Reducing extrapolation bias of area-based k-nearest neighbor predictions by using individual tree crown approaches in areas with high density airborne laser scanning data

Johannes Breidenbach$^{1,2}$, Erik Næsset$^{2}$ & Terje Gobakken$^{2}$

$^1$Norwegian Forest and Landscape Institute, job@skogoglandskap.no
$^2$Norwegian University of Life Sciences

K-nearest neighbor (kNN) approaches are popular statistical methods for predicting forest attributes in airborne laser scanning (ALS) based inventories. Their main upsides are the simplicity to predict multivariate response variables and their freeness of distributional assumptions on the conditional response. One of their largest drawbacks is that predictions outside the range of the reference data inherently result in an under- or overestimation. This property of kNN approaches is known as extrapolation bias and aggravates with an increasing number of neighbors (k) used for the prediction.

This study presents one possibility to reduce extrapolation biases of predictions based on the area-based approach (ABA) by using individual tree crown (ITC) approaches within those specific areas of a low density ALS acquisition where the point density might be sufficiently high for using ITC methods. In the proposed strategy, additional (or artificial) reference plots augmented field measured plots. Artificial plots were created by applying ITC segmentation to a canopy height model derived from high density ALS data. The response variable biomass per hectare was predicted for every segment following a semi-ITC approach. The segment predictions were aggregated on the artificial plot level. The artificial plots were then treated in the same way as the original reference data to make predictions in areas with low density ALS data based on the ABA. It was hereby assumed that the predicted plot level response on the artificial plots is equivalent with the observed plot level response on the original reference data.

The data consisted of 110 reference plots with a smaller data range than the 201 independent validation plots. Considerable extrapolation bias was visible if only the reference plots were used for the prediction. Almost no extrapolation bias was found if the prediction was based on reference plots augmented by artificial plots. The root mean squared error (RMSE) of the biomass predictions based on the reference plots was 39.1%. The RMSE reduced to 29.8% if the reference plots were augmented by artificial plots.

Keywords: Nonparametric regression, kNN, MSN, accuracy, precision, lidar, ALS

1. Introduction

K-nearest neighbor (kNN) approaches are popular statistical methods for predicting forest attributes in airborne laser scanning (ALS) based inventories (e.g., Breidenbach et al. 2010b; Hudak et al. 2009; Packalén and Maltamo 2006). Their main upsides are the simplicity to predict multivariate response variables and their freeness of distributional assumptions on the conditional response. One of their largest drawbacks is that predictions outside the range of the reference data inherently result in an under- or overestimation (McRoberts 2009). This property of kNN approaches is known as extrapolation bias (Magnussen et al. 2010) and aggravates with an increasing number of neighbors (k) used for the prediction. Magnussen et al. (2010) proposed a general, model assisted method to dampen extrapolation biases.

This study presents one possibility to reduce extrapolation biases of predictions based on the area-based approach (ABA) (Næsset 2002) by using individual tree crown (ITC) approaches within those specific areas of a low density ALS acquisition where the point density might be
sufficiently high for using ITC methods. In the proposed strategy, additional (or artificial) reference plots augmented field measured plots. Artificial plots were created by applying ITC segmentation to a canopy height model derived from high density ALS data. The response variable biomass per hectare was predicted for every segment following a semi-ITC approach (Breidenbach et al. 2010a). The segment predictions were aggregated on the artificial plot level which resulted in the “observation” of the response at the artificial plot. The artificial plots were then treated in the same way as the original reference data to make predictions in areas with low density ALS data based on the ABA.

Even if low density ALS data are acquired, areas of high density ALS data exist where additional flight lines are flown perpendicular to the main flying direction for the purpose of calibrating the stripes against each other. Lower flying altitudes may even be considered for these additional flight lines in order to increase point density (Næsset et al. 2006). This makes the proposed approach easily applicable under operational settings. Since the ALS campaign and the field work were carried out independently, an additional ALS campaign was necessary to acquire the high resolution ALS data for some flight strips in this study.

2. Material and methods

2.1 Study area and field data

The study area was located in the municipality of Aurskog-Høland, in southeast Norway (Figure 1). The forest in the area is dominated by Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*). The topography is smooth with heights above sea level between 120 and 390 m.

The field data consisted of 110 reference plots and 201 independent validation plots. On these field plots, diameter at breast height (dbh) and species were recorded for all trees. Tree height was measured for a subsample of trees on every plot to predict the heights of the unmeasured trees. Single tree biomass was derived using species-specific biomass functions based on dbh and height (Marklund 1988).

The reference plots were sectors (i.e., quarters and halves) of 40 large sample plots with an area of 500 and 1000 m², respectively. Following a purposive sampling design, the 40 large sample plots were located under strips of high density ALS data (Figure 1, lower right hand side). The field work in which also tree coordinates were recorded was carried out in 2008. See Breidenbach et al. (2010a) for more details of the reference plot setup. The original number of sectors was 152, each with an area of 250 m². However, in order to better illustrate the strength of using artificial plots, only sectors with a measured biomass between 50 and 170 Mg ha⁻¹ were selected. Thus these selected sectors constituted a sample of 110 reference plots. The reference plots were used to create the artificial plots and for fitting the statistical model using the ABA.

The circular validation plots had an area of 200 m² and stemmed from an operational forest inventory that took place in 2006. The sample plots were distributed according to a stratified systematic design (Figure 1, upper right hand side). Table 1 summarizes characteristics of the reference and validation plots.
Table 1: Characteristics of measured biomass on the sample plots (Mg ha\(^{-1}\))

<table>
<thead>
<tr>
<th>Plot type</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference plots</td>
<td>52.09</td>
<td>92.69</td>
<td>168.30</td>
<td>33.08</td>
</tr>
<tr>
<td>Validation plots</td>
<td>18.47</td>
<td>111.30</td>
<td>314.30</td>
<td>62.64</td>
</tr>
</tbody>
</table>

2.2 Remote sensing data

The whole municipality of Aurskog-Høland was covered with low-density ALS data in summer 2005 in order to provide auxiliary information for an operational forest inventory. The Optech ALTM 3100 sensor was operated from a fixed-wing aircraft with a flying height above ground of approximately 1850 m and a half scanning angle of 15 degrees. The pulse repetition and mirror frequencies were 50 kHz and 31 Hz, respectively. Only the first and last returns per pulse were recorded. While the main flight direction was north-south oriented, two flight strips were flown almost perpendicular to the main block in order to calibrate the other strips against each other (Figure 2). The pulse density under these strips was frequently above 1.5 m\(^{-2}\) whereas the pulse density in general was around 0.7 m\(^{-2}\).
High density ALS data were acquired along five non-overlapping flight strips covering the 40 large sample plots in summer 2006 (Figure 2). The flight direction was east-west oriented and the acquisition settings resulted in flight strips with a swath of 150 m in south-north direction. Up to four returns per pulse and the intensity of the reflected signal were recorded for approximately seven pulses per square meter. The Optech ALTM 3100 sensor was also used for this data acquisition but was operated from a flying height above ground of approximately 800 m and a half scanning angle of 5 degrees. The pulse repetition and mirror frequencies were 100 kHz and 70 Hz, respectively. Both ALS data sets were processed by the data vendor and delivered in UTM coordinates with ellipsoidal heights.

2.3 Creation of artificial plots and area-based predictions

The proposed approach can be subdivided into two phases. In the first phase, artificial plots are generated in the areas with high density ALS data. The second phase is basically the well-known ABA (Næsset 2002). However, the artificial plots are used in addition to the reference sectors to increase the amount of training data.

In the first phase, the positions of the artificial plots were established such that they were
aligned along the flight lines approximately in the middle of the high resolution ALS strips. An artificial plot center was established every 50 m along the flight lines. It should be noted that the numbers given here were chosen for technical and practical reasons. In theory, artificial plots could be anywhere within the area of high resolution ALS data.

Using an inverse distance weighted algorithm and four neighbors, digital surface models (DSMs) with 0.5 m edge length were created from the high density first return ALS data for areas including a 5 m buffer around the artificial plot center. Tree crowns or small clusters of trees were segmented in the inverted DSMs using a watershed algorithm. Segments were assumed to belong to an artificial plot, if their centroid was within a radius of 8.92 m of the artificial plot center. The mean and coefficient of variation of ALS height and intensity, proportion of first returns, proportion of returns lower than 30% of the maximum height (density3) of the ALS returns within the segments as well as the segment area were derived as predictor variables for the semi-ITC approach.

The same strategy as for the artificial plots was followed to segment tree crowns on the reference plots. The segments were intersected with the tree coordinates in order to link the field measurements with the remote sensing data. Due to omission and commission errors of the segmentation algorithm, there can be no, one, or several trees within a segment.

A nonparametric kNN method based on the normalized Euclidean distance (Crookston and Finley 2008) was used to predict the tree properties of crown segments on the artificial plots. Biomass associated with the closest segment on a reference plot was imputed to the target segments on the artificial plots.

In the second phase, the predicted biomass for the segments was aggregated on the artificial plot level. In the case of the reference plots, the measured tree biomass was aggregated. Both data sets were merged to one large table used as reference in the ABA. Targets in terms of kNN were the 201 independent validation plots. The following seven metrics served as predictor variables in the ABA and were derived from the height distribution of the low resolution ALS data: Maximum, mean, standard deviation, coefficient of variation, interquartile distance, kurtosis, first quartile. The metrics were derived for the artificial, reference and validation plots.

As for generating the artificial plots, a kNN approach was used in the ABA. However, the distance metric was based on canonical correlation. With k=1, the method is known as most similar neighbour inference (MSN) (Moeur and Stage 1995). In MSN, the distance between observations, say a target with index \( t \) and a reference with index \( r \), is calculated by multiplying the differences in the explanatory variables (\( X \)), \( d_{tr} = X_t - X_r \), with a weighting matrix. The weighting matrix (\( W \)) is derived using canonical correlation analysis that maximizes the correlation between response (\( Y \)) and explanatory variables by linear transformation \( U_k = \alpha_k Y \) and \( V_k = \gamma_k X; \ k = 1, ..., s \). The variables \( \alpha_k \) and \( \gamma_k \) are the canonical coefficients of the response and the explanatory variables respectively. The weighing matrix is then given by \( W = \Gamma \Lambda^2 \Gamma' \) where \( \Gamma \) = the matrix of all \( \gamma_k \) and \( \Lambda^2 \) = diagonal matrix of the squared \( \lambda_k \). This results in the distance function \( D^2_{tr} = d_{tr} W d_{tr}' \).

The goodness of fit of the MSN models was analysed by plotting observed versus predicted values on the validation plots against each other and by root mean squared errors (RMSE) and mean residuals.

3. Results

The range of the predicted values based on the reference plots alone was between 60 and
150 Mg ha\(^{-1}\). Small observed values were overestimated and large observed values were underestimated (Figure 3, left hand side).

Almost no under- and overestimation of small and large observed values was recognized if artificial plots were used to augment the reference plots (Figure 3, right hand side). The use of artificial plots also resulted in considerably smaller root mean squared errors and mean residuals (Table 2).

![Figure 3: Observed vs. predicted biomass based on the reference plots alone (left) and based on the reference plots augmented by artificial plots (right). The artificial plots were created using the segmented reference plots.](image)

<table>
<thead>
<tr>
<th></th>
<th>RMSE (Mg ha(^{-1}))</th>
<th>RMSE (%)(^{a})</th>
<th>Mean residual (Mg ha(^{-1}))</th>
<th>Mean residual (%)(^{a})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference plots only</td>
<td>43.46</td>
<td>39.06</td>
<td>6.94</td>
<td>6.24</td>
</tr>
<tr>
<td>Reference plots and artificial plots</td>
<td>33.14</td>
<td>29.78</td>
<td>-2.60</td>
<td>-2.34</td>
</tr>
</tbody>
</table>

\(^{a}\) Relative to the mean of the observed values on the validation plots.

4. Discussion and conclusion

The range of the predicted values based on the reference plots alone (between 60 and 150 Mg ha\(^{-1}\)) was smaller than the range of the observed values in the reference plots (between 50 and 170 Mg ha\(^{-1}\)). This was a result of the regression to the mean effect which occurs if the number of neighbors (k) is larger than one (McRoberts 2009). The use of artificial plots levelled out the regression to the mean effect and the extrapolation bias (Magnussen et al. 2010) that occurred due to the smaller data range of the reference plots compared with validation data. Another less obvious effect was that gaps within the data range were reduced by using the artificial plots which also increased the accuracy of the prediction. The distances in the feature space between a target and the closest neighbors increase with increasing gap size. Gaps are
especially large within areas of the data range that is only sparsely covered with observations (McRoberts 2009).

This study showed that semi-ITC approaches can be useful also if low density ALS data are acquired in order to carry out an ABA forest inventory. It would be an easy task to increase the point density on the flight lines that are flown to calibrate the flight lines against each other in a low density ALS acquisition. A good coordination between planning of the flight and field campaigns is, however, essential. Since field and remote sensing data acquisition were carried out independently in this study, high density ALS data over the reference sectors had to be acquired separately. The necessary ALS data do not result in higher costs if two separate flight campaigns can be avoided. The tree coordinates that are necessary for the semi-ITC approach followed here (Breidenbach et al. 2010a) increase the survey costs compared to field measurements sufficient for the ABA. However, the field survey costs for semi-ITC could also be reduced compared to the field data acquisition necessary for the ABA since the measurement of complete sample plots is not required in semi-ITC. In fact, tree coordinates could be measured for single segments or for segments that are organized in polymorph groups. A survey design suitable for semi-ITC could for example be organized in two steps. i) ITC delineation of all areas with high density ALS data. ii) Selection of segments where tree coordinates can easily be measured. III) Measuring of all tree coordinates within the selected segments. The selected segments should cover the whole range of variation in the data. Some kind of randomization scheme during the selection can reduce subjective influence. It is also of importance that the selected segments are not influenced by the fact that they are easy to measure (e.g., by near-by roads). In terms of precision (sometimes denoted as bias), ITC approaches that do not require tree coordinates are not yet in an operational state (Yu et al. 2010).

In operational inventories, predictions need to be tree species-specific (e.g., Breidenbach et al. 2010b; Packalén and Maltamo 2008). This will be the focus in further studies. It would also be of interest to assess how different segmentation algorithms affect the quality of the artificial plots.

Acknowledgements

We wish to thank Mr Knut Marius Hauglin, Ms Nadja Thieme and Mr Vegard Lien at the Norwegian University of Life Sciences for the fieldwork and several GIS analyses. Thanks also appertain Blom Geomatics, Norway, for collection and processing of the airborne laser scanner data.

Parts of this study have been funded by the Research Council of Norway (project #184652/I10: “New Technologies to Optimize the Wood Information Basis for Forest Industries - Developing an Integrated Resource Information System (IRIS)”), the Forest Trust Fund, and the Nordic Forest Research Co-operation Committee. The work is a contribution to the WoodWisdom-Net ERA-NET project.

References


Breidenbach, J., Nothdurft, A., & Kändler, G. (2010b). Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes


