Prediction of stem diameter and biomass at individual tree crown level with advanced machine learning techniques

Salim Malek (1-3), Franco Miglietta (1), Terje Gobakken (2), Erik Næsset (2), Damiano Gianelle (3), Michele Dalponte (3)

Knowledge about the aboveground biomass (AGB) and the diameters at breast height (DBH) distribution can lead to a precise estimation of carbon density and forest structure which can be very important for ecology studies especially for those concerning climate change. In this study, we propose to predict DBH and AGB of individual trees using tree height (H) and crown diameter (CD), and other metrics extracted from airborne laser scanning (ALS) data as input. In the proposed approach, regression methods, such as support vector machine for regression (SVR) and random forests (RF), were used to find a transformation or a transfer function that links the input parameters (H, CD, and other ALS metrics) with the output (DBH and AGB). The developed approach was tested on two datasets collected in southern Norway comprising 3970 and 9467 recorded trees, respectively. The results demonstrate that the developed approach provides better results compared to a state-of-the-art work (based on a linear model with the standard least-squares method) with RMSE equal to 81.4 kg and 92.0 kg, respectively (compared to 94.2 kg and 110.0 kg) for the prediction of AGB, and 5.16 cm and 4.93 cm, respectively (compared to 5.49 cm and 5.30 cm) for DBH.

Keywords: Aboveground Biomass, Diameter at Breast Height, Airborne Laser Scanning (ALS), Remote Sensing (RS), Support Vector Machine for Regression (SVR), Random Forests (RF)

Introduction
Forests are considered a major component of the global carbon cycle. A precise characterization of forest ecosystems in terms of carbon stock density and forest structure is an important key in international efforts to mitigate climate change. Carbon density can be estimated directly from the aboveground biomass (AGB) of trees, while the knowledge about the distribution of diameter at breast height (DBH) can be useful in understanding the forest structure (Slik et al. 2010). Having precise information about the distribution of those two parameters can help to understand the structure and the dynamics of forests. In the past, assessing those characteristics was primarily done with field-based inventory data and sometimes combined with conventional remote sensing (RS) data such as aerial photography and optical satellite images (Dalponte & Coomes 2016, Dalponte et al. 2018). ALS sensors, also referred to as airborne LiDAR (Light Detection And Ranging), are nowadays the most accurate remote sensing technology for monitoring forest carbon (Lefsky et al. 2002, Asner et al. 2012), as they can produce highly detailed 3D point clouds pinpointing locations on branches and the forest floor (Dalponte & Coomes 2016) and they measure surface elevation within a precision of a few centimeters, which offers the potential for studying forests at tree level.

In forest inventories in general, DBH and height (H) are measured and registered in the field in order to predict the AGB using allometric models. AGB is then converted to carbon density for each field-reference tree (Chave et al. 2014, Mensah et al. 2016, Zhang et al. 2016, Peng et al. 2017). With ALS, it became possible to measure the heights of trees in large forests in a short time, which makes it more practical compared to filed-measured methods. However, DBH cannot be measured directly with ALS sensors. Therefore, many studies have been carried out to try to predict DBH from airborne remote sensing data (ARS). The work of Gobakken & Naesset (2004) is considered the first where the DBH and also basal area distribution was predicted by using ALS data at the plot level. Their approach was based on a Weibull density function (Weibull 1951) and regression analysis was used to estimate the corresponding parameters. Recent studies have extracted ARS variables from each individual tree crown (ITC) detected in ARS data (Hauglin et al. 2013, 2014, Jucker et al. 2017, 2018, Mareya et al. 2018, Dalponte et al. 2018). Among this last group of studies, Jucker et al. (2017) proposed new allometric models to predict DBH and AGB based on H and crown diameter (CD) extracted from ARS data. Dalponte et al. (2018) in a recent study successfully linked field-reference DBH and AGB with H and CD extracted from ALS data using linear models. The objective of the current study is to analyze the use of machine learning meth-
ods, such as Support Vector Machines for Regression (SVR) and Random Forest (RF) to predict DBH and AGB at the ITC level using metrics extracted from ALS data. In the first part of the experiment, only H and CD are used as input in order to compare with the work of Dalponte et al. (2018). After that, additional ALS metrics are used as input in order to see their impact on the quality of the prediction.

Materials and methods

Datasets description

In this study, two datasets located in boreal forests of southeastern Norway were used: Hadeland and Våler (Fig. 1). The main tree species in the two areas are Norway spruce (Picea abies [L.] H. Karst), Scots pine (Pinus sylvestris L.), and deciduous tree species, such as birch (Betula spp. L) and aspen (Populus tremula L.). A summary of the field data of the two datasets is presented in Tab. 1 where only the trees chosen for the experiments are showed and grouped according their species.

Hadeland dataset

The field data acquired in the Hadeland district (Fig. 1) were collected on 13 circular sample plots of size 500 m$^2$ and 21 circular sample plots of size 1000 m$^2$ over a total area of about 1300 km$^2$. Within each sample plot, tree species, DBH, and tree coordinates were recorded for all trees with DBH > 3 cm. A total of 3970 trees were recorded. AGB of each tree was calculated using the allometric models of Marklund (1988).

The ALS data were acquired on 9th September 2011 using a Leica ALS70 system operating with a pulse repetition frequency of 180 kHz. The flying altitude was of 1500 m above ground level. Up to four echoes per pulse were recorded and the resulting density of single and first echoes was 2.4 m$^{-2}$.

Methods

In Fig. 2 the architecture of the prediction system used is provided, and in the following paragraphs each step is detailed.

ITC delineation

ITCs were delineated using an approach based on the ALS data and the delineation algorithm of the R package “itcSegment”. The algorithm starts first by finding the local maxima within a rasterized Canopy Height Model (CHM) and designates them as tree tops, and then uses a decision tree method to grow individual crowns around the local maxima. The different steps for this adopted approach are as follows (Dalponte & Coomes 2016):
1. apply a 3 × 3 low-pass filter to the rasterized CHM in order to smooth the surface and reduce the number of local maxima;
2. localization of local maxima by using a circular moving window of variable size. The user provides a minimum and maximum size of the moving window; the window size is adapted according to the central pixel of the window: the size of the window is linearly related to the CHM height. A pixel of the CHM is considered as local maximum if its value is greater than all other values in the window, and if it is greater than some minimum height above ground. The window size is adapted according to the height of the central pixel of the window;
3. labeling each local maximum as an “initial region” around which a tree crown can grow;
4. extraction of the heights of the four neighboring pixels from the CHM and

Tab. 1 - Summary statistics of the field data for all datasets. For the tree height, DBH and AGB the data range and the mean (in brackets) are provided. For the species the number of trees and the percentage (in brackets) are provided.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Species</th>
<th>Hadeland</th>
<th>Våler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree height (m)</td>
<td>Spruce</td>
<td>5.8 - 25.4 (15.8)</td>
<td>3.5 - 33.3 (18.6)</td>
</tr>
<tr>
<td>DBH (cm)</td>
<td>Pine</td>
<td>4.7 - 23.1 (16.0)</td>
<td>4.4 - 26.0 (15.1)</td>
</tr>
<tr>
<td></td>
<td>Broadleaves</td>
<td>5.1 - 22.9 (13.8)</td>
<td>5.8 - 26.3 (15.3)</td>
</tr>
<tr>
<td></td>
<td>Spruce</td>
<td>5.1 - 44.1 (19.3)</td>
<td>4.3 - 50.3 (21.1)</td>
</tr>
<tr>
<td></td>
<td>Pine</td>
<td>4.7 - 51.1 (25.6)</td>
<td>4.0 - 47.9 (20.2)</td>
</tr>
<tr>
<td></td>
<td>Broadleaves</td>
<td>4.0 - 49.5 (14.8)</td>
<td>4.7 - 38.9 (16.2)</td>
</tr>
<tr>
<td>Tree AGB (kg)</td>
<td>Spruce</td>
<td>6.1 - 681.4 (155.1)</td>
<td>4.2 - 1232.9 (216.3)</td>
</tr>
<tr>
<td></td>
<td>Pine</td>
<td>3.1 - 691.0 (214.3)</td>
<td>2.4 - 728.4 (146.3)</td>
</tr>
<tr>
<td></td>
<td>Broadleaves</td>
<td>2.4 - 738.2 (103.2)</td>
<td>4.1 - 680.4 (125.5)</td>
</tr>
<tr>
<td>Species</td>
<td>Spruce</td>
<td>737 (59.7%)</td>
<td>1326 (50.2%)</td>
</tr>
<tr>
<td></td>
<td>Pine</td>
<td>315 (25.5%)</td>
<td>956 (36.2%)</td>
</tr>
<tr>
<td></td>
<td>Broadleaves</td>
<td>182 (14.8%)</td>
<td>361 (13.6%)</td>
</tr>
</tbody>
</table>
Single-tree stem diameter and biomass prediction using ALS data

let us consider the output

vector $y = [y_1, y_2, ..., y_n]$ associated with $X$ and corresponds to the measured DBH and the AGB. The aim of our proposed method is to estimate the relationship between the input vectors $x_i$ and their target values $y_i$.

Support Vector machine for Regression (SVR – Vapnik 1998, Smola & Schölkopf 2004) performs linear regression in a feature space using an epsilon-insensitive loss ($\varepsilon$-SVM). This technique is based on the idea of deducing an estimate $\hat{g}(x)$ of the true but unknown relationship $y = g(x)$ ($i = 1, ..., N$) between the vector of observations $x_i$ and the target value $y_i$ such that: (i) $g(x)$ has, at most, $\varepsilon$ deviation from the desired targets $y_i$; and (ii) it is as smooth as possible. This is performed by mapping the data from the original feature space of dimension $d$ to a higher $d'$-dimensional transformed feature space (kernel space), i.e., $\Phi(x) \in \mathbb{R}^{d'} (d' > d)$, to increase the flatness of the function and, by consequence, to approximate it in a linear way as follows (eqn. 1):

$$g'(x) = \langle x, \omega \rangle + b$$

Therefore, SVR is formulated as minimization of the following cost function (eqn. 2):

$$\psi(\omega, \xi) = \frac{1}{2} \|\omega\|^2 + c \sum_i (\xi_i + \xi_i^*)$$

subject to (eqn. 3):

$$\begin{align*}
    y_i - \langle \omega, \Phi(x_i) \rangle - b & \leq \varepsilon + \xi_i^* \\
    \langle \omega, \Phi(x_i) \rangle + b - y_i & \leq \varepsilon + \xi_i \\
    0 & \leq \xi, \xi^* 
\end{align*}$$

where $\xi$ and $\xi^*$ are the slack variables that measure the deviation of the training sample $x_i$ outside the $\varepsilon$-insensitive zone. $c$ is a parameter of regularization that allows tuning the tradeoff between the flatness of the function $g'(x)$ and the tolerance of deviations larger than $\varepsilon$.

The aforementioned optimization problem can be transformed through a Lagrange function into a dual optimization problem expressed in the original dimensional feature space in order to lead to the following dual prediction model (eqn. 4):

$$g'(x) = \sum_i (\alpha_i - \alpha_i^*) K(x, x_i) + b$$

where $K$ is a kernel function, $U$ is a subset of indices ($i = 1, ..., N$) corresponding to the nonzero Lagrange multipliers $\alpha_i$'s or $\alpha_i^*$'s. The training observations that are associated to nonzero weights are called SVs. The kernel $K(\cdot, \cdot)$ should be chosen such that it satisfies the condition imposed by the Mercer's theorem, such as the Gaussian kernel functions (Vapnik 1998, Smola & Schölkopf 2004). In this study, the SVM implemented

### Table 2 - Metrics extracted from the ALS points.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zmax</td>
<td>Maximum Z</td>
</tr>
<tr>
<td>Zmean</td>
<td>Mean Z</td>
</tr>
<tr>
<td>Zsd</td>
<td>Standard deviation of Z distribution</td>
</tr>
<tr>
<td>Zskew</td>
<td>Skewness of Z distribution</td>
</tr>
<tr>
<td>Zkurt</td>
<td>Kurtosis of Z distribution</td>
</tr>
<tr>
<td>Zentropy</td>
<td>Entropy of Z distribution</td>
</tr>
<tr>
<td>ZP</td>
<td>Ph percentile of height distribution, with P from 5 to 95 at steps of 5</td>
</tr>
<tr>
<td>ZPcumP</td>
<td>Cumulative percentage of points in the Pth layer, with P from 5 to 95 at steps of 5</td>
</tr>
<tr>
<td>Itot</td>
<td>Sum of intensities for each return</td>
</tr>
<tr>
<td>Imax</td>
<td>Maximum intensity</td>
</tr>
<tr>
<td>Imean</td>
<td>Mean intensity</td>
</tr>
<tr>
<td>Isd</td>
<td>Standard deviation of intensity</td>
</tr>
<tr>
<td>Iskew</td>
<td>Skewness of intensity distribution</td>
</tr>
<tr>
<td>Ikurt</td>
<td>Kurtosis of intensity distribution</td>
</tr>
<tr>
<td>IPcumzP</td>
<td>Percentage of intensity returned below the Pth percentile of Z, with P from 5 to 95</td>
</tr>
<tr>
<td>PRth</td>
<td>Percentage of Rth return, with R from 1 to 4</td>
</tr>
</tbody>
</table>

---

iForest 12: 323-329
Malek S et al. - iForest 12: 323-329

in the “kernlab” library of the software R was used.

Random Forest (RF)
The Random Forest (RF) method, which was proposed by Breiman (2001), is considered as one of the most effective machine learning method for predictive analytics. It consists of many decision trees trained on different parts (selected randomly) of the same training set at each node with the goal of reducing the variance. Each node is split using the best among a subset of its corresponding predictors. The strategy of randomness used in RF has been demonstrated to be robust against over-fitting problems (Breiman 2001, Liaw & Wiener 2002).

Given training set $X = \{x_1, x_2, \ldots, x_n\}$, with response $y = \{y_1, y_2, \ldots, y_n\}$ where $N$ is the number of observations which is used in building a forest. Inside the forest a set of $K$ trees $T_1, \ldots, T_K$ is constructed. The output of each tree predicts the outputs for the actual value $\{y_{1t}, y_{2t}, \ldots, y_{nt}\} = T_m(x)$, where $m = 1, \ldots, K$. The final result of the RF is the average of all tree predictions and is calculated as follows (Hannan et al. 2017, Liu et al. 2015, eqn 5):

$$y_{\text{RF}}(x) = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_k(x)$$ (5)

The evaluation of the Random Forest regression is done through the minimization of the mean square error (MSE) in order to select the optimum trees in the forest. In this study, the Random Forest classifier implemented in the “randomForest” library of the software R was used.

Parameter setting
In order to evaluate our methods, each dataset (i.e., Hadeland, and Våler) was divided in two subsets (same as in Dalponte et al. 2018). The first set was used for calibration and the second for the validation phase. The split in the two sets was carried out in order to have similar characteristics in both sets in terms of spatial distribution, and DBH and AGB variation. For the Hadeland dataset, 607 observations were used for calibration and 627 observations were used for validation, while for the Våler dataset, 1398 and 1245, respectively.

Regarding the SVR, the Radial Basis Function (RBF) was used as kernel functions. To compute the best parameter values, we use a cross-validation technique with a number of folds equal to 3. During the cross validation, the parameter of regularization of the SVR $c$ and the width of its kernel function $\gamma$ were varied in the range $[10^{-3}, 10^0]$, respectively. The $c$ value of the insensitive tube was fixed to $10^1$.

For the RF method, we fix the number of trees to grow to 100, while the number of variables which will be randomly sampled as candidates at each split is fixed to 1 when using only CD and H as features and to 25 when using all the features (CD+H+ALS data).

Performance evaluation
In order to evaluate the developed method of prediction and perform a direct comparison with results of the state-of-the-art methods, we adopted the Root Mean Square Error (RMSE) which measures the differences between values predicted by our model and the ground-reference values (eqn 6):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$ (6)

where $N$ is the total number of test observations, $y_i$ is the ground-reference target value and $\hat{y}_i$ is the predicted value of the developed regression method (ARS-predicted). Both $y_i$ and $\hat{y}_i$ correspond to the $i^{th}$ test observation $x_i$.

We also adopted the percentage improvement ratio measure (PIR) in order to evaluate the level of improvement of our method compared to those of the state-of-the-art (SOA). It is formulated as follows (eqn 7):

$$PIR = \frac{RMSE_{\text{SOA}} - RMSE_{\text{ARS}}}{RMSE_{\text{ARS}}} \times 100$$ (7)

Results and discussion
In the first part of the analysis, only H and CD were used as input in order to compare the obtained results with those reported in Dalponte et al. (2018). Tab. 3 reports the results obtained with the set of metrics H+CD in terms of RMSE and PIR.

From Tab. 3, it can be seen that the proposed method using SVR and RF provide better results in terms of RMSE compared to the state-of-the-art method (Dalponte et al. 2018) for both DBH and AGB predictions. SVR shows substantial improvements while the RF results are located between those of SVR and the state-of-the-art method (Dalponte et al. 2018).

In greater detail, considering the AGB prediction, RMSE for the Hadeland dataset was 81.43 kg with SVR and 84.35 kg with RF, while it was 92.04 kg with SVR and 95.46 kg with RF in the Våler dataset. The improvement is significant compared to the results obtained by the state-of-the-art method (94.19 kg and 109.99 kg for Hadeland and Våler, respectively – Dalponte et al. 2018) with a maximum PIR equal to 13.55% for Våler dataset and 16.32% for Hadeland dataset. Regarding the DBH prediction, the RMSE by using SVR was 5.16 cm for Hadeland dataset and 4.93 cm for Våler dataset. Those results show also significant improvements compared to the state-of-the-art method (Dalponte et al. 2018) where RMSE was 5.49 cm for Hadeland and 5.30 cm for Våler. However, RF did not show considerable improvement and provided results close to those of the reference method (5.42 cm for Hadeland and 5.19 cm for Våler). In term of PIR, the maximum improvement was equal to 6.98% and 16.32% for Hadeland and Våler datasets, respectively.

To see visually the quality of the results, we show in Fig. 3 the field-reference DBH vs. ARS predicted DBH, and in Fig. 4 the field predicted AGB vs. ARS predicted AGB. The ARS predicted values are close to the regression line only until a certain value, while afterwards the bias is increasing. For example, if we consider the worst case corresponding to the prediction of DBH for the Hadeland dataset, the predicted values were close to the regression line when the DBH was inferior to 35 cm, while above this value, the predictions were giving values around 30 cm. This can be explained by the fact that among the 607 observations presented in the training, there were only 23 observations which have value higher than 35 cm which represents only 3.8% of the total training data. Moreover, the relationship between the tree DBH (and AGB) and its height is not linear as, after a certain age, trees stop to grow in height and they grow mainly in DBH. Thus, models that are based on ALS data have problems in modelling the DBH and AGB of old trees. Additionally, the traditional ground-based inventories require a big effort in terms of time and they may not be the best choice to provide a balanced dataset regarding the tree height, stem diameter and other features. It can be more efficient to select remotely the best samples (trees) to be later annotated in the field by experts. In terms of RMSE, our method showed good results since most of the predicted values fall close to the regression line except a few observations, but those few ones occupy a large range (between 35 and 50 cm which represents 30% of the total occupied range).
In order to improve the results, a set of ALS metrics were used together with the previous metrics (H and CD). The obtained results (Tab. 4) show that the ALS metrics help to improve the results, especially when using the RF method. For example, regarding the prediction of AGB, the RMSE reached a value less than 80 for the Hadeland dataset (78.5 with SVR and 76.0 by RF).

Fig. 3 - Field-reference DBH vs. ARS-predicted DBH for the two datasets using H+CD as input. In red the 1:1 line.

Fig. 4 - Field-reference AGB vs. ARS-predicted AGB for the two datasets using H+CD as input. In red the 1:1 line.

Fig. 5 - Field-reference DBH vs. ARS-predicted DBH using H+CD+ALS data. In red the 1:1 line.

Fig. 6 - Field-reference AGB vs. ARS-predicted AGB using H+CD+ALS data. In red the 1:1 line.
which represents an improvement of 16.62% by using SVR and 19.34% by using RF. For the Våler dataset RMSM was lower than 90 (88.9 with RF) with a PIR equals to 19.15. The same remarks can be given for the DBH prediction part where the improvement was significant and the PIR was 12.75% and 8.11% for Hadeland and Våler datasets, respectively.

In Fig. 5 and Fig. 6, we present the field-reference DBH vs. ARS-predicted DBH and the field-predicted AGB vs. ARS-predicted AGB, respectively, using ALS metrics along with H and CD. From the different graphs, a slight improvement can be noticed compared to the previous ones (using only H and CD metrics in Fig. 3 and Fig. 4). However, the problem of inaccurate prediction of high values of AGB and DBH appeared also in this part of the experiments. We think that such problem can be investigated in a separate work in order to explore better the possible challenges that it may present.

Conclusion

In this work, we proposed an approach to predict DBH and AGB of trees from remote sensing data by using SVR and RF regression methods. The developed approach was tested on two datasets. On the first part of the experiments, the metrics H and CD were used in order to predict DBH and AGB. The obtained results were promising and the improvements were noticeable, especially in terms of RMSE. In order to improve the results, we proposed to introduce ALS metrics and use them together with H and CD. The obtained results were encouraging, especially with RF where the improvement was large. However, our method was not able to predict the isolated samples with highest values of AGB and DBH. This was probably due to the fact that the relationship among the height of a tree (and thus of the majority of the ALS metrics) and the DBH and AGB was saturating, and also to the fact that there was a limited number of training observations for the larger trees.

Finally, in order to improve the quality of the results and to get better predictions for old trees with large values of DBH and AGB, we think it can be more advantageous to use techniques that preprocess the data in order to yield a balance for their distribution over all the scale of the different metrics.

Acknowledgements

This work was supported by the HyperBio project (project 244599) financed by the BIONR program of the Research Council of Norway and TerraTec AS, Norway.

References


Malek S et al. - iForest 12: 323-329

Tab. 4 - Accuracy statistics for DBH and AGB predictions using H+CD+ALS metrics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>RMSE (cm)</th>
<th>PIR (%)</th>
<th>RMSE (kg)</th>
<th>PIR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadeland</td>
<td>RF</td>
<td>4.79</td>
<td>12.75</td>
<td>75.97</td>
<td>19.34</td>
</tr>
<tr>
<td></td>
<td>SVR</td>
<td>4.93</td>
<td>10.20</td>
<td>78.54</td>
<td>16.62</td>
</tr>
<tr>
<td>Våler</td>
<td>RF</td>
<td>4.88</td>
<td>7.92</td>
<td>88.93</td>
<td>19.15</td>
</tr>
<tr>
<td></td>
<td>SVR</td>
<td>4.87</td>
<td>8.11</td>
<td>91.15</td>
<td>17.13</td>
</tr>
</tbody>
</table>

Single-tree stem diameter and biomass prediction using ALS data

50-60. - doi: 10.1111/j.1466-8238.2009.00489.x