www.metla.fi/silvafennica · ISSN 0037-5330
The Finnish Society of Forest Science · The Finnish Forest Research Institute

Using Mixed Estimation for Combining Airborne Laser Scanning Data in Two Different Forest Areas

Aki Suvanto and Matti Maltamo

Suvanto, A. & Maltamo, M. 2010. Using mixed estimation for combining airborne laser scanning data in two different forest areas. Silva Fennica 44(1): 91–107.

Airborne laser scanning (ALS) data have become the most accurate remote sensing technology for forest inventories. When planning new inventories the costs of fieldwork could be reduced if datasets of old inventory areas are effectively reused in the new area. The aim of this study was to apply mixed estimation using a combination of existing and new field datasets in area-based approach. Additionally, combining datasets with mixed estimation was compared with constructing new local models with smaller datasets. The two forest study areas were in Juuka and Matalansalo, which are located about 120 km apart in eastern Finland. ALS-based regression models were constructed using datasets of Matalansalo (472 reference plots) and Juuka (10–212 reference plots). Models were developed for the basal area median tree diameter and height, mean tree height, stem number, basal area and volume. The work was based on a simulation approach which involved five methods for approximating the regression coefficients. The first method merged the datasets using ordinary least squares (OLS) regression models, whereas the second and third methods combined datasets using mixed estimation on different weighting principles, and the final two estimated local models with predetermined and new independent variables. The results indicate that mixed estimation can improve the accuracy of derived stand variables compared with basic OLS models. Additionally, a sample of 40–50 plots was enough to build local models for basal area and volume and produce at least the equal accuracy of results than any other methods in this study.

Keywords airborne laser scanning, area-based method, mixed estimation, regression models

Addresses *Suvanto*, Blom Kartta Oy, Teollisuuskatu 18, FI-80100 Joensuu, Finland; *Maltamo*, University of Eastern Finland, School of Forest Sciences, P.O. Box, FI-80101, Joensuu, Finland
E-mail aki.suvanto@blomasa.com

Received 31 January 2008 **Revised** 27 August 2009 **Accepted** 19 January 2010
Available at http://www.metla.fi/silvafennica/full/sf44/sf441091.pdf

1 Introduction

The area-based method has been the main approach to using airborne laser scanning (ALS) for large forest inventories because of its cost efficiency and the accuracy of the forest data obtained (Eid et al. 2004). This has made it one of the principal operational methods used to collect information on forest structure and resources in Norway (Næsset 2004a, 2004b, 2007, Olsson and Næsset 2004). Several studies have been published concerning the accuracy of the forest data produced by this technology and methodology, and demonstrating many different kinds of regression models for predicting total forest characteristics such as mean diameter, dominant tree height, number of stems, basal area and volume from laser point data (Næsset 2002, 2004a, Holmgren 2004, Thomas et al. 2006, Maltamo et al. 2006, Hollaus et al. 2006, Jensen et al. 2006). The results have indicated that the area-based method could produce at least as accurate results as traditional field measurements.

Tree species-specific results can be estimated using laser scanner data and aerial photographs (Packalén and Maltamo 2006). However, reliability and accuracy of the area-based method in certain types of forest can be challenging and difficult. Broadleaf forests, in particular, might produce overestimated results because the canopy differs in size and shape from that in coniferous forests and is more closed (Næsset 2004a, 2005). According to his studies, it is highly recommended to treat deciduous stands as separate strata when applying the area-based method. In addition, there have been doubts about the transferability of ALS-based regression models, so that the models tend to be estimated locally for continuous geographical areas.

There are many factors which can affect ALS-based regression models and their estimation results. The flying parameters of the laser scanner, for example, including altitude, point density, type of scanner and field of view (Holmgren et al. 2003, Næsset 2004c, Holmgren 2004, Maltamo et al. 2006), can influence the properties of a single point and then the whole point cloud. Furthermore, a major difference in forest structure, e.g., growing stock, between inventory areas can cause significant over-estimation or under-estimation of forest characteristics. According to Thomas et al. (2006), the best laser quantile is different for low- and high-density laser data and for homogeneous and mixed stands. As a result, transferring ALS-based regression models between forest areas could be a cumbersome process. According to Næsset et al. (2005), it is possible to combine reference data and laser scanner data from different inventories, but it is necessary to check the compatibility of the forest conditions and laser scanners. In Næsset (2007), ALS data of two inventory areas including the same scanner and flying parameters were combined, and the inventory areas were taken into consideration using dummy variables in derived regression models. In his study, dominant height was the only dependent variable which was significantly different to the other forest characteristics.

Jensen et al. (2006) concluded that some biophysical properties can be estimated accurately in large forest areas. They estimated and tested ALS-based regression models in five geographically separate forest areas with mixed-conifer stands but few deciduous species. Therefore, ALS models cannot be generalised to other forest areas without additional testing. Lefsky et al. (2005) suggested there might be a unique relationship between ALS measurements and stand structure in forests where the main tree species are coniferous trees because the canopy is more or less conical. An application of this above mentioned unique relationship to canopy height estimation was attempted by Hopkinson et al. (2006), who demonstrated that the standard deviation of first and last pulse returns is a robust variable for estimating canopy height for different vegetation types and heights and for different lidar survey configurations. Their study included 13 separate datasets representing five sites in Canada with measured average heights ranging from <1 to 24 m. These datasets were collected using four different ALTM (airborne laser terrain mapper) sensors during 2000 and 2005 (Hopkinson et al. 2006). However, their study involves only height estimation, which is probably the most accurate variable to predict using ALS data. Finally, Breidenbach et al. (2007) used mixed linear modelling to combine ALS datasets from Germany and the USA.

Under Finnish conditions, Uuttera et al. (2006) tested different remote sensing methods for two separate forest areas to estimate stand-level forest characteristics. They concluded that using the area-based method with regression models produced the most accurate results. These models were constructed in a separate forest area and models were not calibrated or re-estimated for the two test areas, which were located 150 and 300 km away from the model area. The regression models produced fairly accurate stand-level results, but the major concern was that these entailed a significant bias for some stands. However, these stands were dominated by deciduous trees and this bias might be because of the difference between coniferous and deciduous trees. Moreover, structural differences between forest areas can influence a model's accuracy. Uuttera et al. (2006) conjecture that ALS-based regression models are transferable between inventory areas but bias might be a major problem and this must be taken into consideration. Therefore, a calibration for local reference data might help produce more accurate and unbiased results. Another aspect is reducing fieldwork costs in a new inventory area by considering whether datasets of old inventory areas could be effectively reused in modelling and constructing new models in the new target forest area.

One possible method for combining different datasets in ALS-based forest inventories is to apply mixed estimation. This method has mainly been employed in economics, but there has been some research into its applicability to forestry (Korhonen 1993, Roesch 1999). Korhonen (1993), for instance, used mixed estimation to calibrate volume functions for Scots pine sample tree material from two national forest inventories (NFI7 and NFI8). In his study, the pine volume model was constructed in sampling simulations for the NFI8 data by taking information for the mixed estimation from the NFI7 data and comparing the result with an OLS model constructed without prior information. In this case, the main result was that mixed estimation produced better results when there were fewer sample trees, but the OLS estimator was better when there were several hundred sample trees. According to Korhonen (1993), the weight assigned to the prior information could influence the final results. Roesch (1999) found mixed estimation to be favourable method to estimate basal area when compared to moving averages and imputation.

The aim of this study was to apply mixed estimation using a combination of existing and new field datasets. Additionally, combining datasets with mixed estimation was compared with constructing new local models with smaller datasets. All the ALS-based regression models are constructed for plot-level total forest characteristics, such as basal area median tree diameter and height, mean tree height, stem number, basal area and volume.

2 Materials and Methods

2.1 Study Areas

The two sites studied here, Matalansalo and Juuka, are located approximately 120 km apart in eastern Finland and represent typical managed forest of the southern part of the boreal forest zone. Matalansalo represents an old inventory area and Juuka a new target area for which we wished to estimate plot-level forest characteristics. The two forest areas are fairly similar in structure. The tree species composition at Juuka consists of Scots pine 59%, Norway spruce 30% and deciduous trees, mainly birch, 11%. These proportions are fairly similar at Matalansalo (58% Scots pine, 34% Norway spruce, 8% deciduous). The proportions of the fertility classes are moist sites (*Myrtillus* type) 48%, dry sites (*Vaccinum* type) 47% and poor sites (*Calluna* type) 5% in the Juuka forest area and grass-herb sites 8%, moist sites (*Myrtillus* type) 49%, dry sites (*Vaccinum* type) 42% and poor sites (*Calluna* type) 1% at Matalansalo. The stand development classes of the plots were young 37%, middle-aged 39% and mature 24% at Juuka, whereas Matalansalo had 27% young forests, 42% middle-aged and 31% mature. Topography range was 80–160 metres above mean sea level in Matalansalo, whereas in Juuka the range was 145–250 metres.

The field data and remote sensing material were acquired in 2004 for Matalansalo and in 2005 for Juuka. The reference sample plot material was similar in both test areas. Circular field plots with a radius of nine metres were measured; the total number of reference plots was 472 at Matalansalo and 212 at Juuka. However, the principle

Table 1. Plot-level forest characteristics in the Juuka test area. Min = Minimum value, Max = Maximum value, Mean = arithmetic mean value, Sd = Standard deviation. Dgm = diameter of basal area median tree, Hgm = height of basal area median tree, h_ mean = mean height of all trees in a plot, N = stem number, G = basal area, V = volume.

	Min	Max	Mean	Sd
Dgm, cm	9.00	32.80	18.23	5.01
Hgm, m	6.23	25.33	14.87	3.87
h_mean, m	5.65	21.44	11.85	2.86
N, ha^{-1}	393	4126	1508	672
G, m^2 ha^{-1}	4.40	55.18	23.76	8.82
V, m^3 ha^{-1}	15.75	506.40	173.12	90.22

Table 2. Plot-level forest characteristics in the Matalansalo test area. Min = Minimum value, Max = Maximum value, Mean = arithmetic mean value, Sd = Standard deviation. Dgm = diameter of basal area median tree, Hgm = height of basal area median tree, h_mean = mean height of all trees in a plot, N = stem number, G = basal area, V = volume.

	Min	Max	Mean	Sd
Dgm, cm	7.60	43.60	19.83	6.46
Hgm, m	6.00	30.55	17.03	5.13
h_mean, m	5.69	24.34	12.78	3.42
N, ha^{-1}	275	4048	1507	692
G, m^2 ha^{-1}	4.45	48.96	24.68	8.05
V, m^3 ha^{-1}	16.05	601.68	203.37	103.51

for sampling the reference plots differed between the forest areas because they formed a systematic network inside a single stand at Matalansalo but were randomly distributed at Juuka. The position of the centre of each plot was determined using GPS (Global Positioning System). All trees with a diameter at breast height (DBH) of more than 5 cm were measured and their species and storey class were recorded. The height of a sample tree was measured for each tree species and storey class in a plot, and the heights of tally trees were calculated using the species-specific height models of Veltheim (1987). Model tree heights were calibrated so that the measured and estimated heights of each sample tree were equal. Finally, the same correction factor was applied consistently to all trees of the same species and storey class within one plot. Tree volumes were calculated using the volume models of Laasasenaho (1982). The plot-level forest characteristics for the two inventory areas are presented in Tables 1 and 2.

2.2 Laser Scanner Data

The remote sensing data for the Matalansalo test area were acquired in August 2004 using an Optech ALTM 1233 laser scanner, which produced a point cloud georeferenced in terms of the x, y and z coordinates. The accuracy was 0.75 m for x and y and 0.25 m for z. The flying altitude was 1500 m above ground level and the field of view was 30 degrees. These flying parameters gave a swath width of approximately 800 m and a point density of 0.7 laser pulses per square metre. The scanning in the Juuka test area was performed in August 2005 using an Optech ALTM 3100C laser scanner at a flying altitude of 2000 m and a field of view of 30 degrees. The pulse repetition frequency was 50 kHz at Juuka and 33 kHz at Matalansalo. The swath width at Juuka was approximately 1070 m and the pulse density 0.56 pulses per square metre. First and last pulse data were recorded in both areas.

A digital terrain model (DTM) was generated for each area from the laser point data, which required separation of the points into ground and vegetation hits. The classification was performed on the Terrascan software (http://www.terrasolid.fi) using a method based on Axelsson (2000). Classified ground points were calculated as averages for each raster cell. Raster values for cells containing no data were calculated using Delaunay triangulation and the bilinear interpolation method. This led to the creation of a DTM raster with a cell size of 1 m for Matalansalo and 2.5 m for Juuka.

Thus, the laser data were recorded using two different types of laser scanner and with different flying parameters. The manufacturer of the scanners were the same in both cases, but the model is different (the ALTM 3100C has a higher pulse frequency). According to previous research, point density is not a critical issue for applying the area-based method and its results (Holmgren 2004, Maltamo et al. 2006, Gobakken and Næsset 2007). However, Næsset et al. (2005) mention that each individual instrument has a unique specification even when belonging to the same production series, and as such they can produce different canopy heights and degrees of canopy penetration. Another major difference of these laser scanners is the technique for recording the first and last pulse data. The ALTM 1233 typically records these two pulse types because it contains two separate electronic circuits. You normally get two echoes for every pulse but it is not always possible to distinguish between them at fairly low height ranges (Næsset, personal communication 2007, Hopkinson et al. 2006). The ALTM 3100C can record four pulse types: only echo, first only echo, last only echo and intermediate echo. In our processing stage, the laser data of the ALTM 3100C were classified into first and last pulse data. The first pulse data consist of the only echoes and first only echoes and the last pulse data of the only echoes and last only echoes. The intermediate echoes were eliminated entirely. All these technical laser scanner issues can affect the results of this research, but the laser scanner data can be assumed to be similar in quality throughout.

Various laser height metrics were calculated for both inventory areas and for each reference plot. Laser point heights less than 2 m were classified as ground points and all the other points as vegetation hits (Næsset 2002, 2004a). Laser canopy height percentiles such as 5, 10, 20, 30…, 90, 95 and 100% were calculated from the vegetation hits (Næsset 2004a), and proportional densities were calculated for these height quantiles. Standard deviation, mean values, coefficients of variation and the proportions of vegetation hits were also computed from the laser point data. This was done separately for the first and last pulse data.

2.3 Methods

Mixed estimation can be employed whenever there are two datasets available: a sample from the current target population and an auxiliary dataset, which should be fairly similar to it (Lappi et al. 2006). When combining the datasets, less weight should be given to the observations from the auxiliary dataset than to those from the target population (Lappi et al. 2006). In general, mixed estimation produces smaller a mean square error than the ordinary least squares (OLS) method, but can also produce biased results (Lappi et al. 2006).

Let y_1 and y_2 be the vectors of the dependent variables for the target and auxiliary populations, respectively, and X_1 and X_2 the model matrices from the target and auxiliary datasets. In mixed estimation, the regression coefficient $\hat{\beta}$ is estimated using Eq. 1 instead of the OLS Eq. 2.

$$\hat{\beta} = \left(X_1'X_1 + \lambda X_2'X_2\right)^{-1}\left(X_1'y_1 + \lambda X_2'y_2\right) \tag{1}$$

$$\hat{\beta} = \left(X_1'X_1\right)^{-1}\left(X_1'y_1\right) \tag{2}$$

Mixed estimation is a weighted least squares method, with the weights of the target population equal to one and the observations of the auxiliary dataset amounting to λ, which describes the weight of auxiliary data (Lappi et al. 2006). According to Theil and Goldberger (1961), λ should be estimated using the models' residuals ratio. This ratio is the model residual error for the target population divided by the auxiliary data model residual error. In this study, such weighting might place too much weight on the auxiliary data. Therefore, this hypothesis will be examined during the study. It is expected that it would be better to use weights which provide the same number of plots for both inventory areas otherwise the reference plots at Juuka, which correspond to the target area, will obtain more weight than the auxiliary data at Matalansalo.

Weights for the auxiliary observations in the mixed estimation were defined using the proportions of reference plots from the target population: 0.5, 0.6, 0.7, 0.8 and 0.9. The actual λ value for mixed estimation is solved from these proportions. A proportion of 0.6 and a chosen sample size from Juuka of 30 plots denote that the proportion of the Juuka target data is 60% and the proportion of Matalansalo reference plots is 40%, corresponding to 20 sample plots. Therefore, the actual λ value in mixed estimation and the weight of auxiliary data would be 20/472. This weighting principle was applied to all proportions and is derived from the fact that the differences in plot-level forest characteristics seemed to be fairly clear and the regression models constructed from Matalansalo data produced some overestimations in the case of the Juuka forest area.

The first stage in this study was to develop linear regression models for the six dependent forest characteristics: basal area median tree diameter (Dgm) and height (Hgm), mean tree height (h_mean), stem number (N), basal area (G) and volume (V). These abbreviations are not perhaps generally accepted but used only as acronyms in the result tables of this study. The most difficult task was defining forms of regression models and finding explanatory variables which would be significant at the 5% level for both inventory areas. The structure of the regression models needed to be fairly simple and the number of explanatory variables need to be minor because of the major multicollinearity of ALS-based height quantiles and percentiles. All regression models and calculations for this study were constructed using R statistical software (R Development Core Team 2009). Forms of the final regression models were tested carefully using different combinations and transformations of independent and dependent variables such as square root, natural logarithm, square and inverse. R-program's regsubset-package and StepAIC-procedure was utilised on a selection of independent variables for each dependent variable.

This research work was implemented using a simulation approach. In mixed estimation, the reference data of Matalansalo were always included as auxiliary data, whereas the number of target population plots in Juuka and the weight of auxiliary data changed during the simulation. Simulation started with a sample of 10 target area plots from the Juuka test site, and plots were added one by one until they had all been included. Each sample size iterated 100 times and plots were selected randomly for each sample. Five different methods were used to approximate the regression coefficients in the simulation procedure. The first method is described in the results as a normal estimation, which implies that the plot sample from the Juuka test area was added to the Matalansalo plots and models' coefficients were solved using OLS regression. The second and the third approximating options involved the use of mixed estimation. One selected the weights for the auxiliary data using the residual ratio. The other method was a more heuristic approach. This denotes that the weights in mixed estimation treat the inventory areas as equal or assign more weight to the target population. The fourth method was constructed from a local OLS model, and estimated the regression coefficients using the selected sample of plots from the Juuka test area and the original, predetermined explanatory variables (Table 3). In the fifth method, only the sample plots from the Juuka forest area were used as reference data and new independent variables were selected using an automatic variable selection method.

The automatic variable selection method was implemented using R-program's regsubset-package. This package is included in the leaps library and a more detailed description is explained in Miller (2002). In the automatic variable selection method, the level of significance for each explanatory variable was 5% and the maximum number of independent variables was three. The automatic variable selection method occasionally produced regression models which were not significant, especially in the early stages of the simulation and when the number of reference plots was small. Therefore, it was necessary to ensure that all independent variables were significant, but in some cases at least one independent variable had to be left in the regression model.

The accuracies of the derived plot-level forest characteristics are presented here in terms of the relative root mean square error (RMSE) and relative bias (bias).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2}{n}} \tag{3}$$

$$\mathrm{bias} = \frac{\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)}{n} \tag{4}$$

where n is the number of sample plots, X_i is the observed value and $\hat{X}_i$ the predicted value for plot i. The relative RMSE and bias for each forest characteristic were calculated as percentages by dividing the absolute RMSE and bias by the true mean characteristics. These accuracy figures are calculated for the Juuka test area as a whole, which contains 212 reference plots, and the final results are an average of 100 simulations for each plot sample size.

3 Results

The ALS-based regression models were originally constructed for the Matalansalo test area and so it was necessary to check that these models and their independent variables were also significant at the 5% level in the Juuka test area (Table 3). These regression models were used throughout the simulations and new regression coefficients were solved in each iteration. Therefore, there was no need to present the exact regression coefficients because our main focus was to demonstrate how these results could change when transferring these models to other inventory areas. Moreover, the structure of the models remains constant in normal estimation, mixed estimation and estimating local model parameters. In the automatic variable selection method, new independent variables were selected during these simulations. All these ALS-based regression models are fairly simple in structure and the number of independent variables is from one to four.

The results are presented here as separate graphs for the various forest characteristics, with each graph containing regression coefficients estimated in the five different ways (Figs. 1–5). An abbreviation *normal* in the graphs denotes that the plot sample from the Juuka forest area was added to the reference plots for Matalansalo. This is the normal procedure for merging datasets and constructing regression models. *Mixed estimation 0.5* denotes calibration of the original ALS-based regression models using the mixed estimation methodology. In other words, the two inventory areas are taken as equal in spite of the differences in the number of sample plots in the data. *Mixed estimation 0.9* refers to mixed estimation in which the proportion of the target data is 0.9 relative to the auxiliary dataset and λ is solved according this assumption. Corresponding proportions of 0.6, 0.7 and 0.8 of the target data were also used in mixed estimation during the simulation process. The major finding was that the result curves obtained by mixed estimation approached each other and the local model results. Therefore, the major results were achieved using proportions of 0.5 and 0.9 in the mixed estimation procedure. A model's residual ratio weighting was also investigated in connection with the mixed estimation method. The final result was that this weighting was not powerful enough and produced somewhat worse results than the other weighting method. The abbreviation *local model* in the graphs denotes that the regression model parameters were estimated using predetermined explanatory variables (Table 3) and the sample plots of the Juuka forest area. The fifth method, automatic variable selection, is described in the curve labelled *local variable selection*. This implies that the new independent variables were selected and new regression coefficients estimated based on the reference plot sample from the Juuka test area.

Table 3. Explanatory variables selected from the laser point data to predict the plot-level forest characteristics. The explanatory variables are abbreviated as follows: f or l denotes first or last pulse data, the prefix h describes a particular height quantile in the laser point data, hmean is a mean value of the laser pulses and veg denotes the proportion of vegetation hits which describe the laser pulses reflected from the canopy compared to the total number of the laser pulses.

	ln(Dgm)	ln(Hgm)	ln(h_mean)	ln(N)	ln(G)	ln(V)
f_{veg}	x					
f_{veg}^2	x			x		
$\sqrt{f_{h20}}$	x					
$\sqrt{f_{h60}}$	x					
$\ln(f_{h80})$		x				
l_{hmean}^2				x		
l_{veg}^2				x		
$\ln(f_{veg})$					x	x
$\ln(f_{hmean})$					x	x
l_{veg}					x	
$\sqrt{1_{veg}}$						x
l_{hmean}			x			

Mixed estimation resulted in slightly more accurate results than normal OLS estimation. The results also

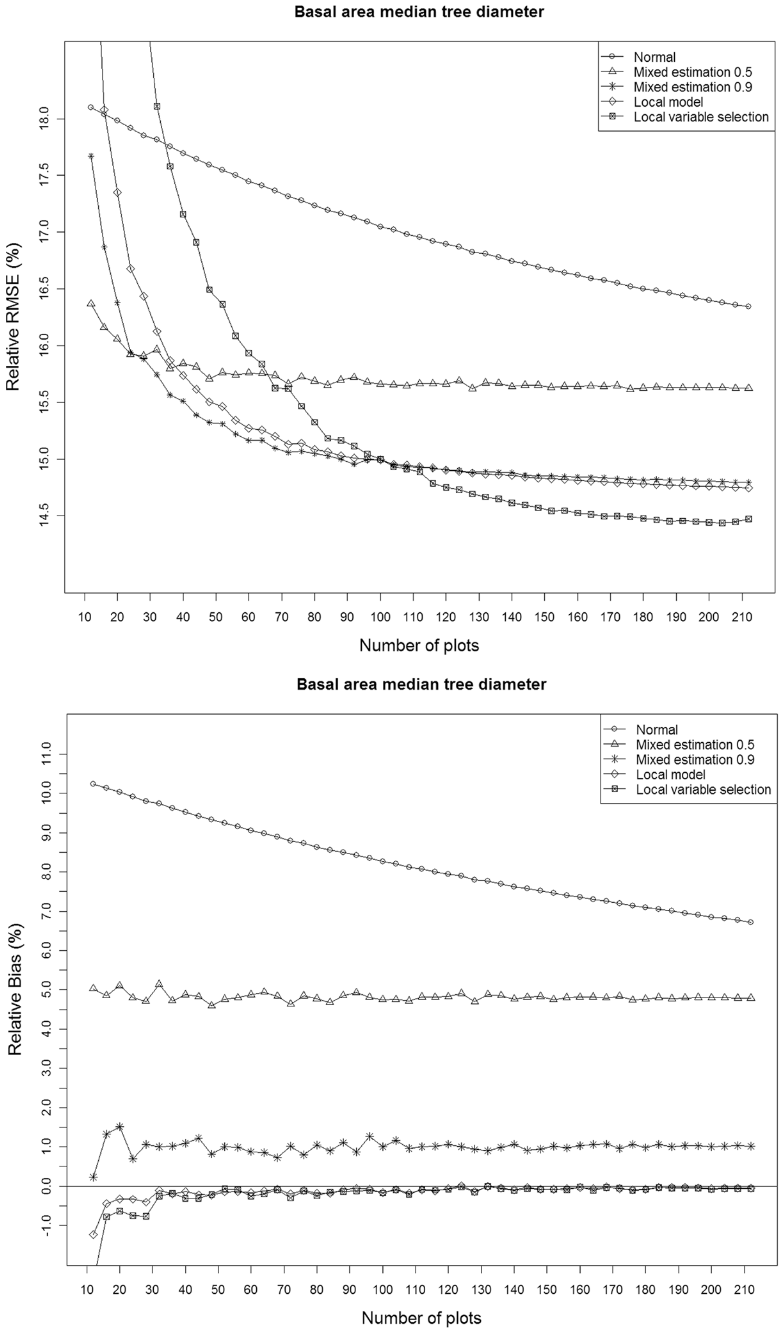

Fig. 1. Results for basal area median tree diameter obtained using the five estimation methods.

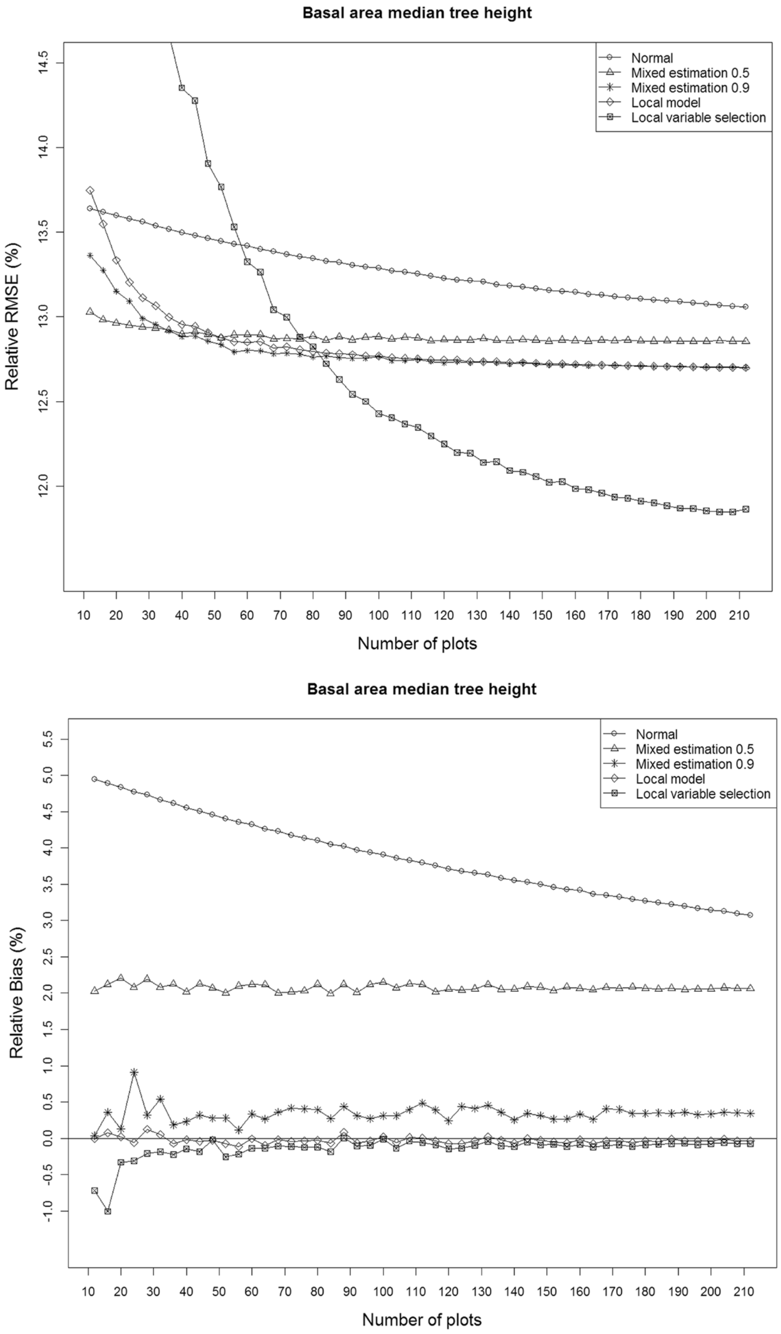

Fig. 2. Results for basal area median tree height obtained using the five estimation methods.

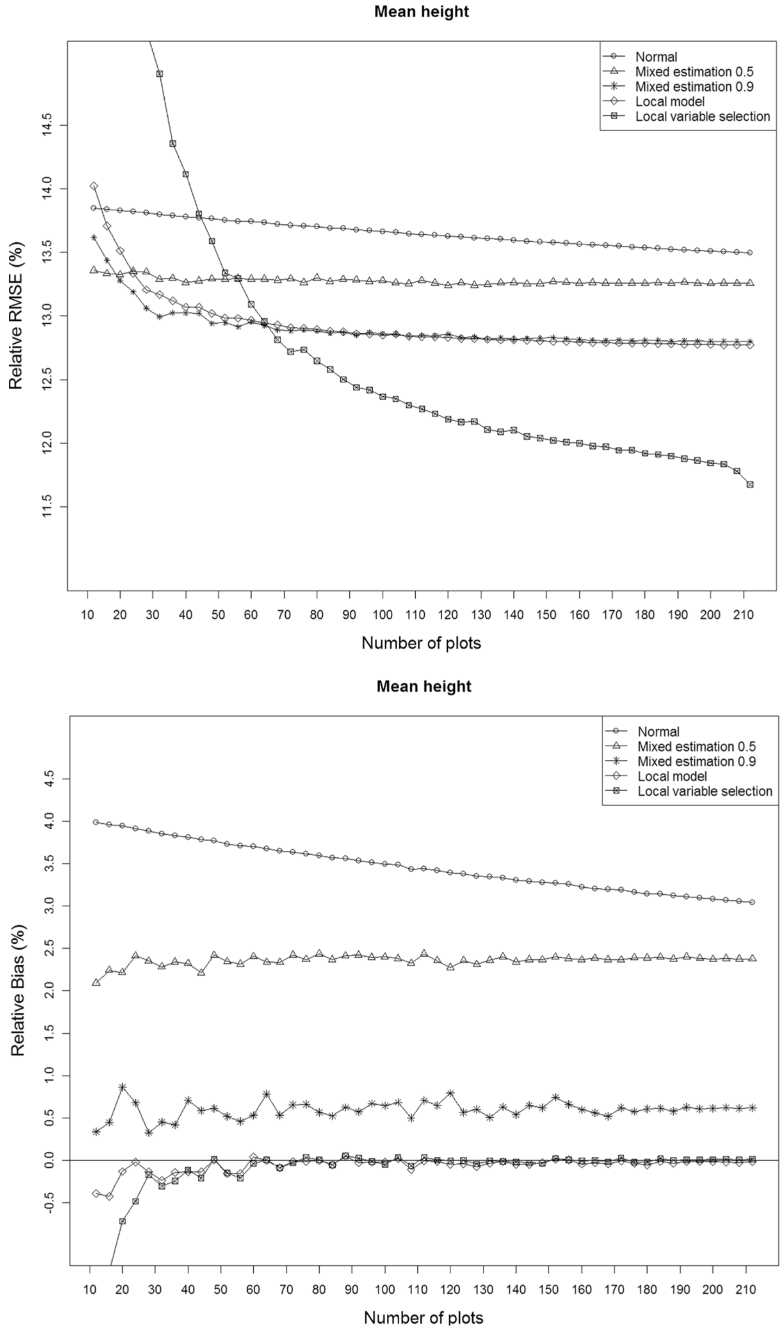

Fig. 3. Results for mean tree height obtained using the five estimation methods.

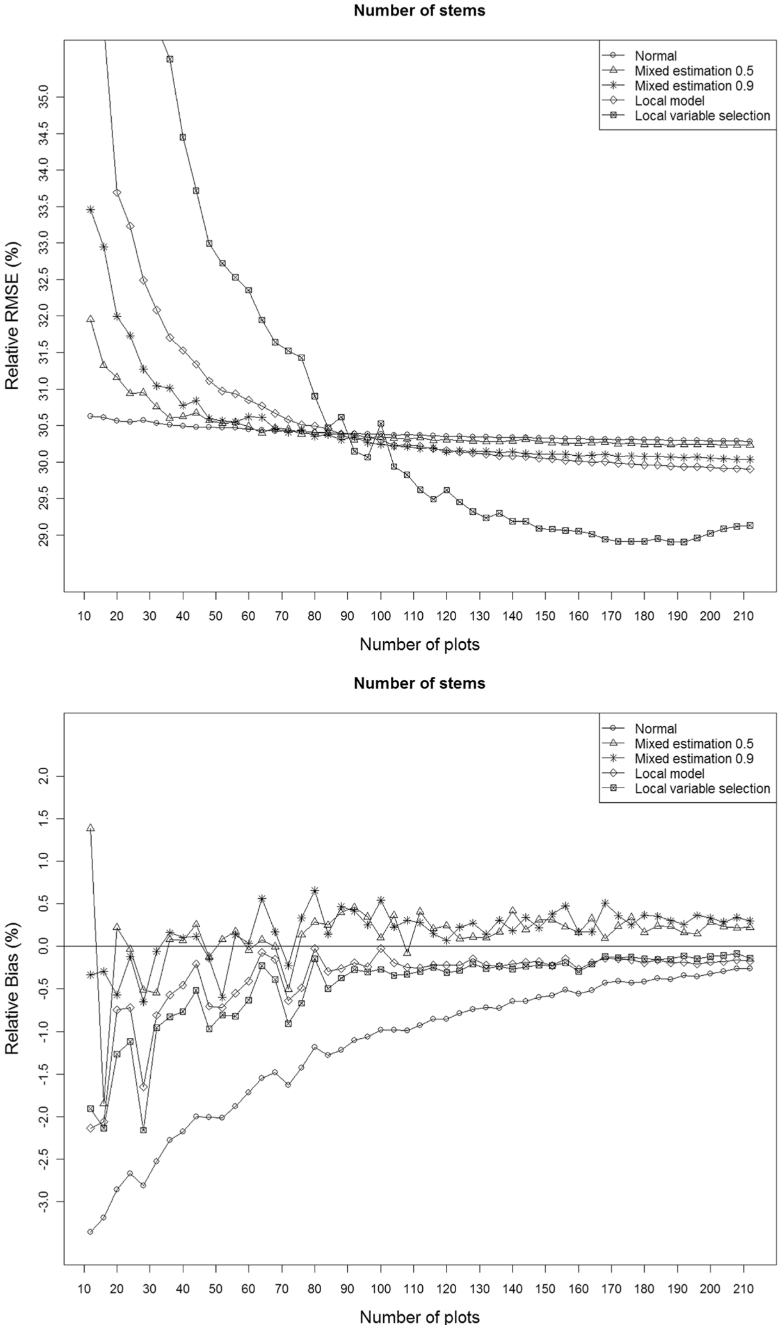

Fig. 4. Results for stem number obtained using the five estimation methods.

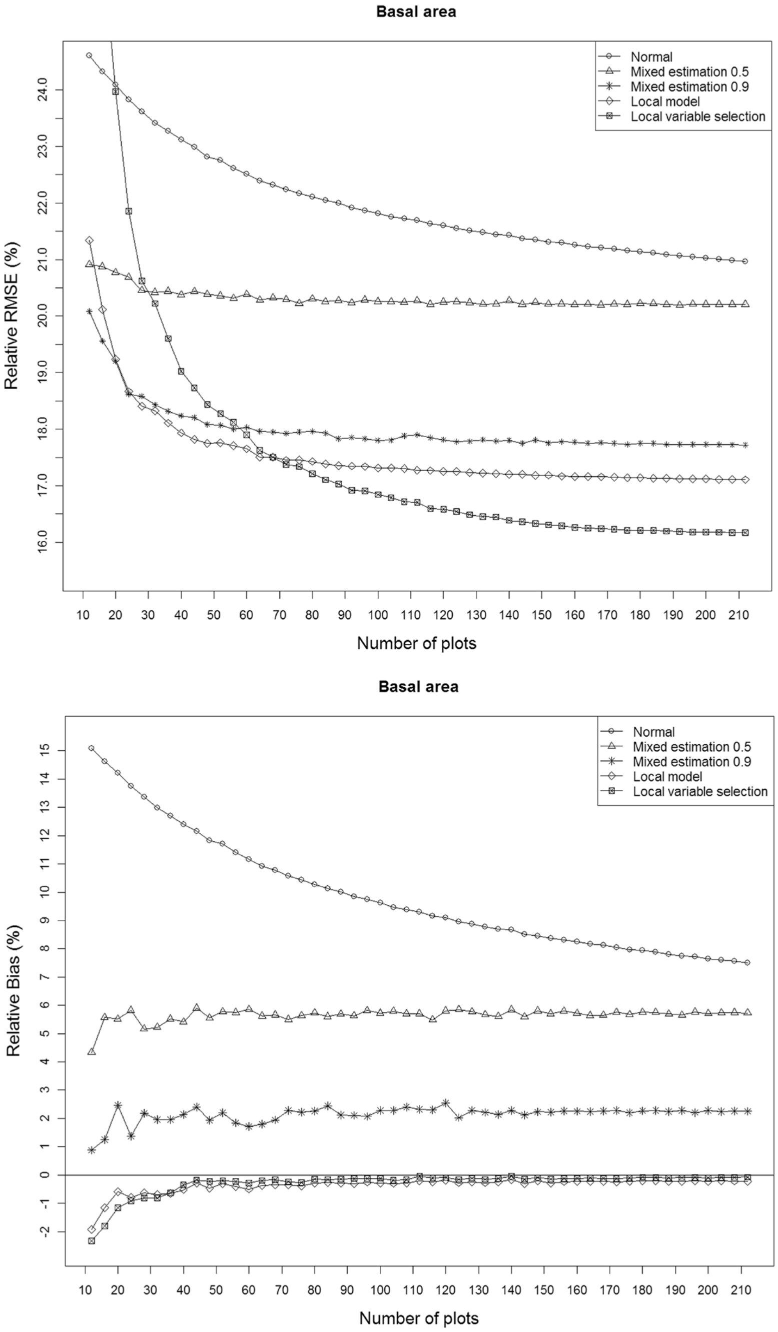

Fig. 5. Results for basal area obtained using the five estimation methods.

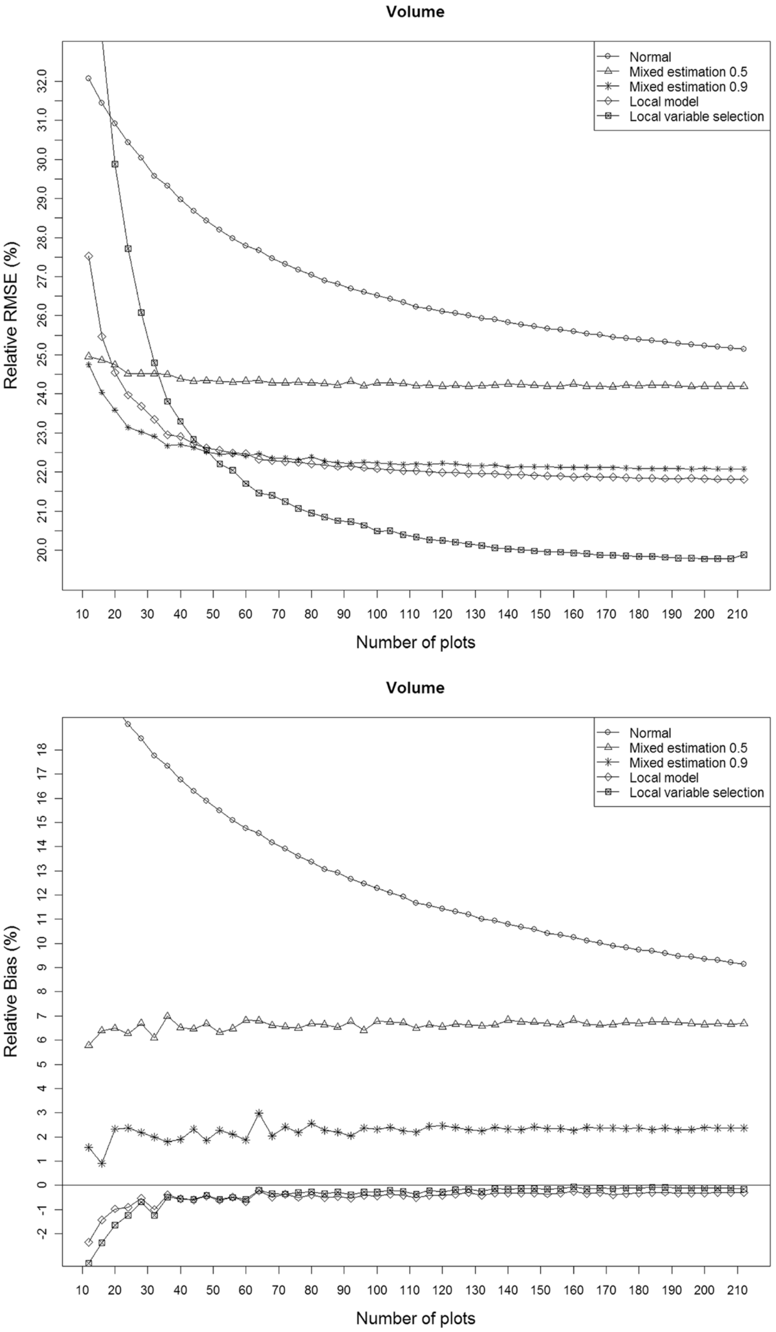

Fig. 6. Results for total volume obtained using the five estimation methods.

suggest that local regression models can produce more accurate results than mixed estimation even when the number of sample plots is fairly small. Mixed estimation produced better results for basal area and volume when the number of reference plots was 20–30 and for the other forest characteristics when the number was approximately 40–50. The major difference between mixed estimation and local models or the automatic variable selection method was in estimating the number of stems. There have to be at least 80 reference plots before local models produce results that are at least as good as those obtained by mixed estimation. It is also worth noting that mixed estimation with a proportion of 0.5 produced better results than mixed estimation with a proportion of 0.9 at the beginning of the simulation. This was seen for Dgm, Hgm and N.

The bias is generally smaller in local models than in mixed estimation because of the mixed estimation methodology. Mixed estimation produced more accurate results than local models for small sample sizes. A sample size of 40 or 50 plots for basal area and volume was enough to achieve estimates equal to the local model parameters. Using the proportion of 0.9 from target data in mixed estimation produced almost the same results as the local model, a major difference being seen only in basal area.

4 Discussion

The results of this study indicated that in ALS-based forest inventory, mixed estimation with a combination of datasets from two inventory areas improved the accuracy of derived plot-level characteristics compared with OLS-based regression models. Additionally, although there were some differences between the various forest characteristics, for instance, plot-level volume, 50 plots was in this study enough to construct local models and achieve as accurate results. To summarize, the number of sample plots to be measured could be lower than in present inventories. The same has been found also in the studies by Junttila et al. (2008) and Maltamo et al. (2009).

In this study the structure of the constructed regression models was fairly simple. For instance, the number of explanatory variables was low. More sophisticated regression models will give better results, but the transferability and combination of models between two different forest areas would be cumbersome and the significance of the independent variables and even the whole regression model could not be guaranteed. There is also an interesting issue about explanatory variables because these models do not include any independent variables which describe the proportional density of laser height percentiles. By contrast, the proportion of vegetation hits and the mean height of laser pulses are explanatory variables which describe forest size and structure in general and are, therefore, suitable for ALS-based regression models.

Using new local models, the automatic variable selection method produced huge errors for all forest characteristics at the beginning of the simulation procedure because the size of the sample was too small for the selection of new, distinguishable independent variables. This principle might in some cases cause serious extrapolation of local regression models. This issue was emphasised because the reference plots were selected randomly, meaning the sample could have included forests in which the growing stock was fairly small, for instance. The automatic variable selection method yielded the best results for plot-level volume, for which they were similar to those obtained by the other methods when the size of the sample was about 40–50 reference plots. For the other forest characteristics, however, it requires at least 80 or even 100 reference plots to obtain at least the same results as the predetermined local models. It is worth emphasising that the automatic variable selection method was an automatic algorithm and was based totally on simulation. Therefore, results should not be too optimistic with the achieved results perhaps improving still if these models were constructed manually.

For some dependent variables, the ALS-based regression models, which were constructed in the Matalansalo forest area, produced systematically underestimated results in the Juuka test area because of the structural differences between these two forests. Hgms are greater in the Matalansalo area than at the Juuka test site, for instance (Tables 1 and 2), i.e., a tree at Matalansalo will usually be taller than one of the same DBH at Juuka because of the allometric differences between trees in different forest areas. This issue is also influenced by biological factors such as temperature. To clarify the systematic differences between the two forest areas, mean height and Hgm (weighted height characteristics) were chosen as dependent variables. The results of both height variables were similar for RMSE, but highlighted minor differences in the case of bias in normal estimation, which was somewhat higher for Hgm than for mean height. The difference in growing stock between the areas will definitely influence the ALS-based regression models, implying that the same independent variables cannot accurately predict forest characteristics in different inventory areas. From this point of view, our results are reasonable because the mixed estimation is not powerful enough to modify and calibrate the regression coefficients relative to local models and new independent variables. However, it should be remembered that at least part of the results might be because we used different laser scanners in two considered inventory areas.

The number of stems was the only exception among the chosen forest characteristics. This was difficult to approximate using mixed estimation or even local models. The final results of N were fairly similar with all estimation methods. The automatic variable selection method produced somewhat better results when the number of reference plots was over 120. One significant issue was that the best estimates of N were achieved when the number of reference plots was smaller than the total number of Juuka reference plots. Moreover, number of stems is the only forest characteristic for which the normal OLS estimation method produces fairly similar results for relative RMSE to mixed estimation or any other estimation method considered here. This might be because the reference range and average N are almost equal in the Juuka and Matalansalo test areas.

The results of this study differed from the study of Korhonen (1993), which applied mixed estimation to tree-level volume. The major difference between these two studies was the number of observations needed for OLS estimation to achieve better results than the mixed estimation. According to Korhonen, the OLS estimator was better when using several hundreds of sample trees but otherwise mixed estimation produced more accurate results. Perhaps the difference is dependent on the scale because the variation in the modelling datasets is higher at the tree-level than plot-level.

In general, using mixed estimation models demands a unique relationship which could be defined as “natural law”. For instance, tree height is usually modelled using tree diameter and this is a fairly stable relationship. Model coefficients certainly change in different forest areas but independent variables remain the same. When estimating tree height using single and fixed models is reasonable, and mixed estimation is one method for generalising this kind of information. In ALS-based forest inventories, the diversity is higher and, therefore, the dependent and independent variables are better chosen locally. In this study, the uncertainty and the heterogeneous of these two datasets will also increase because of the influence of the two different laser scanning systems, for instance.

According to this study and these forest areas, using 50 field plots to construct the local volume model from ALS measurements was enough. Therefore, it would be useful in the future to apply these results in practice because a smaller sample size for a given target inventory area would increase the cost efficiency of the fieldwork. This would improve fieldwork planning and the design of the sampling approach. The plot-level reference data should be appropriate and sufficiently extensive to provide samples of the different forest types and structures in the target inventory area. According to Lefsky et al. (2005), a modified sampling design including a range of structures and ages for the reference plots is needed, but a complete sequence for every forest type is unnecessary. This question of sampling design becomes more complicated when tree species-specific results are needed because the number of reference plots will probably need to be higher to take account of variations in tree species composition.

It would be interesting in the future to gather reference field plots and laser scanner data from more than two inventory areas representing geographically different locations, e.g., in different countries (Breidenbach et al. 2007). The method for combining ALS inventory areas might then be different. In this study, the field reference data sampling principles between the two forest areas were not the same and, therefore, it was not possible to invoke, e.g., the mixed model prediction theory (see Breidenbach et al. 2007). ALS-based measurements for forest inventory purposes can also utilise unique relationships in volume, for instance. In the area-based method, using a proportion of vegetation hits and the mean height of laser pulses for plot or stand might describe the forest total volume. Using this kind of simple relationship between different countries and smaller sub-areas might produce accurate estimates about the growing stock.

Finally, it would be relevant in future studies to note the differences between laser scanning instruments and consider the influence of these devices on the final forest inventory results. Technical issues related to these devices, particularly the footprint diameter of the laser pulse, pulse repetition frequency and the capability of the laser scanner to record and transmit pulse reflections and pulse power, might also crucially affect the results of ALS-based regression models (Hopkinson 2007). In this study, the effect of different laser scanning instruments could have influenced the final results of forest characteristics. However, we had no chance to investigate this issue, and for that reason we would need to scan the same forest area several times using different flying parameters of the laser scanner such as flying altitude and pulse repetition frequency. One interesting approach would be to compare different laser scanning instruments by different manufacturers.

References

Axelsson, P. 2000. DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing 33(B4): 110–117.

Breidenbach, J., McGaughey, R.J., Andersen, H.-E., Kändler, G. & Reutebach, S.E. 2007. A mixed-effects model to estimate stand volume by means of small footprint airborne lidar data for an American and a German study site. In: Rönnholm, P., Hyyppä, H. & Hyyppä, J. (eds.). Proceedings of ISPRS Workshop Laser Scanning 2007 and Silvilaser 2007, September 12–14, 2007, Finland. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVI (Part 3/W52). p. 77–83.

Eid, T., Gobakken T. & Næsset, E. 2004. Comparing stand inventories for large areas based on photo-interpretation and laser scanning by means of cost-plus-loss analyses. Scandinavian Journal of Forest Research 19: 512–523.

Gobakken, T. & Næsset, E. 2007. Assessing effects of laser point density on biophysical stand properties derived from airborne laser scanner data in mature forest. In: Rönnholm, P., Hyyppä, H. & Hyyppä, J. (eds.). Proceedings of ISPRS Workshop Laser Scanning 2007 and Silvilaser 2007, September 12–14, 2007, Finland. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVI (Part 3/W52). p. 150–155.

Hollaus, M., Wagner, W., Eberhöfer & C. & Karel, W. 2006. Accuracy of large-scale canopy heights derived from LIDAR data under operational constraints in a complex alpine environment. ISPRS Journal of Photogrammetry & Remote Sensing 60(5): 323–338.

Holmgren, J. 2004. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. Scandinavian Journal of Forest Research 19: 543–553.

— , Nilsson, M. & Olsson H. 2003. Simulating the effects of lidar scanning angle for estimation of mean tree height and canopy closure. Canadian Journal of Remote Sensing 29(5): 623–632.

Hopkinson, C. 2007. The influence of flying altitude, beam divergence and pulse repetition frequency on laser pulse return intensity and canopy frequency distribution. Canadian Journal of Remote Sensing 33(4): 312–324.

— , Chasmer, L., Lim, K., Treitz, P. & Creed, I. 2006. Towards a universal lidar canopy height indicator. Canadian Journal of Remote Sensing 32(2): 139–152.

Jensen, J., Humes, K., Conner, T., Williams, C. & DeGroot, J. 2006. Estimation of biophysical characteristics for highly variable mixed-conifer stands using small-footprint lidar. Canadian Journal of Forest Research 36: 1129–1138.

Junttila, V., Maltamo, M. & Kauranne, T. 2008. Sparse Bayesian estimation of forest stand characteristics from ALS. Forest Science 54: 543–552.

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

Laasasenaho, J. 1982. Taper curve and volume function for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae 108. 74 p.

Lappi, J., Mehtätalo, L. & Korhonen, K.T. 2006. Generalizing sample tree information. In: Kangas, A. & Maltamo, M. (eds.). Forest inventory. Methodology and applications. Managing Forest Ecosystems. Vol 10. Springer, Dordrecht. p. 85–106.

Lefsky, M., Hudak, A., Cohen, W. & Acker, S.A. 2005. Geographic variability in lidar predictions of forest stand structure in the Pacific Northwest. Remote Sensing of Environment 95: 532–548.

Maltamo, M., Eerikäinen, K., Packalén P. & Hyyppä, J. 2006. Estimation of stem volume using laser scanning-based canopy height metrics. Forestry 79(2): 217–229.

— , Bollandsås, O.M., Næsset, E., Gobakken, T. & Packalén, P. Different sampling strategies for field training plots in ALS inventory. Proceedings of the Silvilaser 2009 conference.

Miller, A. 2002. Subset selection in regression. 2nd ed. Monographs on statistics and applied probability 95. Includes bibliographical references and index. ISBN 1-58488-171-2.

Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment 80: 88–99.

— 2004a. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research 19: 164–179.

— 2004b. Accuracy of forest inventory using airborne laser scanning: Evaluating the first Nordic full-scale operational project. Scandinavian Journal of Forest Research 19: 554–557.

— 2004c. Effects of different flying altitudes on biophysical stand properties estimated from canopy height and density measured with a small-footprint airborne laser scanning. Remote Sensing of Environment 91: 243–255.

— 2005. Assessing sensor effects and effects of leaf-off and leaf-on canopy conditions on biophysical stand properties derived from small-footprint airborne laser data. Remote Sensing of Environment 98: 356–370.

— 2007. Airborne laser scanning as a method in operational forest inventory: Status of accuracy assessments accomplished in Scandinavia. Scandinavian Journal of Forest Research 22: 433–442.

— , Bollandsås, O. & Gobakken, T. 2005. Comparing regression method in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. Remote Sensing of Environment 94: 541–553.

Olsson, H & Næsset, E. 2004. Preface. Scandinavian Journal of Forest Research 19: 481.

Packalén, P. & Maltamo, M. 2006. Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. Forest Science 52(6): 611–622.

R Development Core Team. 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roesch, F.A. 1999. Mixed estimation for a forest survey sample design. USDA Forest Service Southern Research Station 160A Zillicoa Street, P.O. Box 2750, Asheville, NC 28802. Reprinted from the 7999 Proceedings of the section on Statistics and the environment of the American Statistical Association. 6 p.

Theil, H. & Goldberger, A.S. 1961. On pure and mixed statistical estimation in economics. International Economic Review 2: 65–78.

Thomas, V., Treitz, P., McCaughey, J.H. & Morrison, I. 2006. Mapping stand-level forest biophysical variables for a mixed wood boreal forest using lidar: an examination of scanning density. Canadian Journal of Forest Research 36: 34–47.

Uuttera, J., Anttila, P., Suvanto, A. & Maltamo, M. 2006. Yksityismetsien metsävaratiedon keruuseen soveltuvilla kaukokartoitusmenetelmillä estimoitujen puustotunnusten luotettavuus. Metsätieteen aikakauskirja 4/2006: 507–519. (In Finnish).

Veltheim, T. 1987. Pituusmallit männylle, kuuselle ja koivulle. Metsänarvioimistieteen pro gradu -tutkielma. Helsingin yliopisto. 59 p. (In Finnish).

Total of 33 references