may not be the best validation method. Thus, indices, such as the Reynolds index (Reynolds et al., 1988), have also been used to display the absolute deviation between predicted and observed number of stems. Deviations between predicted and observed number of stems on diameter class level have also been used to show the accuracy of predictions (Bollandsås and Næsset, 2007).

The precision of diameter distribution predictions based on ALS data is dependent on several factors. For example, if the diameter distributions to be modelled are multi-modal or irregular, which is often the case in unmanaged uneven-aged forests, the percentile method may be superior in order to reproduce accurate numbers of stems in each diameter (Borders et al., 1987; Maltamo et al., 2000, 2006; Bollandsås and Næsset 2007). Utilizing the properties of a predefined density function like the Weibull may still produce an accurate mean volume but may be inferior in order to reproduce an accurate stem number because the Weibull is a uni-modal distribution in its basic form (Maltamo et al., 2006). However, it is possible to scale such a distribution to also reproduce irregular forms (Deville and Särndal, 1992; Kangas and Maltamo, 2000; Maltamo et al., 2007; Mehtätalo et al., 2007). In any case, the precision is dependent on the correlation between the data used for model calibration and the stands that are predicted. Furthermore, the ALS data must also reflect the same relationships to the forest attributes over both the calibration areas (e.g. sample plots) and the prediction areas. It is also important that the range of the calibration data in terms of forest attributes covers the range of the predicted stands, at least when relationships between attributes are non-linear. For the use of nearest neighbour approaches, such as the k-NN and the MSN, this is of particular importance because the forest attributes of interest are imputed from stands that are similar and these methods never go outside of the range of the calibration data, i.e. they cannot be used for extrapolation. If the data fulfill this requirement, nearest neighbour methods are effective in order to predict forest attributes from ALS data. A study by Packalén and Maltamo (2008), which used the MSN method, reported an accuracy of total volume at stand level of ≏11 per cent RMSE of total observed volume.

The MSN method requires that dependent and independent variables are chosen so that canonical correlations can be established to determine the distance between different observations (Moeur and Stage, 1995). The observations with the shortest distance in terms of the canonical correlation are then the MSNs. The choice of dependent and independent variables will affect the correlations, and it is therefore important that they are selected to reflect strong and stable relationships. Laser data applications usually have very many potential-independent variables, so many that some selection procedure has to be used. Some studies (e.g. Maltamo et al., 2006) have relied on an iterative process to select variables based on RMSE. The MSN method is then run several times using different variables, always evaluating on RMSE of some dependent variable. However, there may be a threat of constructing models that are too detailed and complex by utilizing this approach if there is no restriction of the maximum number of independent variables. Such models may fit really well to the training data, but may not be valid for the whole population from which the training data were sampled. The variable selection process is therefore an important aspect of utilization of non-parametric methods, as for parametric methods.

Parametric models like the Seemingly Unrelated Regression (SUR) have the capability of simultaneous prediction of several dependent variables. The SUR method has therefore been used (Gobakken and Næsset, 2005) to predict diameters at percentiles of the cumulative basal area distributions to reproduce diameter distributions by scaling with a predicted stand basal area. Parametric models are in general versatile if they are calibrated properly, and compared with the MSN method, such models are not limited by the range of the developmental data. However, extrapolation beyond the range of the model development data should be carried out with caution for any model. In many cases, the relationships between dependent and independent variables can be different towards the extreme values where data may be scarce or missing compared with the mean values.

No previous studies have compared the accuracy of non-parametric versus parametric modelling of diameter distributions on the same data. Both of these approaches have successfully been applied for diameter distribution modelling in separate studies. In this study, we used the MSN method (Moeur and Stage, 1995) and the SUR method to model diameter distributions on a study site in a managed forest in southeastern Norway. The objective of the study was to compare the accuracy of the two methods by independent validation. The tree measurements in the validation data were carried out by means of harvesters.
Methods

Study area

This study was conducted in the Aurskog-Høland municipality (59° 50’ N 11° 30’ E, 120–390 m a.s.l.), southeastern Norway. The total area was 960 km². The dominant tree species in the area were Norway spruce (Picea abies (L.) Karst.) and Scots pine (Pinus sylvestris L.), but a large portion of especially younger stands was dominated by deciduous species, mainly birch (Betula pubescens Ehrh.). The study took advantage of an operational stand-based forest inventory utilizing ALS data that was conducted at the time we were planning this research. Thus, in this study, we utilized aerial photographs, ALS data and field data collected for the operational inventory.

Photo interpretation and forest stratification

Aerial photographs were acquired on 28–29 June 2005 using a digital Vexcel UltraCam D camera. The pixel size on the ground was 25 cm. Stereo photogrammetry based on the aerial photographs was used to delineate the forest stands and to determine tree species, age and site productivity of each stand by manual photo interpretation.

Training data collection

The field data used in the training of the MSN model were collected from 201 circular sample plots systematically distributed throughout the study area. The training dataset is identical to that used by Maltamo et al. (2009). The majority of the plots comprising the training dataset were also used by Næsset and Gobakken (2008) in a study on above- and below-ground biomass estimated by using ALS data.

The sample plot inventory was carried out during fall 2006. Each plot had an area of 200 m² in young forest and 400 m² in mature forest. The forest was considered young if it could be characterized as maturity class I in the Norwegian maturity class system that labels forest stands from class I (clear cuts) to V (mature stands). The discriminant ages between maturity classes are dependent on site index, but as a rule of thumb stands are in class II when the dominant height <8 m. On each plot, diameters at breast height (DBH) ≥4 cm were callipered and tree heights were measured on sample trees selected with a relascope. We aimed for 10 sample trees per plot. The selection procedure was first to estimate stand basal area using the relascope and then calculate a quotient to know which relascope trees to select. Number of sample trees per plot ranged between 4 and 13 with an average of 9. The stem diameters were recorded in 2-cm classes. Ground-truth values for Lorey’s mean height (hLorey), mean basal area diameter (Dg), stand basal area per hectare (G), number of trees per hectare (N) and total plot volume per hectare (V) were computed from the field measurements according to the procedures outlined by Næsset (2001).

Validation data collection

The validation data were collected during winter and fall in 2007. It originates from a large clear felling operation of 29 different forest stands over a total area of 40 ha. The data were collected using John Deere 1070D harvesters, which recorded length and diameter of each stem for every 10-cm section along the stem. Volume of each stem was calculated section-wise and summed. No exact tree coordinates were registered by the harvester, but each tree could be attributed to a delineated stand. Furthermore, the harvester targeted mostly merchantable trees and some small trees were therefore left standing. Some trees were also left according to biodiversity standards for forest operations. Thus, a field effort was carried out to measure single trees left by the harvester. These trees were added to the data. If there were several remaining trees in a clump, the clump was geo-referenced by dGNSS. These areas were kept out of the further analyses. This yielded a complete inventory of every tree in the study area. However, it is likely that some of the small trees were registered neither by the harvester, nor by the field effort because harvesters sometimes run them over.

Breast height was set to 1.1 m from the stump, and DBH was measured at this height. The diameter distribution of each stand was calculated in 2-cm classes, corresponding to the training data. The validation data are dominated by pine and a summary of the data appear in the lower part of Table 1.

Laser scanner data

Laser data for this study were acquired under leaf-on conditions on 8–10 June 2006. Additional data were acquired on 6 September 2006 to fill in a minor gap in the data acquired in June. A fixed-wing piper PA31 – 310 aircraft was used in the acquisition. Laser data were collected with an Optech ALTM 3100 laser scanner operated from an altitude of ~1850 m a.g.l. The flight speed was 75 m s⁻¹, and the pulse repetition frequency was 50 kHz. The scan frequency was 71 Hz, resulting in a point density on the ground of ~0.7 m⁻². The maximum scan angle was 15°, but pulses emitted at an angle >13° were discarded during the subsequent data processing.

Initial processing of the data was accomplished by the contractor (Blom Geomatics, Norway). Planimetric coordinates (x and y) and ellipsoidal height values were computed for all echoes. The entire flight campaign was flown as a block of parallel strips with a swath overlap of 15 per cent, and three flight lines were flown orthogonal to the other flight lines and used in matching and correction for systematic errors between swaths. Ground echoes were found and classified using the progressive Triangular Irregular Network (TIN) densification algorithm (Axelsson, 2000) of the TerraScan software (Anon, 2005). A TIN model was created from the

<table>
<thead>
<tr>
<th>Data</th>
<th>Characteristic</th>
<th>Range</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration (201 plots)</td>
<td>V, m³ha⁻¹</td>
<td>28.00–668.10</td>
<td>208.75</td>
<td>123.56</td>
</tr>
<tr>
<td></td>
<td>G, m²ha⁻¹</td>
<td>5.34–50.71</td>
<td>24.25</td>
<td>9.67</td>
</tr>
<tr>
<td></td>
<td>N, ha⁻¹</td>
<td>200–3575</td>
<td>946</td>
<td>536.73</td>
</tr>
<tr>
<td></td>
<td>Dg, cm</td>
<td>9.72–35.20</td>
<td>22.57</td>
<td>5.45</td>
</tr>
<tr>
<td></td>
<td>Ps ( per cent)</td>
<td>0–100</td>
<td>38.3</td>
<td>35.5</td>
</tr>
<tr>
<td></td>
<td>Pp ( per cent)</td>
<td>0–100</td>
<td>55.6</td>
<td>37.4</td>
</tr>
<tr>
<td></td>
<td>Pd ( per cent)</td>
<td>0–11</td>
<td>5.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Validation (29 stands)</td>
<td>V, m³ha⁻¹</td>
<td>139.96–786.26</td>
<td>313.55</td>
<td>122.27</td>
</tr>
<tr>
<td></td>
<td>G, m²ha⁻¹</td>
<td>17.58–86.93</td>
<td>34.12</td>
<td>13.12</td>
</tr>
<tr>
<td></td>
<td>N, ha⁻¹</td>
<td>206–2276</td>
<td>713</td>
<td>416.94</td>
</tr>
<tr>
<td></td>
<td>Dg, cm</td>
<td>18.23–33.36</td>
<td>24.84</td>
<td>3.91</td>
</tr>
<tr>
<td></td>
<td>Ps ( per cent)</td>
<td>0–100</td>
<td>79.9</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>Pp ( per cent)</td>
<td>0–100</td>
<td>18.1</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>Pd ( per cent)</td>
<td>0–12</td>
<td>2.0</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Total area of the validation stand is 41.08 ha.
planimetric coordinates and corresponding heights of the laser echoes classified as ground points. The heights above the ground surface were calculated for all echoes by subtracting the respective TIN heights from the height values of all echoes recorded.

The ALTM 3100 sensor is capable of recording up to four echoes per pulse. In this study, we used the two echo categories classified as ‘first of many’ and ‘single’. Echoes of these two categories were pooled into one dataset; and this aggregated dataset was denoted as ‘first’ echoes.

The height distributions created from the first echoes were used to calculate point-level percentiles for 5, 10, 20, . . . , 80, 90, 95 and 100 percent of the heights (H5, H10, H20, . . . , H80, H90, H95, H100) (see Naesset, 2002), and cumulative proportional canopy densities (D5, D10, D20, . . . , D80, D90, D95) were calculated for the respective percentiles. The height distributions contained only those laser points that were classified as above-ground echoes, using a threshold value of 2 m. H95, for example, denotes the height at which the accumulation of laser echoes in the vegetation is 5 percent, and correspondingly, D90 denotes the proportion of laser echoes accumulating at 5 percent of the maximum height relative to the total number of echoes. Other variables calculated for the sample plots were the proportion of ground echoes versus canopy echoes (VEG) using a threshold value of 2 m, and the average height (Hmean), the standard deviation (Hsd) and coefficient of the variation (Hcoeffvar) of echoes > 2 m above ground level.

### The k-MSN method

The k-MSN method is a non-parametric method that uses canonical correlation analysis to produce a weighting matrix used for the selection of the k MSN variables from the training data. Most similar neighbours are observations that are similar to the target observation (Moeur and Stage, 1995). By using canonical correlations, it is possible to find the linear transformations Uk and Vk of the set of dependent variables (Y, biophysical variables) and independent variables (X, ALS-derived variables) that maximize the correlation between them:

\[
U_k = a_k Y, \quad V_k = \gamma_k X,
\]

where \(a_k\) is the canonical coefficient of the dependent variables \((k = 1, \ldots, s)\) and \(\gamma_k\) is the canonical coefficient of the independent variables. There are \(s\) possible pairs of canonical variates, where \(s\) is equal to the number of \(Y\) or \(X\) variables that is smallest (Moeur and Stage, 1995).

The MSN distance metric derived from canonical correlation analysis is as follows:

\[
D^2_{ij} = (X_i - X_j)' \Lambda^2 \Gamma' (X_i - X_j),
\]

where \(X_i\) is the vector of the known search variables from the target observation, \(X_j\) is the vector of the search variables from the training observation, \(\Gamma\) is the matrix of canonical coefficients of the independent variables, and \(\Lambda^2\) is the diagonal matrix of squared canonical correlations.

### Variable selection and choice of k for the MSN method

The dependent variables used in this study were number of stems per hectare (N), volume per hectare (V) and mean diameter (D0.2) because of their strong correlation with the diameter distribution. These variables could also be regarded as proxies for scale, shape and location parameters of the distribution. Furthermore, the independent variables were selected among the first return laser variables by means of Partial Least Square Regression (PLSR) (Wold et al., 1983; Martens, 2001; Martens and Martens, 2001), which like the k-MSN relies on linear transformations of the independent variables for explaining the variation in the dependent variables. The analysis was performed using the pls-Package (Mevik and Wehrens, 2007) of the R software version 2.8.1 (R Development Core Team, 2008).

Firstly, all candidate-independent variables were simultaneously regressed against the three dependent variables. The PLSR is related to Principal Component Regression and produces orthogonal principal components that are relevant for explaining the variation in the dependent variables. In our analysis, we used two principal components. In each principal component, each variable is weighted according to its importance for explaining the dependent variable variation. These weights are called loading weights. We ranked the loading weights of the independent variables in each principal component and used the top-ranked ones as independent variables in the MSN analysis. To be able to set a threshold for which loading weights were considered large, we plotted the ranked loading weights and set the threshold where the weights displayed a distinct drop.

The MSN analysis was performed using the yalmpute-Package (Crokosten and Finley, 2008) of the R software (R Development Core Team, 2008). The number of neighbours (k) used in this study was \(k = 1\). The rationale for using only one neighbour was that we wanted to be consistent in the way we selected variables. Other variables than the top-ranking ones after the PLSR analysis may be more relevant if \(k\) is set > 1.

### Accuracy assessment

The accuracies of the two different methods at modelling diameter distributions were assessed by means of independent validation. The MSN model was used to impute the MSNs to grid cells in the validation data from the plots of the training data. Then, a difference between the imputed and observed number of stems in each diameter class was computed for each cell. Since we had no exact position of each tree, the observed diameter distribution of each cell within a stand was equal to the average diameter distribution of the respective stand. Similarly, the parametric
SUR model was used to predict the diameter distribution of all grid cells and then differences between the predicted and observed number of stems were calculated by diameter class in the same way as for the MSN method. For both methods, average differences for each diameter class were then calculated for each stand (mean over grid cells) and then average stand differences on diameter class level were calculated and reported. Differences were considered statistically significant if \( P < 0.05 \). Differences between predicted and observed basal area by diameter class were also calculated for both methods. This was done to further differentiate the differences on tree sizes. Small differences in number of very large trees will, for example, yield large differences between predicted and observed volume and basal area, whereas the differences in number are irrespective of size. Furthermore, the error index proposed by Reynolds et al. (1988) was calculated, allowing comparisons to other studies reporting this reliability measure. The Reynolds error index is the sum of the absolute differences between imputed/predicted and observed number of stems over diameter classes, relative to the total number of observed stems (equation 3):

\[
e = \frac{\sum_{j=1}^{n} |n_{ij} - n_{oj}|}{N} \times 100. \tag{3}
\]

where \( n_{ij} \) and \( n_{oj} \) are the imputed and observed number of trees, respectively, in diameter class \( j \), \( j = 1, 2, \ldots, k \), and \( N \) is the total number of trees according to the field inventory.

Results

The MSN model calibration and accuracy

The selection of variables was based on PLSR. The ranking of the variables according to their loading weight from the PLSR for principal component one and two indicated 10 important variables in each component (Table 2). The most important variables in the first principal component were all height variables whereas the most important ones in the second principal component were mostly density variables. Some of the height variables were important in both components.

Table 3 displays differences between imputed and observed number of stems for each diameter class. The smallest diameter classes were greatly overpredicted (up to 62 trees per hectare), and the numbers of stems in the larger classes were more accurately determined. For trees in diameter classes larger than 12 cm, the largest average difference is 12 trees per hectare (overprediction) and most of the differences are non-significant. The accuracy as indicated by the standard deviation of the differences on diameter class level averaged over stands, ranged between 8 and 50 trees per hectare. The MSN method significantly underestimated the number of stems by an average of 144 stems per hectare \( (P = 0.0483) \). The mean Reynolds index value was 59.2 for the MSN method (Table 4).

SUR model calibration and accuracy

Variables used were different for most of the 12 \( \times \) 4 regressions, but both density and height variables were used. The results of this method are reported in Table 3 and like for the MSN method, it overpredicted the number trees in the smallest diameter classes up to 54 trees per hectare. For the diameter classes above 12 cm, there were mostly minor non-significant differences between predicted and observed number of stems. On average, the accuracy of the parametric models, as reported by the standard deviation of the differences that ranged between 10 and 64, seems to be of the same magnitude as for the MSN approach. The average of predicted minus observed total number of stems was \(-70 \) \( (P = 0.304) \), and the mean Reynolds index was 55.1.

Differences between predicted and observed basal area

Table 3 also displays differences between predicted and observed basal area in the different diameter classes. It can be seen that the overprediction of the number of stems in the small diameter classes had little impact on the basal area difference because of the small basal areas of these trees. However, the differences for the smallest classes \( (DBH < 12 \text{ cm}) \) were still significant. As opposed to the total number of stems using the MSN method, which was just below the significance threshold of 0.05 \( (P = 0.0483) \), the total basal area difference of \(-0.6 \text{ m}^2 \text{ ha}^{-1} \) was not significant \( (P = 0.84) \). The mean difference of basal area for the parametric model was larger \((-3.9 \text{ m}^2 \text{ ha}^{-1}) \) even though the difference in total stem number was smaller than for the MSN approach.

Discussion and conclusions

In this study, we tested two different approaches for modelling diameter distributions. These approaches were a non-parametric MSN approach and a parametric regression approach. The comparison between the two modelling approaches was carried out by calculating differences between predicted and observed number of stems for diameter classes in 29 forest stands using both approaches.

The development of the MSN model involved selecting among explanatory variables derived from the laser point cloud. These should be as strongly correlated to some selected stand variables as possible. The stand variables should in turn be correlated to the diameter distribution. Our choice of stand variables was based on the reasoning that number of stems \( (N) \), volume per hectare \( (V) \) and mean basal area diameter \( (D_{bas}) \) together could be regarded as proxies for scale, shape and location of the diameter distribution. Maltamo et al. (2009) also found these variables highly significant in the canonical correlations in their MSN analysis, imputing diameter distributions using the same training dataset as the current study. In fact, their study tested different combinations of dependent variables in the canonical correlation analysis, and \( N, V \) and basal-area-weighted mean diameter were among the best when different models were assessed according to relative
When explanatory variables are selected for the canonical correlations in MSN analyses, one can rely on minimizing RMSE for the stand variables used in the canonical correlation analysis (Maltamo et al., 2006). Also in studies where diameter distributions (Gobakken and Næsset, 2004; Maltamo, 2007) and other stand properties like timber volume (Næsset, 2002; Maltamo et al., 2006) and stand biomass (Means et al., 1999) are modelled directly from laser data using regression techniques, selecting the explanatory variables by means of minimization of RMSE is common. In this study, we tested a selection procedure that does not rely on such minimizations. Instead, we took advantage of the PLSR in which response relevant variance-covariance structures in the data are compressed into principal components (Wold et al., 1983; Martens, 2001; Martens and Martens, 2001). By interpreting the importance of each variable by its respective loading weight in the PLSR, we made our choice of which variables to include in the canonical correlation analysis. Importance values to evaluate variables have also been used by Crookston and Finley (2008) and Hudak et al. (2008). This way of selecting variables may have an advantage because the chance of over-fitting the model is lower compared with relying on an iterative process of selecting the variables that minimize RMSE. Such models, if not carefully evaluated, may seem valid for the calibration data, but may be spurious in relation to the whole population because the established relationships only are valid for just this particular dataset. Minimizing RMSE may give this effect for example if an unsupervised stepwise regression is performed. Such analyses may cause too many variables to be selected in relation to the number of observations (the model is too complex) and selection of the ‘wrong’ variables (spurious relationships). Thus, by relying on ranking variables according to the loading weight of each variable in a PLSR analysis, it is more likely that only the most important variables, providing a stable correlation valid for the whole

<table>
<thead>
<tr>
<th>DBH class</th>
<th>N⁰</th>
<th>N⁰ effective</th>
<th>G⁰ effective</th>
<th>N⁰ effective</th>
<th>G⁰ effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>713</td>
<td>-144 (375)</td>
<td>-0.6 (15.2) ns</td>
<td>-70 (354) ns</td>
<td>-3.9 (13.4) ns</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>18.3 (21.4)</td>
<td>0.0 (0.04)</td>
<td>15.6 (20.3)</td>
<td>0.0 (0.04)</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>47.2 (66.3)</td>
<td>0.2 (0.18)</td>
<td>49.9 (39.6)</td>
<td>0.2 (0.15)</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>61.5 (60.2)</td>
<td>0.4 (0.26)</td>
<td>53.8 (31.5)</td>
<td>0.3 (0.20)</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>27.6 (42.5)</td>
<td>0.3 (0.40)</td>
<td>18.6 (40.5)</td>
<td>0.2 (0.39)</td>
</tr>
<tr>
<td>13</td>
<td>57</td>
<td>10.3 (49.7) ns</td>
<td>0.1 (0.66) ns</td>
<td>-6.8 (64.2) ns</td>
<td>-0.1 (0.85) ns</td>
</tr>
<tr>
<td>15</td>
<td>62</td>
<td>-1.5 (42.5) ns</td>
<td>0.0 (0.75) ns</td>
<td>-12.3 (43.7) ns</td>
<td>-0.2 (0.77) ns</td>
</tr>
<tr>
<td>17</td>
<td>57</td>
<td>1.0 (31.9) ns</td>
<td>0.0 (0.72) ns</td>
<td>-0.6 (31.5) ns</td>
<td>0.0 (0.71) ns</td>
</tr>
<tr>
<td>19</td>
<td>63</td>
<td>-1.6 (42.2) ns</td>
<td>0.0 (1.20) ns</td>
<td>-3.8 (35.8) ns</td>
<td>-0.1 (1.02) ns</td>
</tr>
<tr>
<td>21</td>
<td>63</td>
<td>12.0 (42.7) ns</td>
<td>0.4 (1.48) ns</td>
<td>-1.84 (39.5) ns</td>
<td>-0.1 (1.37) ns</td>
</tr>
<tr>
<td>23</td>
<td>63</td>
<td>-9.9 (45.4) ns</td>
<td>-0.4 (1.88) ns</td>
<td>-5.3 (43.3) ns</td>
<td>-0.2 (1.80) ns</td>
</tr>
<tr>
<td>25</td>
<td>57</td>
<td>-2.7 (36.0) ns</td>
<td>-0.1 (1.77) ns</td>
<td>-0.4 (33.0) ns</td>
<td>0.0 (1.62) ns</td>
</tr>
<tr>
<td>27</td>
<td>62</td>
<td>-7.0 (34.1) ns</td>
<td>-0.4 (1.95) ns</td>
<td>-11.4 (34.1) ns</td>
<td>-0.7 (1.95) ns</td>
</tr>
<tr>
<td>29</td>
<td>45</td>
<td>-0.9 (27.1) ns</td>
<td>-0.1 (1.80) ns</td>
<td>-0.6 (25.5) ns</td>
<td>0.0 (1.69) ns</td>
</tr>
<tr>
<td>31</td>
<td>37</td>
<td>-4.0 (19.1) ns</td>
<td>-0.3 (1.44) ns</td>
<td>-0.5 (20.6) ns</td>
<td>0.0 (1.56) ns</td>
</tr>
<tr>
<td>33</td>
<td>32</td>
<td>-11.8 (21.5)</td>
<td>-1.0 (1.84) ns</td>
<td>-4.2 (20.3) ns</td>
<td>-0.4 (1.74) ns</td>
</tr>
<tr>
<td>35</td>
<td>22</td>
<td>7.0 (17.6)</td>
<td>0.7 (1.67)</td>
<td>-0.7 (11.3) ns</td>
<td>-0.1 (1.09) ns</td>
</tr>
<tr>
<td>37</td>
<td>14</td>
<td>-1.7 (10.5) ns</td>
<td>-0.2 (1.13) ns</td>
<td>-0.5 (12.3) ns</td>
<td>-0.1 (1.32) ns</td>
</tr>
<tr>
<td>39</td>
<td>10</td>
<td>4.6 (18.3) ns</td>
<td>0.6 (2.19)</td>
<td>-3.2 (9.5) ns</td>
<td>-0.4 (1.14) ns</td>
</tr>
<tr>
<td>41</td>
<td>7</td>
<td>-4.4 (7.9)  ns</td>
<td>-0.6 (1.04) ns</td>
<td>-5.0 (9.9) ns</td>
<td>-0.7 (1.30) ns</td>
</tr>
<tr>
<td>43</td>
<td>8</td>
<td>0.0 (8.9) ns</td>
<td>-0.1 (1.33) ns</td>
<td>-10.6 (11.1)</td>
<td>-1.6 (1.69) ns</td>
</tr>
</tbody>
</table>

Table 3: Mean difference between predicted and observed (predicted minus observed) number of stems per hectare (N⁰) and basal area per hectare (G⁰) for different diameter classes as a result of using the non-parametric method (MSN) and the parametric method (SUR).

Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>59.2</td>
<td>25.5</td>
<td>232.0</td>
<td>38.1</td>
</tr>
<tr>
<td>SUR</td>
<td>55.1</td>
<td>27.2</td>
<td>222.7</td>
<td>37.3</td>
</tr>
</tbody>
</table>

Table 4: Mean, minimum and maximum Reynolds index value with corresponding standard deviation by using the MSN and SUR methods for retrieving the diameter distributions of 29 validation stands.
population, are selected. However, it should be remembered that in the case of non-parametric methods, it is possible to select more variables than in the case of regression models. It is still quite safe to use, say, 10–20 variables in non-parametric predictions regarding to model over-fitting (Packalén et al., 2012). In the k-MSN approach, canonical correlation analysis orthogonalizes the large number of predictor variables. Thus, it avoids the problems often encountered in regression (Moeur and Stage, 1995).

The variables that were selected by utilizing our approach (Table 2) were logical given prior expectations. In the first principal component, the top-ranking variables were only height variables and in the second there were mostly density variables. It is also important to point out that the height variables were more important when they were close to the maximum canopy height, and the density variables received a higher importance value as they were affected by more of the vertical structure of the canopy. Since all our dependent variables were correlated to the diameter distribution and therefore also volume (one actually being volume), this seems logical as volume generally is a direct function of a basal area and a height.

In our study, we applied MSN with \( k = 1 \). This differs from previous studies (Maltamo et al., 2006; Packalén and Maltamo, 2008) carried out with ALS data where more accurate results were obtained by using, say, \( k = 5 \) MSNs. On the other hand, in the first publication of the MSN, the methodology was presented based on just one neighbour (Moeur and Stage, 1995). The use of only one neighbour in this study can also be justified as the analysis carried out at grid level and then validated on stand level by aggregating grid level predictions. Each stand can in fact include more than 100 grid cells, and this averaging effect is similar to the effect of using higher values of \( k \) at smaller scale.

The reason for the overprediction of number of stems in the small diameter classes is that the harvester for obvious reasons targeted tree sizes that were merchantable. Most of the small trees were left in the forest. Some large trees were also left, for example for biodiversity reasons. Trees that were left in the forest were measured manually and added to the diameter distribution. However, very few small trees were found standing in the different stands after harvest. The reason is most likely that the harvester had run them over or cut these trees while manoeuvring, leaving them on the ground. For this reason, the diameter distributions in the validation data probably have fewer small trees than the actual number before harvest. Thus, the positive biases in the small diameter classes are most likely less than those reported in Table 3.

It was somewhat surprising that the percentile-based regression model approach led to better accuracies in small diameter classes than the MSN. In previous studies (Kangas and Maltamo, 2000; Bollandsás and Næsset, 2007), which are based on modelling of basal-area-weighted percentiles distribution, the resulting stem frequency distributions have often been considerable overestimates on plot level, leading even to decreasing distribution forms. The likely reason for better accuracy is the effect of averaging over cells that yields stand level results, which probably is stronger in the case of percentiles based distributions than the MSN approach.

In several previous diameter distribution modelling studies (Maltamo, 1997; Kangas and Maltamo, 2000; Maltamo et al., 2007; Packalén and Maltamo, 2008), different estimates of sawlog proportions have been considered. These estimates can be theoretical sawlog volumes that can be calculated by using taper curves or they can be just proportion of sawlog-sized trees. If the results of this study (Table 3) are examined in terms of sawlog-sized trees (DBH > 18 cm), it can be seen that both alternatives yield some underestimates but the MSN was more accurate in predicting the large trees. For trees larger than the 18 cm limit, the MSN predicted unbiased number of stems with an average difference \( (P > 0.05) \) of –1.5 trees per hectare (not reported in the tables). The corresponding number using the SUR approach was –4.9 trees per hectare \( (P = 0.039) \). Also at the very upper end of the distribution, we found that the largest trees were significantly underpredicted by the parametric approach, while the MSN approach yielded an average difference close to zero. This may be explained by the fact that with the MSN approach, the predictions are more likely to stay in range. This means that even if there is a smaller proportion of sampling units with large trees in the calibration data compared with the validation data, the MSN approach is still able to impute these large trees in those validation stands where they appear. With the parametric regressions, this smaller proportion of sampling units with large trees will be reflected in the model. This may have yielded the underprediction of the large trees in the validation data. In fact, in regression the observations close to the maximum range of the data are likely to have the greatest leverage (Montgomery et al., 2001) and thus have greatest influence on the regression.

Both methods produced non-significant mean differences between predicted and observed basal area (Table 3). For the MSN approach, the mean difference was \(-0.6 \text{ m}^2 \text{ ha}^{-1} (P = 0.84)\) or 1.9 per cent of observed mean basal area. If we then assume that heights and volume are predicted with the same models for both the calibration dataset and the validation dataset, this percentage applies also for volume. Compared with other studies, this percentage is quite low. Bollandsás and Næsset (2007) reported volume differences between predicted and observed values of 0.6–12.6 per cent for different strata defined according to stand structure. The corresponding difference from using the parametric model was \(-3.9 \text{ m}^2 \text{ ha}^{-1} (P = 0.14)\), which translates to 12.4 per cent of observed volume using the same reasoning as above. Two other studies (Gobakken and Næsset, 2004, 2005) that have evaluated diameter distribution predictions in mature forests on volume report values of –4.8 to 1.8 per cent and –1.2 to 2.1 per cent. However, both these studies used cross-validation in the evaluation of the models that most likely will yield better figures than an independent validation (Montgomery et al., 2001).

The average standard deviation of the differences between predicted and observed number of stems was lowest for the SUR. The reason for this is that since the SUR is based on regression, it will tend predict average number of stems in a certain diameter class given the explanatory variables. The MSN imputes a complete distribution from a training plot to a validation grid cell and is therefore more likely to impute some extreme values on diameter class level. This effect will decrease with increasing number of training plots.

Table 4 shows the mean, minimum, maximum and standard deviation of the stand level Reynolds index. The mean number for the MSN approach was 59.2 and 55.1 for the parametric method. For both approaches, the results are well in line with cross-validation results of previous studies. Gobakken and Næsset (2004, 2005) reported values from 30 to 77.6 on plot level in even-aged forests using parametric methods. Packalén and Maltamo (2008),
that imputed species-specific diameter distributions of a mixed-species forest, reported the accuracy using a slightly different error index comparing relative stem frequencies for each diameter class and summing the absolute deviations. However, it seemed that their stand level cross-validation gave approximately the same magnitude of the absolute deviations. Bollandsås and Næsset (2007) reported Reynolds index values in the range of 59.7–111.3 based on independent validation results in a size-diverse forest.

This study demonstrated that both non-parametric imputation of diameter distribution and parametric prediction yielded good results when tested using an independent dataset. Even though there were some differences in precision of the predicted number of stems of each diameter class for both approaches, it must also be remembered that the calibration- and the validation-datasets had slightly different species distributions (Table 1). More specifically, the validation data comprise more pine than the validation dataset. The latter mostly comprised spruce. The crown shapes of spruces and pines are somewhat different. Spruce trees are more strictly conical than pines that tend to be more rounded as they mature. Thus, the relationship between laser data and the diameter distribution is different for these two tree species (Næsset, 2004, 2005). This may have affected the results to some degree, and it is likely that the results for both methods would have improved if the datasets comprised the same species distributions.

We conclude that both the non-parametric and the parametric approach predicted diameter distributions well in line with the observed distributions. The difference between the results of the two approaches was mainly that the non-parametric approach seemed to better stay in the range of the observed distribution.

Acknowledgements
We thank Viken Skog BA for giving us access to the field data and Blom Geomatics AS for collection and processing of the airborne laser scanner data. We also thank two anonymous reviewers who have provided excellent comments and suggestions that have improved the article.

Conflict of interest statement
None declared.

Funding
This work was supported by the Research Council of Norway (project #173911/I10: ‘Assessing wood properties of forest resources’) and the Norwegian Forest Trust Fund.

References


Comparing parametric and non-parametric modelling


Thomas, V., Oliver, R.D., Lim, K. and Woods, M. 2008 LiDAR and Weibull modeling of diameter and basal area. For. Chron. 84, 866–875.
