Title: A Growth-Model Driven Technique for Tree Stem Diameter Estimation by using Airborne LiDAR Data

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2018

Author(s): Claudia Paris, Lorenzo Bruzzone,

Volume:

Page(s):

DOI: 10.1109/TGRS.2018.2852364

A Growth-Model Driven Technique for Tree Stem Diameter Estimation by using Airborne LiDAR Data

Claudia Paris, *Member, IEEE*, Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Diameter at breast height (DBH) is one of the most important tree parameter for forest inventory. In this work, we present a novel method for the adaptive and the accurate DBH estimation of trees characterized by small and large stems. The method automatically discriminates among different tree growth-models by means of a data-driven technique based on a clustering procedure. First, the method detects young trees belonging to the lowest forest layer by simply considering the vertical structure of the forest. Then, different clusters of mature trees that are expected to share the same growth-model are identified by analyzing the environmental factors that can affect the stem expansion (e.g., topography, forest density). For each detected growth-model cluster, a tailored regression analysis is performed to obtain accurate DBH estimation results. Experiments have been carried out in an homogeneous coniferous forest located in the Alpine mountainous scenario characterized by a complex topography and a wide range of soil fertility. The method was tested on two datasets characterized by different Light detection and ranging (LiDAR) point densities and different forest properties. The results obtained demonstrate the effectiveness of having multiple regression models adapted to the different growth-models.**

***Index Terms*—Forestry, Light detection and ranging (LiDAR), Remote Sensing, Tree Stem Attributes, Forest Parameters, Regression Analysis.**

I. Introduction

Diameter at breast height (DBH), together with the height of the tree, is one of the most relevant tree parameter for the characterization of the forest structure. At single tree level, DBH is fundamental to estimate parameters such as tree stem volume, basal area or carbon storage. At plot level, an accurate prediction of the DBH distribution (number of stems per ha across a set of DBH classes) is necessary to characterize the structure, the growth and the economic value of the forest stand. Although many studies in the literature present area-based approaches to estimate the DBH distribution [1]–[3], to accurately characterize the forest, DBHs should be predicted at individual tree level. While the height of the trees can be directly retrieved by subtracting the Digital Terrain Model (DTM) from the LiDAR measurements, DBHs are not directly measured and should be extracted from the data by means of regression models.

To address this issue, several works estimate DBH considering the crown geometry measured by the LiDAR data [4]–[7]. Besides the correlation between the shape of the crown and the DBH, these parameters are not sufficient to accurately model the variability of the DBH especially in heterogeneous forest scenarios [8]. To obtain a more detailed characterization of the canopy structure, some studies extract LiDAR point cloud metrics calculated from the volume of the segmented crowns [9]–[11]. In [9] the variables extracted from the multireturn LiDAR data represent the distribution of the laser pulses within the crown, thus modeling the height of the tree, the horizontal and vertical shape of the crown, the internal structure of the crown and the forest species. In [12], the authors performed stem attribute prediction on continuous waveform LiDAR datasets acquired in leaf-off/leaf-on conditions. The crown area, crown volume, tree height and tree crown height were used to estimate the DBH. The trees were a-priori divided into species types (coniferous and deciduous) and foliage conditions (leaf-on/leaf-off seasons). While good estimation results were obtained for conifers regardless of the season, poor accuracies were achieved on broadleaves specially in leaf-on condition, when the detection and segmentation of the deciduous crowns are more complicated. In [7], the authors extended the stem attributes analysis by comparing on the same datasets four regression models. The models were applied to each group of trees separately to compare the performance of the different methods. Support Vector Regression (SVR) yielded the best DBH prediction regardless of foliage condition or tree species.

Despite the possibility of accurately representing the crown strcuture, the characterization of the tree shape is not sufficient to obtain accurate DBH estimates (see [8], [13]). To addrees this issue, recent studies explored the possibility of extracting variables from the circular area around the tree to model the immediate forest neighbourhood [13]–[15]. Indeed, the stand density plays a fundamental role in the expansion of the DBH in terms of availability of water and sunlight. In [13], the authors introduced a competition index to evaluate the influence of the surrounding trees (i.e., competitors) on the DBH growth in old-aged forest. The height and the distance of the competitors are evaluated to quantitatively estimate their pressure on the growth of the considered tree. The growth competition index analysis has been presented in [16], where the authors employed a bitemporal LiDAR data acquisition to monitor and improve the understating on the individual tree growth.

In [17], the authors present a method to perform the accurate reconstruction of DBHs of free-standing or partly occluded trees considering very high density LiDAR data (i.e., > 50 pts/m^2). The method automatically extracts the DBH from LiDAR data by using a skeleton measurement technique. Although the stem extraction allows an accurate estimation of the tree DBH, the method cannot be applied to dense forest

scenario and requires very high density LiDAR data.

Much effort has focused on spatial statistical models [18]–[20], which take advantage from the spatial correlation for improving the accuracy of the predicted DBHs. In [21] the authors demonstrate that accounting for spatial correlation of LiDAR model errors can improve the accuracy of the parameters retrieved. Indeed, it is reasonable to assume that the dendrometric variables of trees growing in the same forest area are more similar with respect to trees belonging to separate forest stands. In [18], the authors compared different statistical regression models and found that the linear mixed-effects model (LME) allows a more accurate DBH estimation with respect to the geographically weighted regression (GWR), the ordinary least squares (OLS) and the generalized least squares (GLS) methods with a non-null correlation structure. Although LME does not directly incorporate the spatial information, the inclusion of random effects permits to focus on each individual tree by taking into account the lack of independence among trees belonging to same forest stand. These models improve the DBH estimation accuracy, however, in mountainous forest areas the properties of forest stands are not uniform due to the complex terrain morphology. Thus, the spatial distribution of the trees is not homogeneous and the terrain properties rapidly change when considering close trees due to the steep slopes.

From the analysis of the literature it turns out that even though height and DBH are correlated within the same forest area, there is a high variability in their relationship due to the terrain properties (e.g., fertility, soil class, altitude, slope) and the forest properties (e.g., stem density, management history of the stand). Accordingly, regression models based only on tree variables achieve good performances on medium size DBHs but are highly sensitive to the outlayers, thus causing poor model fitting at the tails of the distribution. In particular, these models tend to overestimate small DBHs and underestimate large DBHs. While the underestimation of the DBH strongly affects the tree (or stand) volume estimates, the overestimation of the small DBH is problematic for predicting the future growth of the stand plot. Moreover, recent studies have proposed approaches for precise forest mapping, by estimating the forest age [22], monitoring the carbon dynamic [23] or accurately modeling the forest structure via synthetic models [24]. In this context, the precise characterization of the environmental conditions can be employed to improve the DBH estimation. In this paper we propose a data-driven process that dynamically detects classes of trees characterized by different DBH growth-models. Instead of considering the spatial correlation, the proposed approach takes into account the vertical forest structure and all the major environmental factors (which can be computed from LiDAR data) that affect the DBH growth. The aim of the proposed approach is to detect classes of trees characterized by different growth-models. Indeed, trees belonging to the same stand plot but affected by different environment conditions (e.g., stand density, terrain slopes) present different growth-models. In contrast, trees located in different forest areas but sharing similar forest conditions are characterized by comparable stem expansion rates. The environmental variables are used in the framework of a clustering technique to aggregate trees sharing the same growth-model in the same class. Moreover, the variables that mainly discriminate different growth-model classes are identified by a feature ranking method. Finally, a regression model specific for each class is defined and adopted, thus increasing the estimation accuracy. The main contributions of this work are: (i) to employ the LiDAR data to accurately represent both the crown structure and the local forest environment of a tree, (ii) to dynamically detect classes of trees sharing similar growth-models in a considered forest scenario and, (iii) to estimate the DBH of different growth-model classes by using tailored regression analyses. The proposed method has been tested in two homogeneous coniferous forests located in the Southern Italian Alps, a complex mountainous scenario characterized by a wide range of soil classes and DBHs. Results obtained demonstrate the effectiveness of having multiple regression models tailored to each growth-model class. This allows a sharp improvement in the estimation accuracy of tree attributes related to both small and large DBHs.

The paper is organized into five sections. Section II illustrates the architecture of the proposed DBH estimation approach and describes all the phases of the method in detail. Section III presents the two LiDAR datasets used in the experimental analysis. Section IV presents the obtained experimental results. Finally, Section V draws the conclusion of the work.

II. Proposed Estimation Method

In this paper, we present a novel method for accurately estimating tree DBHs by using the 3-D LiDAR data to account for both the tree crown structure and the local forest environment. The proposed method (PM) is separated into four main phases: (i) preprocessing, (ii) extraction of variables potentially correlated to the stem growth, (iii) data-driven clustering of the trees belonging to different growth-model classes, and (iv) data-driven DBH estimation. Fig. 1 shows the architecture of the proposed DBH estimation technique. In the following we describe in detail each phase of the PM.

A. Preprocessing

The *preprocessing* phase seeks to delineate the individual tree crowns and their surrounding local forest environment from the LiDAR data. The method used to perform the individual tree crown segmentation is described in detail in [25]. First, the DTM is subtracted from the LiDAR data to obtain the relative height value of each laser point with respect to the ground. To this end, the DTM provided by the company that acquired the LiDAR data was used. The nominal precision of the DTM is of about 30 cm for altimetry and of about 1 m for planimetry. Then, the normalized LiDAR point cloud is rasterized to generate the Canopy Height Model (CHM) in order to identify most of the trees present in the scene by using a Level Set Method (LSM). To recover possible missed crowns, the method further analyzes the forest area around each tree-top directly in the LiDAR data. Finally, the trees are delineated in the 3D space by exploiting the geometrical shape of their crowns.

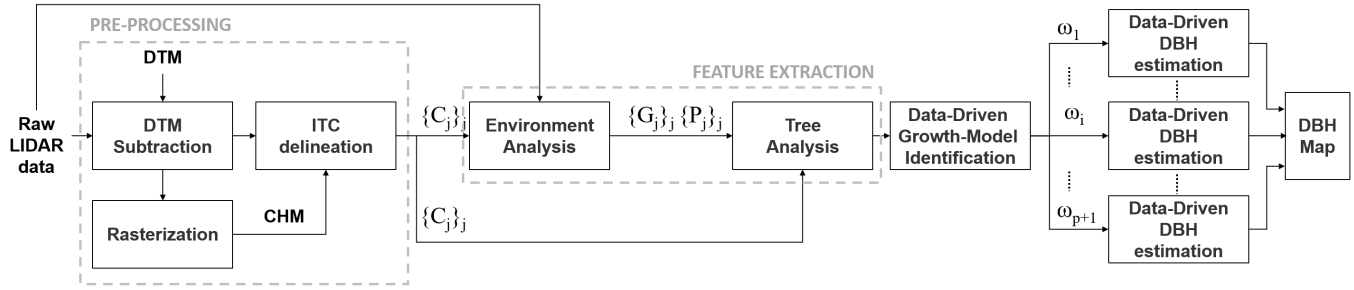
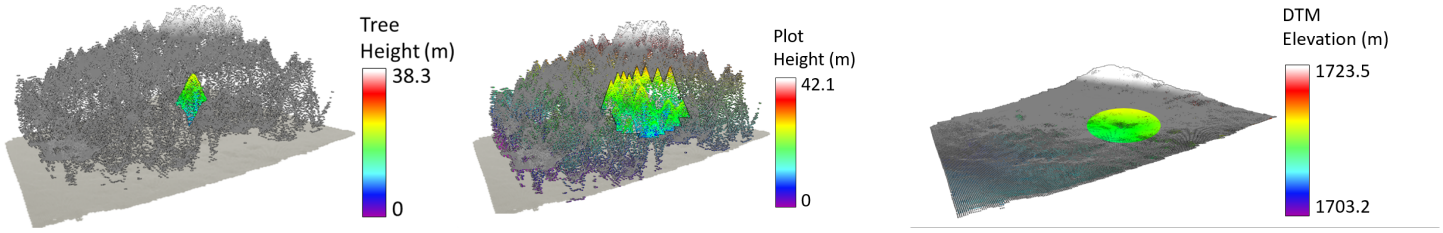


Figure 1. Architecture of the proposed method based on a data-driven clustering of trees belonging to different growth-models for a multi-regression based stem DBH estimation.



(a) (b) (c)

Figure 2. Visual representation of the variables extracted to model the growth of the tree stems in terms of: (a) structure of the crown $\mathbf{x}^{\text{Tree}}$, (b) forest stand $\mathbf{x}^{\text{Plot}}$, (c) topography $\mathbf{x}^{\text{Dtm}}$.

In addition to the segmented crown, around each tree-top $\mathbf{t}_j = \{x_j^t, y_j^t, z_j^t\}$ the method extracts from the LiDAR data both the forest stand and the ground topography within a given radius r_{env}. The gaps present in the terrain are interpolated by using the two-dimensional Laplacian elliptic partial differential equation (PDE) that guarantees accurate results even in cases where many holes are present [26]. At the end of this step, for each tree-top $\{\mathbf{t}_j\}_j$ we obtain the related segmented LiDAR point clouds representing the crowns $\{C_j\}_j$, the forest plot $\{P_j\}_j$ and the 3D ground topography $\{G_j\}_j$ (see Fig. 2).

B. Feature Extraction

The PM aims to identify clusters of trees belonging to the same growth-model class directly from the data. Thus, we need to properly model both the vertical forest structure and the environmental variables (in terms of stand density and topography) which may affect the stem growth. Fig. 2 presents a visual representation of the main factors that affect the DBH. Note that in our implementation, we do not consider the species information in the DBH estimation as from an operational view point it is not always feasible to assume classification maps available over large forest areas. For each detected tree $\mathbf{t}_j$ we extract variables in order to consider: (i) the structure of the tree crown $\mathbf{x}^{\text{Tree}}$, (ii) the local and global stand densities $\mathbf{x}^{\text{Plot}}$, and (iii) the topography $\mathbf{x}^{\text{Dtm}}$. Accordingly, the considered feature vector $\mathbf{x}_j \in \mathcal{R}^d$ associated to the tree $\mathbf{t}_j$ is defined as follows:

$$\mathbf{x}_j = (\mathbf{x}_j^{\text{Dtm}} \cup \mathbf{x}_j^{\text{Plot}} \cup \mathbf{x}_j^{\text{Tree}}) \tag{1}$$

Tab. I reports the features extracted from $\{C_j\}_j$ to represent the crown structure $\mathbf{x}^{\text{Tree}}$. Because of the availability of the multireturn LiDAR data, we extract both a set variables which represent the statistical distribution of returns within the crown (e.g., H^1_{max} is the maximum height value measured among all the first returns collected within the same crown) and a set of variables being able to model the crown geometry (e.g., c_a, r_1, r_2). The internal structure of the tree crowns is accurately characterized by the differences between the mean height values of different returns (e.g., H^1_{av} - H^2_{av}), whereas the vertical profile is represented by the height percentiles H_p and the difference between the maximum height of the 1^{st} return and the minimum height of the 3^{rd} return, H^1_{max}-H^3_{max}. It is worth noting that this set of variables has been widely used for modelling the tree structure for both stem volume estimation [27]–[29] and forest species classification [30], [31].

To consider the forest density, we extract features from $\{P_j\}_j$ representing both the local and the global stand density (Tab. II). While the local features allows us to account for the immediate neighbourhood around the tree, the global stand density models the forest environment (e.g., forest stand density). The Local Canopy Cover (LCC) is calculated as the ratio between canopy cover and ground area within a radius larger than 1m with respect to the crown radius (i.e., $r_1 + 1$m). For the stand density we calculate the Global Canopy Cover (GCC) as the ratio between canopy cover and ground area around the tree within a given radius r_{env}. In the same area we calculate the ratio p_2/p_1 between the number of 2^{nd} and 1^{st} returns to evaluate the vertical density of the forest stand. Moreover, we extract the Canopy Reflection Sum (CRS) index, which has been proved to be an effective metric to model the forest density [32], [33].

Finally, a proper representation of the topography around each tree is obtained by extracting from $\{G_j\}_j$ the variables

Table I
SET OF VARIABLES USED FOR MODELING THE CROWN STRUCTURE.

Variable	Description
H^1_{max}	Maximum height value among the 1^{st} returns
H^2_{max}	Maximum height value among the 2^{nd} returns
H^3_{max}	Maximum height value among the 3^{rd} returns
H^4_{max}	Maximum height value among the 4^{th} returns
H^1_{range}	Height range of the 1^{st} returns
H^2_{range}	Height range of the 2^{nd} returns
H^3_{range}	Height range of the 3^{rd} returns
H^4_{range}	Height range of the 4^{th} returns
H^1_{av}	Average height value among the 1^{st} returns
H^2_{av}	Average height value among the 2^{nd} returns
H^3_{av}	Average height value among the 3^{rd} returns
H^4_{av}	Average height value among the 4^{th} returns
H^1_{var}	Height variance of the 1^{st} returns
H^2_{var}	Height variance of the 2^{nd} returns
H^3_{var}	Height variance of the 3^{rd} returns
H^4_{var}	Height variance of the 4^{th} returns
H^1_{skw}	Height skewness of the 1^{st} returns
H^2_{skw}	Height skewness of the 2^{nd} returns
H^1_{kurt}	Height kurtosis of the 1^{st} returns
H^2_{kurt}	Height kurtosis of the 2^{nd} returns
H^1_{max}-H^3_{max}	Max height 1^{st} - Min height 3^{rd}
H^1_{av} - H^2_{av}	Average height 1^{st} - Average height 2^{nd}
H^1_{av} - H^3_{av}	Average height 1^{st} - Average height 3^{rd}
H^1_{av} - H^4_{av}	Average height 1^{st} - Average height 4^{th}
H^2_{av} - H^3_{av}	Average height 2^{nd} - Average height 3^{rd}
H^2_{av} - H^4_{av}	Average height 2^{nd} - Average height 4^{th}
H^3_{av} - H^4_{av}	Average height 3^{rd} - Average height 4^{th}
H_p	pth height percentile, with p ={25,50,75,90,95}
c_a	Crown area
r_1	Radius of the circle circumscribed to the crown
r_2	Radius of the ellipse circumscribed to the crown

Table II
SET OF VARIABLES USED FOR MODELING THE FOREST DENSITY.

Variable	Description
LCC	Local Canopy Cover
GCC	Global Canopy Cover
p_2/p_1	Ratio of 2^{nd} and 1^{st} returns
CRS	Sum of intensity

presented in Tab. III. It is worth mentioning that even though the topography is often not considered, the terrain properties play a fundamental role in the DBH growth (e.g., water drainage, soil fertility, sunlight exposure). Let us focus the attention on the tree crown C_j and let us define the partial derivatives of $z = G_j(x, y)$ along the orthogonal directions x and y at the distance r_{env} in the horizontal plane and let us

Table III
SET OF VARIABLES USED FOR MODELING THE TOPOGRAPHY.

Variable	Description
S_{west}	Slope between (x^t_j, y^t_j) and $(x^t_j - r_{env}, y^t_j)$
S_{east}	Slope between (x^t_j, y^t_j) and $(x^t_j + r_{env}, y^t_j)$
S_{south}	Slope between (x^t_j, y^t_j) and $(x^t_j, y^t_j - r_{env})$
S_{nord}	Slope between (x^t_j, y^t_j) and $(x^t_j, y^t_j + r_{env})$
γ	Aspect (degrees clockwise from north)
φ	Profile Curvature: direction of max slope
ϕ	Plan Curvature: transverse to the max slope
w	Wetness Index
A_{min}	Minimum Altitude
A_{max}	Maximum Altitude
A_{av}	Average Altitude

assume that the second-order partial derivative exist:

$$g_x = \frac{\partial z}{\partial x},\ g_y = \frac{\partial z}{\partial y},\ g_{xx} = \frac{\partial^2 z}{\partial x^2},\ g_{yy} = \frac{\partial^2 z}{\partial y^2},\ g_{xy} = \frac{\partial^2 z}{\partial x \partial y} \tag{2}$$

The standard topographic metrics usually employed in the DBH estimation (which are slope, altitude and aspect within a given radius r_{env} around the tree) are extracted, where the sun exposure γ has been calculated as follows:

$$\gamma = 180 - arctan\left(\frac{g_y}{g_x}\right) + 90\frac{g_x}{|g_x|} \tag{3}$$

Moreover, differently from the literature, an accurate characterization of the terrain morphology is performed by using variables introduced in the hydrological modeling of a terrain [34], [35]. This is done to obtain a proper characterization of hydrological topographic attributes that allow a better estimation of the relation between DBH and tree height. Let us define:

$$p = (g_x^2 + g_y^2) \text{ and } q = (g_x^2 + g_y^2 + 1) \tag{4}$$

The considered terrain variable are the profile curvature (φ), the plan curvature (ω) and the wetness index (w), defined as follows:

$$\varphi = \frac{g_{xx}g_x^2 + 2g_{xy}g_xg_y + g_yg_y^2}{pq^{3/2}} \tag{5}$$

$$\omega = \frac{g_{xx}g_x^2 - 2g_{xy}g_xg_y + g_yg_x^2}{q^{3/2}} \tag{6}$$

$$w = ln\left(\frac{A_s}{\sqrt{g_x^2 + g_y^2}}\right) \tag{7}$$

where A_s is the circular area around the tree having a given radius r_{env}.

C. Data-Driven Clustering of Trees Belonging to Different Growth-Models

The third phase, *data-driven clustering of trees belonging to different growth-models*, seeks to automatically detect classes

of trees characterized by different growth rates. Let us assume to have Ω growth-models classes. The first growth-model we aim to detect is the one associated to the young trees. While the growth of mature trees is significantly affected by the environment, young trees are characterized by almost linear DBH/height growth rate. Moreover, young trees are characterized by relative low height values with respect to the structure of the forest, since they usually belong to the lowest forest layer. Accordingly, the detection of the young trees is performed by analyzing the vertical structure of the forest considering the variables $\mathbf{x}^{\text{Str}} = [\text{H}^1_{max}, \text{H}^1_{max} - \text{H}^3_{max}, \text{H}^1_{av}, \text{H}^1_{range}]$ derived from the tree feature space $\mathbf{x}^{\text{Tree}}$. In the considered implementation, a data-driven approach is employed to automatically detect the forest layers by using an unsupervised clustering algorithm. Thus, the clustering is completely driven by the employed variables (i.e., the considered feature space). Here, for simplicity we use the k-mean clustering algorithm. However, any other clustering technique can be considered. Let n be the number of forest layers $\{L_i\}_{i=1}^n$ of the considered scenario. The k-mean clustering algorithm initializes randomly the set of centroids and associates each features vector $\mathbf{x}_j^{\text{Str}}$ to the closest centroid considering the euclidean distance metric. By iteratively adjusting the centroid position with respect to the center of the obtained clusters $\{\boldsymbol{\mu}_i\}_{i=1}^n$, the algorithm converges by minimizing the intra-cluster variance in the feature space defined as:

$$\sum_{i=1}^{n} \sum_{\mathbf{x}^{\text{Str}} \in L_i} ||\mathbf{x}^{\text{Str}} - \boldsymbol{\mu}_i||^2 \tag{8}$$

where $\boldsymbol{\mu}_i$ is the cluster centroid of L_i. Since the vertical forest structure is strongly dependent on the type and the age of the forest, in the PM we automatically estimate the number of layers n. This condition allows us to adapt the estimation of the forest structure (and thus, of the young trees) to the specific properties of the considered forest. To this end, the PM employes the Calinski Harabasz (CH) Index, which has been widely used to automatically detect the number of clusters in an unsupervised way [36]–[38]. The cluster validity is evaluated for different values of n by considering the average between- and within- cluster sum of squares, i.e.,

$$\text{CH} = \left[\frac{\sum_{i=1}^{n} n_i ||\boldsymbol{\mu}_i - \boldsymbol{\mu}||^2}{n-1} \right] \Big/ \left[\frac{\sum_{i=1}^{n} \sum_{\mathbf{x}^{\text{Str}} \in L_i} ||\mathbf{x}^{\text{Str}} - \boldsymbol{\mu}_i||^2}{n_t - n} \right] \tag{9}$$

where n_t is the total number of samples, n_i is the number of samples of the ith cluster, $\boldsymbol{\mu}_i$ is the cluster centroid of L_i and $\boldsymbol{\mu}$ is the centroid of the entire dataset. Let ω_1 be the growth-model class associated to the young trees and $\Omega_p = \{\Omega - \omega_1\}$ be the remaining set of classes. Let L_1 be the lowest layer of the forest detected by the k-mean clustering algorithm. The jth crown C_j is classified according to the following rule:

$$\begin{aligned} C_j \in \omega_1 \quad &\text{if} \quad \mathbf{x}^{\text{Str}} \in L_1 \\ C_j \in \Omega_p \quad &\text{if} \quad \mathbf{x}^{\text{Str}} \in \{L_2, .., L_n\} \end{aligned} \tag{10}$$

Although the correct detection of the young trees is important to improve the DBH estimation of the small tree DBH, the main challenge is represented by the DBH estimation of mature trees. Thus, for a given tree height, the DBH considerably varies depending on the age of the tree. However, the stem growth of these trees is strictly related to the environmental condition. Therefore, by considering the feature space $\mathbf{x}^{\text{Env}} = (\mathbf{x}^{\text{Dtm}} \cup \mathbf{x}^{\text{Plot}})$, we can identify classes of mature trees characterized by different growth-models. Accordingly, the data-driven approach automatically determines the remaining p growth-models $\{\omega_i\}_{i=1}^p$ in the feature space $\mathbf{x}^{\text{Env}}$. Note that the clustering analysis is completely independent on the geographic location. Thus, trees widely separated in space can be associated to the same growth-model class given the similarity of the environmental conditions by minimizing the cost function:

$$\sum_{i=1}^{p} \sum_{\mathbf{x}^{\text{Env}} \in \omega_i} ||\mathbf{x}^{Env} - \boldsymbol{\mu}_i||^2 \tag{11}$$

where $\boldsymbol{\mu}_i$ is the cluster centroid of ω_i. Similarly to the previous case, the number of remaining growth-models p is estimated directly from the data by using the CH index.

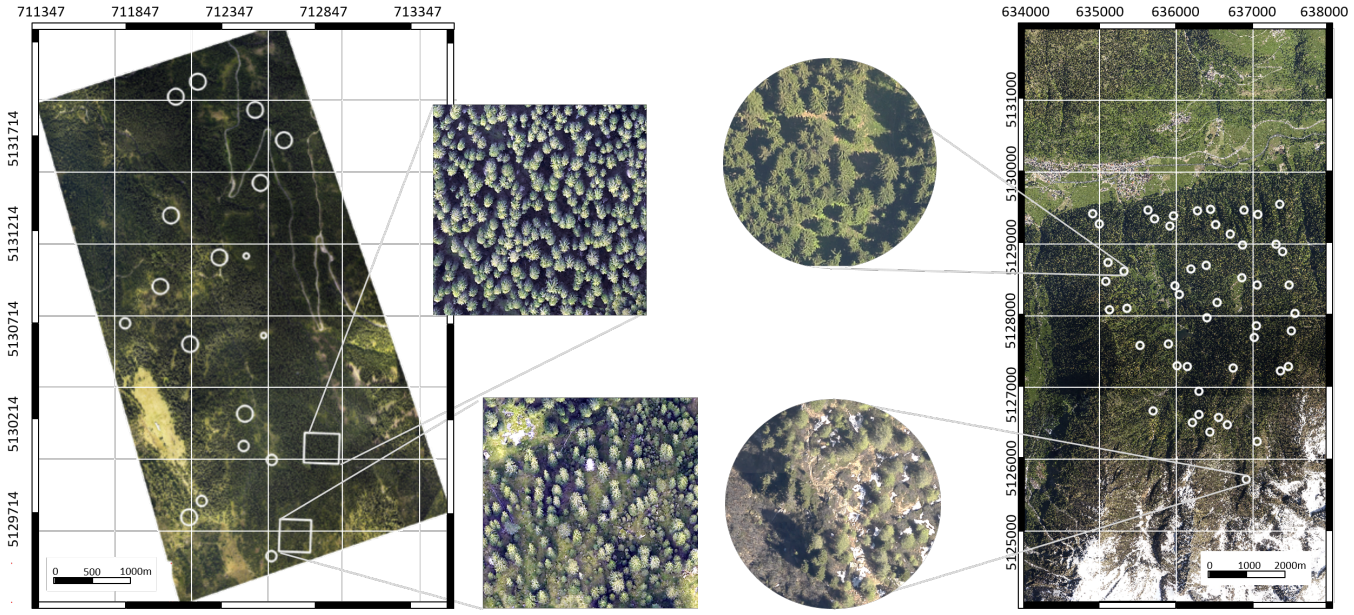
D. Data-Driven DBH Estimation

In the last step of the PM, *data-driven DBH estimation*, different regression models are defined and adopted for each growth-model class to accurately retrieve the DBH. Thus, the dependence of the DBH from the extracted variables varies according to the different environmental conditions. Accordingly, having a regression model tailored to each class allows us: (i) to adapt the regression rule to the class of trees, (ii) to detect the set of most informative variables per regression model and, (iii) to increase the correlation between the predicted variables and the stem attributes.

Let us assume to have a training set made up of q_s samples $T = (\mathbf{y}, \mathbf{X})$, where $\mathbf{X}$ is the $q_s \times d$ matrix of extracted variables and $\mathbf{y} \in \mathcal{R}^{q_s}$ is the vector of the observed values that need to be estimated. According to the clustering results obtained, the considered training set is partitioned into $p+1$ training sets $T_i (i = 1, .., p+1)$, where the i_{th} training set $T_i = (\mathbf{y}_i, \mathbf{X}_i)$ is composed of the q_i training samples associated to the i_{th} growth-model class ω_i. To avoid model overfitting and reduce the computational burden, a feature selection step is performed. This is done separately for each growth-model class in order to optimize the selection of features with respect to the behaviour of the specific class. Note that the feature selection is applied to the whole set of d features in order to identify for each class of trees the most informative features. The search algorithm used is based on the optimization of a multiobjective problem which optimizes both the mean squared error MSE and the determination coefficient R^2 on the validation set. The metrics are jointly optimized according to the concept of Pareto optimality [39]. The most informative set of features is detected by using a Sorting Genetic Algorithm II (NSGA-II), which achieves accurate feature selection results in a reasonable computational time [39], [40].

III. Dataset Description

Experiments were carried out on two LiDAR datasets acquired with different point densities in homogeneous conif-



(a) (b)

Figure 3. Study areas, Trentino region, Italy: (a) Paneveggio, and (b) Pellizzano. The stand plots are highlighted in white and the zooms point out the different forest density and structure conditions. Coordinates are reported in the UTM WGS84 32N system.

erous forests located in the Trentino region, Southern Alps, Italy (see Fig. 3). The considered mountainous scenario is characterized by a wide range of DBHs and complex terrains morphology (steep slopes, wide range of elevation and soil fertilities). The main forest species are Norway Spruce and European Larch, with a small presence of Silver Fir and Swiss Pine.

The first study area is located at Parco Naturale Paneveggio Pale di San Martino (see Fig. 3a). The coordinates of the central point of this area are $46°17'47{,}60''$ N, $11°45'29{,}98''$ E. The area extends approximately 3.68 km^2 and is characterized by a complex topography with hillsides having different inclinations and sun exposition (mainly north north-west aspect). The altitude ranges from 1536 to 2064 m a.s.l., whereas the slopes are up to 30°. Medium density LiDAR data, with a maximum pulse density of 5 pls/m^2, were acquired with an Optech ALTM 3100EA sensor in September 2007. The number of returns acquired per pulse was up to four, the laser pulse wavelength was 1064 nm and the pulse repetition frequency 100 kHz.

The second study area is located in the municipality of Pellizzano (see Fig. 3b). The coordinates of the central point of this area are $46°17'31{,}00''$ N, $10°45'56{,}49''$ E. The area extends approximately 32 km^2 and the altitude ranges from 900 to 2000 m a.s.l.. High-density LiDAR data, with a maximum pulse density of 15 pls/m^2, were acquired between 7th and 9th of September 2012 with a Riegl LMS - Q680i sensor. The pulse repetition frequency was 400 kHz and up to 4 returns were recorded per laser pulse.

Reference data were collected on the ground within 21 stand plots, for Paneveggio, and 52 stand plot, for Pellizzano. The plots are randomly distributed on the entire study area to obtain a uniform statistical representation in terms of topography and forest density. Species, DBH, volume and height were measured for some trees inside each plot of Pellizzano (1930 trees) and for all the trees having DBH larger than 3 cm in Paneveggio (1462 trees). A summary of the field data used in the analysis is reported in Tab. IVa and Tab. IVb for Paneveggio and Pellizzano, respectively. The samples were randomly divided into training, test and validation sets. While the test set allows us to detect both the best set of features and the best parameters of the regression model, the validation set is used to evaluate the performance of the PM. For each set, the species composition and the number of trees are reported. Moreover, the average, the minimum and the maximum values of DBH and tree top height are presented. In the considered experimental analysis we are dealing with homogeneous coniferous forests mainly dominated by Norway Spruce. However, note that the PM is data-driven, and thus adaptive. Therefore, it can be applied to any forest scenario. At regional level, the DBH inventory categories used in Trentino are the following:

- pre-inventory trees (DBH $\leq$ 17.5 cm);
- small trees (17.5 cm $<$ DBH $\leq$ 27.5 cm);
- medium trees (27.5 cm $<$ DBH $\leq$ 47.5 cm);
- large trees (DBH $>$ 47.5 cm).

This categorization aims to model the forest structure for management and planning. Fig. 4a and Fig. 4b report the number of samples for training, test and validation sets divided per DBH class for Paneveggio and Pellizzano, respectively. In Paneveggio, most of the trees have a medium size DBH which ranges between 27.5 cm and 47.5 cm with a significant presence of young trees having DBH $\leq$27.5 cm. Thus, Paneveggio is a forest mainly characterized by mature trees where young trees are still growing. In contrast, Pellizzano is an old-growth forest mainly characterized by old trees having medium or very large DBH values (i.e., $>$ 47.5 cm). Here,

Table IV
DISTRIBUTION OF THE REFERENCE DATA DIVIDED INTO TRAINING, TEST AND VALIDATION SETS. THE SPECIES COMPOSITION OF EACH SET AND THE VALUES OF DBH AND TREE HEIGHT ARE REPORTED FOR: (a) PANEVEGGIO, AND (b) PELLIZZANO.

	N.	Height (m)			DBH (cm)		
		Average	Min	Max	Average	Min	Max
Training Set	215	23.7	2.3	42	38.6	3	109
Silver Fir	1	4	4	4	10	10	10
Norway Spruce	203	23.7	2.3	42	38	3	109
Swiss Pine	4	21.9	19.3	26.8	65.8	42	92
European Larch	7	24.5	13.6	33.5	43.4	12	70
Test Set	142	24	2.3	43	38.1	3	74
Silver Fir	5	27.8	12.3	35.4	43.8	20	60
Norway Spruce	129	23.9	2.3	43	37.98	3	74
Swiss Pine	2	25.2	23.7	26.8	33.7	30.5	37
European Larch	6	24.9	7	31.5	37.33	12	54.5
Validation Set	1105	23.6	2.3	42	35.9	3	78
Silver Fir	10	21.3	3.7	34.6	32.2	8	66
Norway Spruce	1040	23.6	2.3	42	35.5	3	78
Swiss Pine	17	24.2	3.9	39.9	43.9	9	64
European Larch	38	25.2	3.6	38.7	42.7	7	70.5

(a)

	N.	Height (m)			DBH (cm)		
		Average	Min	Max	Average	Min	Max
Training Set	344	28.3	3.5	46	46.5	5	111
Norway Spruce	249	29	4.3	46	46.4	6	111
European Larch	95	26.4	3.5	45.5	46.7	5	105
Test Set	318	28.9	3	44	47.1	7	92
Norway Spruce	224	30.1	3	44	47.58	7	92
European Larch	94	26.1	75	43.34	45.83	10.5	86
Validation Set	1268	30.3	4.5	46	48.7	8	90.5
Norway Spruce	949	30.5	4.5	46	48.29	8	90.5
European Larch	319	29.5	5.7	468	49.71	10	85

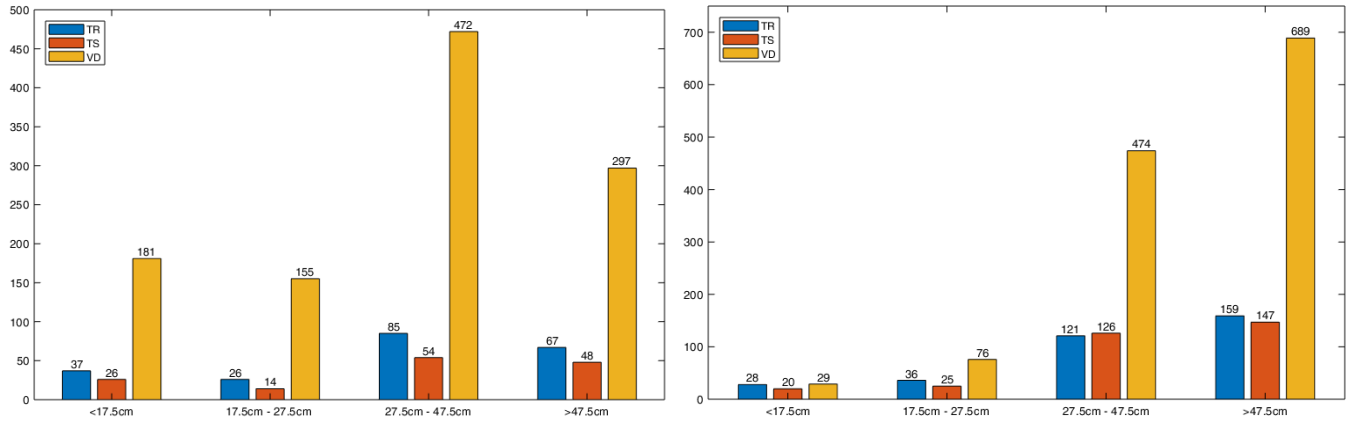
(b)

only few trees present small stems (i.e., $<$17.5 cm).

IV. EXPERIMENTAL RESULTS

In this section we present the experimental results obtained by applying the PM (hereafter referred as PM), to both datasets. The results are presented in terms of: i) number of clusters of trees belonging to different growth-models, ii) feature selection analysis, and iii) accuracy of DBH estimation.

The results obtained with the PM are compared with two reference methods present in the literature [9] (hereafter referred as RM1) and [12] (hereafter referred as RM2). In [9], the authors predict the DBHs by means of a regression analysis applied to the whole set of trees considering only features modelling the tree structure. Experiments were carried out on a medium density LiDAR dataset acquired in the Alpine scenario (i.e., Trentino Region). Different variable selection methods and different estimators were compared. Here, we considered the methods that provided the highest accuracy on the validation set (i.e., Sequential Forward Selection with multiple linear estimator and SVR estimator with linear kernel function). In [12] a multiple linear regression model is trained on a full waveform LiDAR dataset acquired in the Bavarian Forest National Park (which is a sub-alpine spruce forest). The prediction parameters used for the regression analysis aim to describe the crown shape (i.e., crown height, tree-top height, crown area and crown volume). These comparisons allow us to highlight the increase of accuracy obtained by the PM due to the novelties introduced with respect to common



(a) (b)

Figure 4. Number of samples for training (TR), test (TS) and validation (VD) sets divided per DBH classes for: (a) Paneveggio, and (b) Pellizzano.

practice described in the literature, which are: (i) the use of features that characterize the forest environment, (ii) the use of tailored regression models to separately estimate DBH of trees belonging to different growth-model classes. For all the datasets r_{env} was set equal to 10m, which is a standard plot size used to perform area-based LiDAR analysis and field surveys [10], [14], [41].

A. Growth-Models Identification

Fig. 5 shows the clustering results obtained in terms of detected forest layers (Fig. 5a and Fig. 5b) and growth-model classes (Fig. 5c and Fig. 5d) for Paneveggio and Pellizzano, respectively. According to the CH index, the number of mature classes p is equal to 2 for both datasets. In contrast, the number of forest layers detected is $n = 3$ for Paneveggio and $n = 4$ for Pellizzano. While Pellizzano is an old-growth forest characterized by a complex vertical structure, Paneveggio presents younger trees. This is also confirmed by the fact that most of the Pellizzano trees belong to the mature growth-model classes.

For both datasets, the young trees were accurately identified by the unsupervised clustering performed in the feature space that describes the vertical structure of the forest, i.e., $\omega_1 = L_1$. Then, the two classes ω_2 and ω_3 of mature and old trees were retrieved with the clustering applied to the feature space that describes the environment condition. As one can notice from Fig. 5c and Fig. 5d, the considered classes are characterized by different DBH/height growth rates due to the different tree structure, forest stand and topographic conditions. While young trees present almost linear relationship between height and DBH, mature and old trees present more complex relationships between DBH and tree top. Hence, for the same tree height, the DBH strongly varies due to the environmental conditions.

B. Feature Selection Results

Fig. 6 and Fig. 7 show the most informative features selected by the NSGA-II algorithm to train the regression model for the set of trees belonging to ω_1 (Fig. 6a and Fig. 7a), ω_2 (Fig. 6b and Fig. 7b) and ω_3 (Fig. 6c and Fig. 7c) for Paneveggio andn Pellizzano, respectively. For both datasets, the crown radius (r_1 or r_2) and the tree top height (H^1_{max}) are always selected regardless of the growth-model class, together with the terrain altitude features. This confirms the strong correlation that exists between the shape of the crown and the DBH, as well as the impact of the environment on the growth of the stem. Indeed, in addition to the attributes belonging to the $\mathbf{x}^{\text{Tree}}$ feature space, variables modeling both the topography and the stand density are always selected, thus confirming the importance of a proper representation of the terrain morphology and forest structure.

In Paneveggio, for the young tree class ω_1 the local forest density metric LCC, the profile curvature φ, the wetness index w and the terrain altitude A_{min} are selected. Also the DBH growth of the mature tree class ω_2 is influenced by the profile curvature φ and the terrain altitude A_{max}, whereas differently from the young trees, the terrain attribute selected is the slope S_{west}. Finally, for the old tree class ω_3 the vertical forest density p_2/p_1 and CRS, the aspect γ and the terrain altitude, A_{min} and A_{max}, mainly affect the stem expansion. Note that, even though the altitude features may be correlated in flat forest areas, in the considered complex mountainous scenario some trees are located on very steep slopes where the difference between the maximum and the minimum terrain altitude values (in the plot delineated around the tree) can be higher than 10m. In this context, the proposed data-driven method is able to accurately model the specific properties of the forest without requiring any preliminary analysis from the user.

For all the classes height percentiles are always selected. In contrast, in Pellizzano the height percentile (95th percentiles) is selected only for the young trees (ω_1). For the mature (ω_2) and old (ω_3) trees, in addition to the features modeling the crown shape, only variables modeling the topography and the stand density are selected. In particular, the aspect γ, the profile φ and plan curvature ϕ, and the wetness index w are

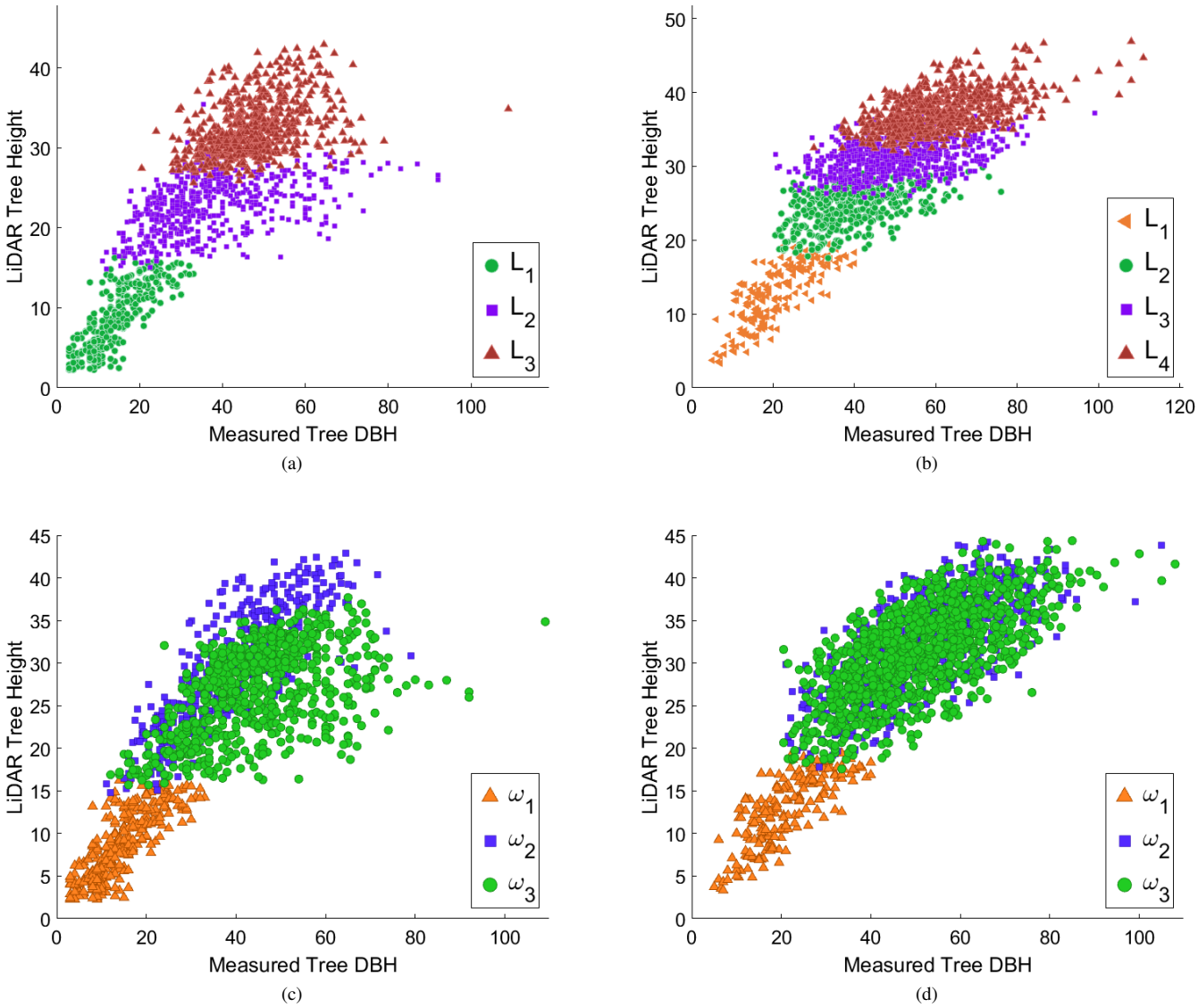


Figure 5. LiDAR tree heights vs measured tree DBH of the identified clusters of trees belonging to: (a) different layers of the vertical Paneveggio forest structure (L_1,L_2,L_3), (b) different layers of the vertical Pellizzano forest structure (L_1,L_2,L_3,L_4), (c) trees belonging to different growth-models of the Paneveggio forest, and (d) trees belonging to different growth-models of the Pellizzano forest. The young trees are classified as ω_1, the mature trees are classified as ω_2 and the old ones as ω_3.

selected together with the GCC for the trees belonging to ω_2. For the old trees ω_3, the vertical forest density p_2/p_1, the GCC, the aspect γ, the wetness index w and the slope S_{est} represent the most informative set of features.

C. Stem DBH Estimation

Tab. V shows the DBH predictions obtained with the PM, the RM1 and the RM2 divided per DBH class and on the entire set of trees for Paneveggio (Tab. Va) and Pellizzano (Tab. Vb).

To quantitatively evaluate the prediction accuracy, the DBH predicted with the RM1, the RM2 and the PM were compared to field measured DBH. In particular, we calculate the Mean Absolute Error (MAE), the Mean Square Error (MSE) and the percentage Root Mean Square Error (%RMSE). Note that the %RMSE is computed as the ratio between the RMSE and the average DBH value. In this framework, the %RMSE allows us to quantify the error on the DBH classes, by weighting the RMSE on the average DBH of each considered class. The results obtained on both datasets (see Tab. V) demonstrate the strong improvement obtained by the PM in estimating small and large stems (tails of the distributions), while similar results are achieved by the PM and both the reference methods on trees having medium size DBH.

In Paneveggio the error metrics were reduced by more than 2 cm on the entire dataset, with an RMSE lower than 7 cm for all the DBH classes. The PM achieved an RMSE of 6.12 cm, while the RM1 and the RM2 obtained 8.82 cm and 8.28 cm, respectively. In particular, the %RMSE of the PM ranges between 18% (on DBH between 27.5 cm$-$47.5 cm) and 28% (on DBH$<$17.5 cm). In contrast, RM1 generated a %RMSE

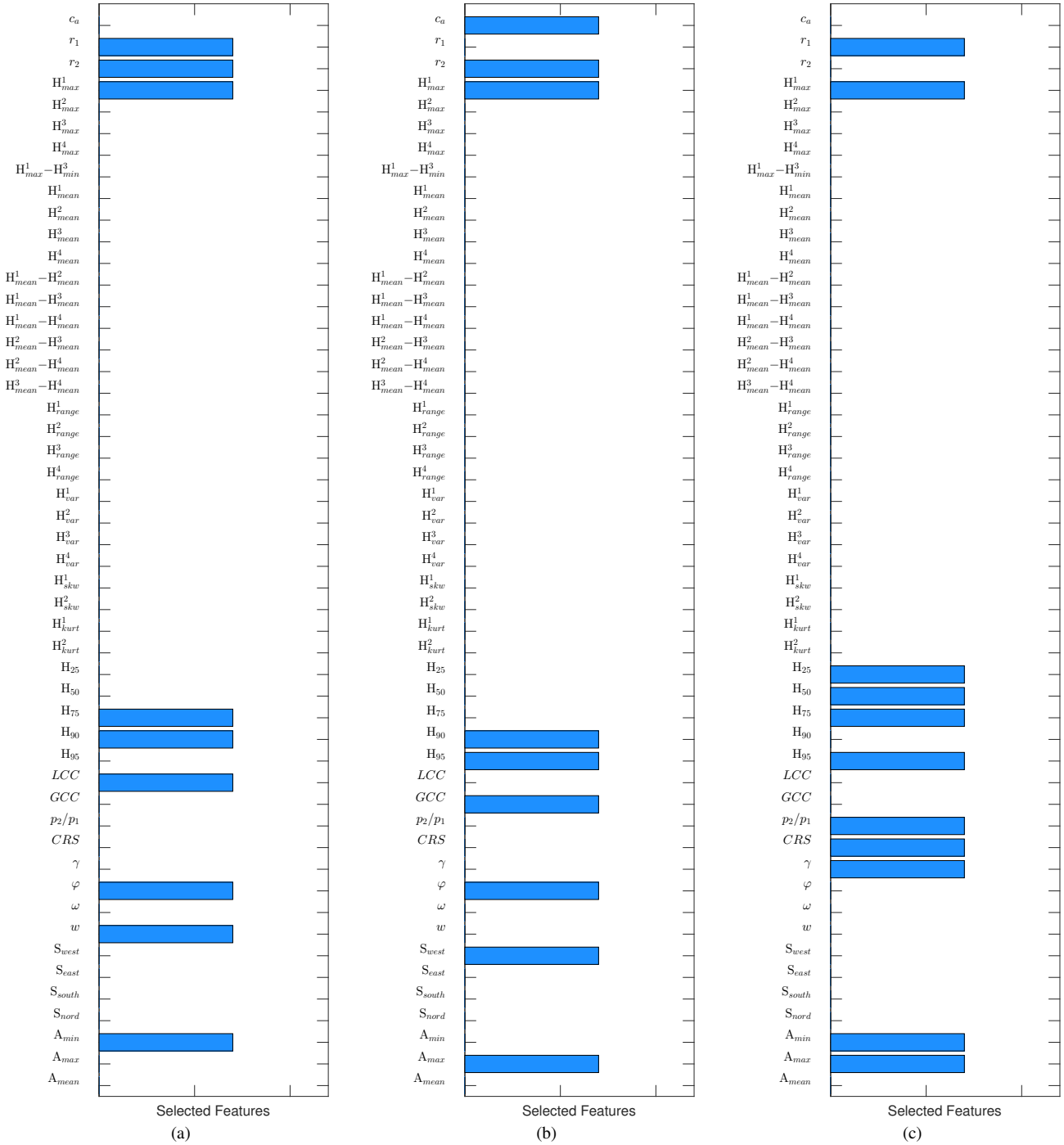


Figure 6. Features selected to train the regression model for: (a) set of young trees (ω_1), (b) set of mature trees (ω_2), and (c) set of old trees (ω_3). (Paneveggio)

that ranges between 24% (on DBH between 27.5 cm−47.5 cm) and 64% (on DBH<17.5 cm), while RM2 obtained a %RMSE that ranges between 21% (on DBH between 27.5 cm−47.5 cm) and 46% (on DBH<17.5 cm). From the results obtained, one can observe that the RM1 allows a better prediction of large stems with respect to the RM2, while the RM2 achieved better results than the RM1 on small stems. In contrast, the PM significantly outperformed both the methods on all the DBH classes.

Also in Pellizzano the PM obtained the best results on the entire dataset with a RMSE of 6.98 cm compared to 8.54 cm and 7.61 cm of the RM1 and RM2, respectively. This improvement is smaller compared to the one obtained in Paneveggio due to the different forest structures. While in Paneveggio we are dealing with a mixed-aged forest, characterized by groups of trees belonging to different growth models, Pellizzano is an old-growth forest mainly characterized by old/mature trees with little percentage of young trees. Similar to the Paneveggio dataset, the PM strongly reduces the RMSE with respect to RM1 for all the DBH classes. In greater detail, the PM reduced

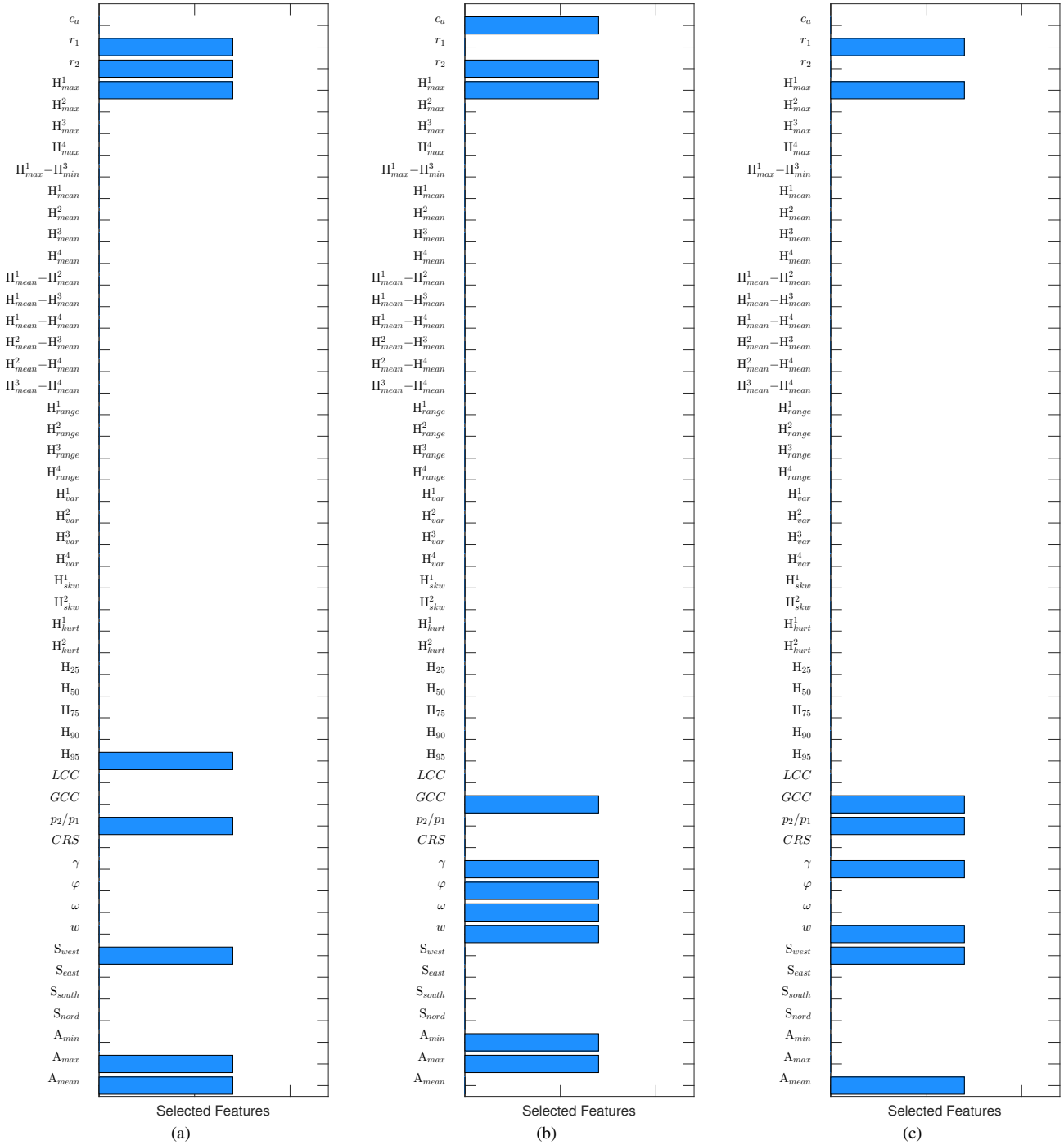


Figure 7. Features selected to train the regression model for: (a) set of young trees (ω_1), (b) set of mature trees (ω_2), and (c) set of old trees (ω_3). (Pellizzano)

the RMSE of more than 3 cm for trees having DBH<27.5 cm, and of 2.5 cm for medium stems (DBH between 27.5 cm−47.5 cm). The RM2 achieved slightly better accuracy than the PM on the pre-inventory stems (DBH<17.5 cm) and the medium stems (DBH between 27.5 cm−47.5 cm) by decreasing the RMSE of 0.36 cm and 0.07 cm, respectively. However, regardless of the forest structure, the PM improved the RMSE of more than 1cm on large stems (DBH>47.5 cm).

The accurate estimation of the DBHs leads to an accurate representation of the DBH distribution. This is confirmed by Fig. 8, which shows the capability of the PM to accurately model the DBH distribution both for Paneveggio (see Fig. 8a) and Pellizzano (see Fig. 8b). The PM achieves a reliable estimation of the size of the tree stems also at the tails of the DBH distribution.

Fig. 9a and Fig. 9b the number of samples in the training, test and validation sets for each growth model class for Paneveggio and Pellizzano, respectively. The DBH predictions obtained by the PM, the RM1 and the RM2 for each growth model class are reported in Tab. VIa and Tab. VIb for

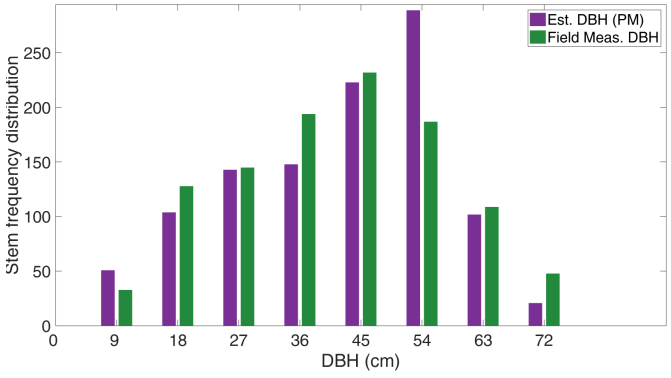
Table V
MAE, RMSE AND RMSE(%) CALCULATED ON THE DBH ESTIMATES FOR THE THREE GROWTH-MODEL CLASSES AUTOMATICALLY IDENTIFIED BY THE METHOD CONSIDERING THE PM, THE RM1 AND THE RM2 IN: (a) PANEVEGGIO, (b) PELLIZZANO.

DBH	PM			RM1			RM2		
class (cm)	MAE	RMSE	%RMSE	MAE	RMSE	%RMSE	MAE	RMSE	%RMSE
<17.5	2.39	3.17	28.80	5.78	7.07	64.18	3.96	5.08	46.13
17.5−27.5	4.43	5.92	26.35	8.08	9.72	43.29	5.78	7.36	32.77
27.5−47.5	5.29	6.53	17.46	7.29	8.94	23.91	5.93	7.76	20.74
>47.5	5.46	6.86	22.48	6.77	9.09	29.79	8.47	10.74	35.2
All Trees	4.74	6.12	17.03	7.01	8.82	24.56	6.27	8.28	23.05

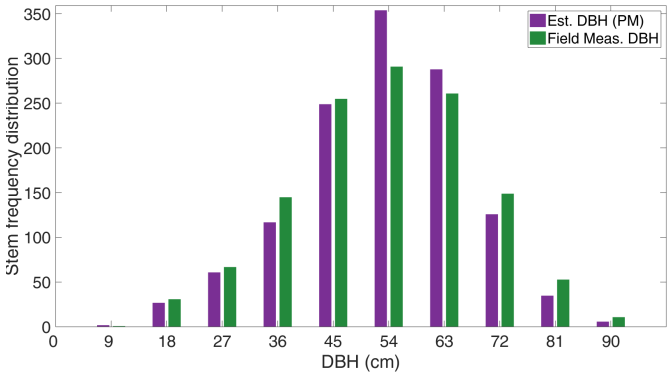
(a)

DBH	PM			RM1			RM2		
class (cm)	MAE	RMSE	%RMSE	MAE	RMSE	%RMSE	MAE	RMSE	%RMSE
<17.5	2.19	2.81	20.16	5.31	6.43	46.19	1.92	2.45	17.64
17.5−27.5	3.80	5.63	24.09	7.35	9.56	40.91	4.62	6.21	26.58
27.5−47.5	5.71	6.92	17.70	7.57	9.45	24.17	5.27	6.85	17.51
>47.5	5.97	7.28	16.92	6.26	7.82	18.17	6.78	8.35	19.43
All Trees	5.66	6.98	14.35	6.79	8.54	17.56	5.97	7.61	15.64

(b)



(a)



(b)

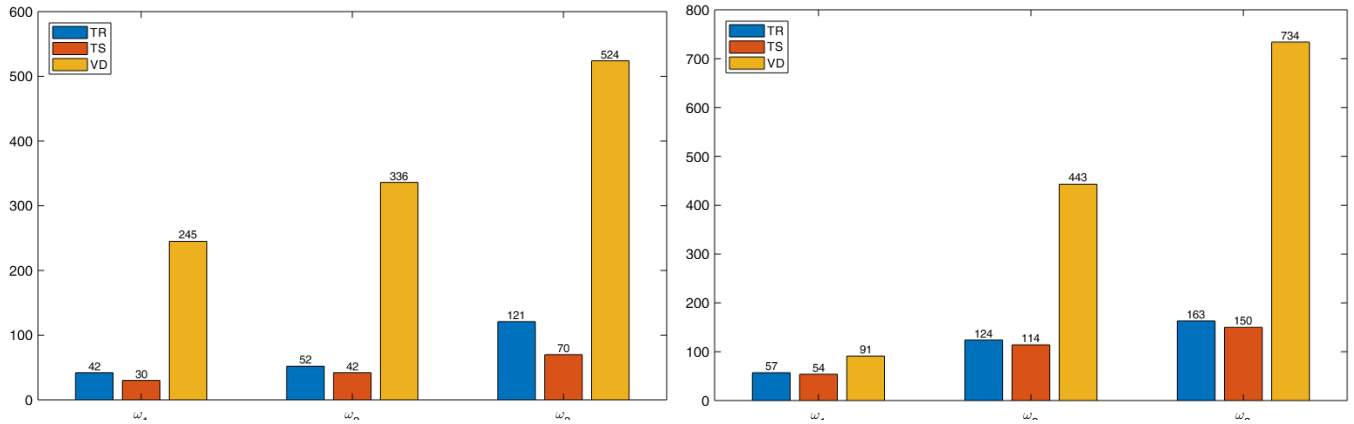
Figure 8. Field measured vs predicted DBH distributions: (a) Paneveggio, and (b) Pellizzano.

Paneveggio and Pellizzano, respectively. The results obtained confirm that the PM accurately retrieves the DBH for all the growth model classes, by reducing the estimation error with respect to both RM1 and RM2. In Paneveggio, the RMSE has been reduced by at least 2 cm for all the growth models. As expected, a smaller improvement is achieved on the old-growth Pellizzano forest. In this case the PM improved by more than 1 cm the DBH estimates compared to RM1 for all the classes. With respect to RM2, it reduced the error of less than 1 cm for ω_1 and ω_3, whereas improved of 1.50 cm the DBH estimates on ω_2.

Fig. 10 depicts the field measured versus the predicted DBHs by the PM (Fig. 10a and Fig. 10b), the RM1 (Fig. 10c and Fig. 10d) and the RM2 (Fig. 10e and Fig. 10f) for Paneveggio and Pellizzano, respectively. The plots show the correlation coefficient R^2 that represents the amount of variability within the estimates. The results confirm the quantitative evaluation of Tab. V, showing that the PM achieved the best estimation for both the datasets regardless of the density of the LiDAR data and the forest conditions.

V. ANALYSIS OF THE MAIN VARIABLES DISCRIMINATING DIFFERENT CLASSES OF MATURE TREES

Let us focus the attention on the growth model classes analysis performed to identify which variables mainly affect the stem growth of the mature trees. To gain a deepest understanding of the growth mechanism of the mature trees, we identify the set of variables that have major influence on discriminating among different growth-models. To this end, we consider the Jeffreys-Matusita distance (JM) in order to evaluate the statistical separability of the growth-model classes,



(a) (b)

Figure 9. Number of samples for training (TR), test (TS) and validation (VD) sets divided per growth model for: (a) Paneveggio, and (b) Pellizzano.

Table VI
MAE, RMSE AND RMSE(%) CALCULATED ON THE DBH ESTIMATES FOR THE THREE GROWTH-MODEL CLASSES AUTOMATICALLY IDENTIFIED BY THE METHOD CONSIDERING THE PM, THE RM1 AND THE RM2 IN: (A) PANEVEGGIO, AND (B) PELLIZZANO.

model	**PM**			**RM1**			**RM2**		
class	**MAE**	**RMSE**	**%RMSE**	**MAE**	**RMSE**	**%RMSE**	**MAE**	**RMSE**	**%RMSE**
ω_1	2.35	3.11	21.86	6.68	8.06	23.58	4.16	5.21	36.65
ω_2	5.34	6.87	17.46	6.92	8.83	29.45	7.06	9.23	23.45
ω_3	5.47	6.64	15.14	7.23	9.15	22.57	6.75	8.78	20.03

(a)

model	**PM**			**RM**			**RM2**		
class	**MAE**	**RMSE**	**%RMSE**	**MAE**	**RMSE**	**%RMSE**	**MAE**	**RMSE**	**%RMSE**
ω_1	2.32	3.07	14.37	6.04	7.51	14.95	2.39	3.27	15.34
ω_2	5.27	6.37	12.40	6.92	8.46	17.76	6.18	7.83	15.25
ω_3	6.30	7.65	15.18	6.81	8.71	17.76	6.29	7.85	15.58

(b)

and of the Sequential Floating Feature Selection (SFFS) search strategy to identify the subset of features that maximizes the separability criterion. This choice is motivated by the need to have a good trade-off between quality of selected features and computational time [42]. Let us consider the classes $\boldsymbol{\omega_h}$ and $\boldsymbol{\omega_t}$, the JM_{ht} between their distributions can be defined according to the Bhattacharyya distance B_{ht}:

$$\mathrm{JM}_{ht} = \sqrt{2\{1 - e^{-\mathrm{B}_{ht}}\}} \tag{12}$$

Under the simplifying assumption that the distributions of the growth-model classes can be modeled with Gaussian distributions, the Bhattacharyya distance can be defined as follows:

$$\mathrm{B}_{ht} = \frac{1}{8}(\mathbf{m}_h - \mathbf{m}_t)^T \left(\frac{\Sigma_h + \Sigma_t}{2}\right)^{-1} (\mathbf{m}_h - \mathbf{m}_t) + \\ + \frac{1}{2} ln\left(\frac{1}{2}\frac{|\Sigma_h + \Sigma_t|}{\sqrt{|\Sigma_h||\Sigma_t|}}\right) \tag{13}$$

where $\mathbf{m}_h$ and $\mathbf{m}_t$ are the mean vectors of the classes $\boldsymbol{\omega_h}$ and $\boldsymbol{\omega_t}$, and Σ_h and Σ_t their covariance matrices. Due to the capability of the JM distance to saturate when the discriminability between the classes does not increases by increasing their Bhattacharyya distance, we automatically detect: (i) the most relevant set of variables, and (ii) the number of variables to select. This analysis allows us: (i) to determine which variables mostly affect the growth of mature stems, (ii) to asses from the quantitative point of view the separability of the classes in the feature space where we perform the growth-models classification.

Tab. VII presents the ordered ranking of the main variables that discriminate between two different classes of stem growth of the mature trees detected in Paneveggio (see Tab. VIIa) and in Pellizzano (see Tab. VIIb). In the considered datasets, the clustering identified a class of mature trees and a class of old trees.

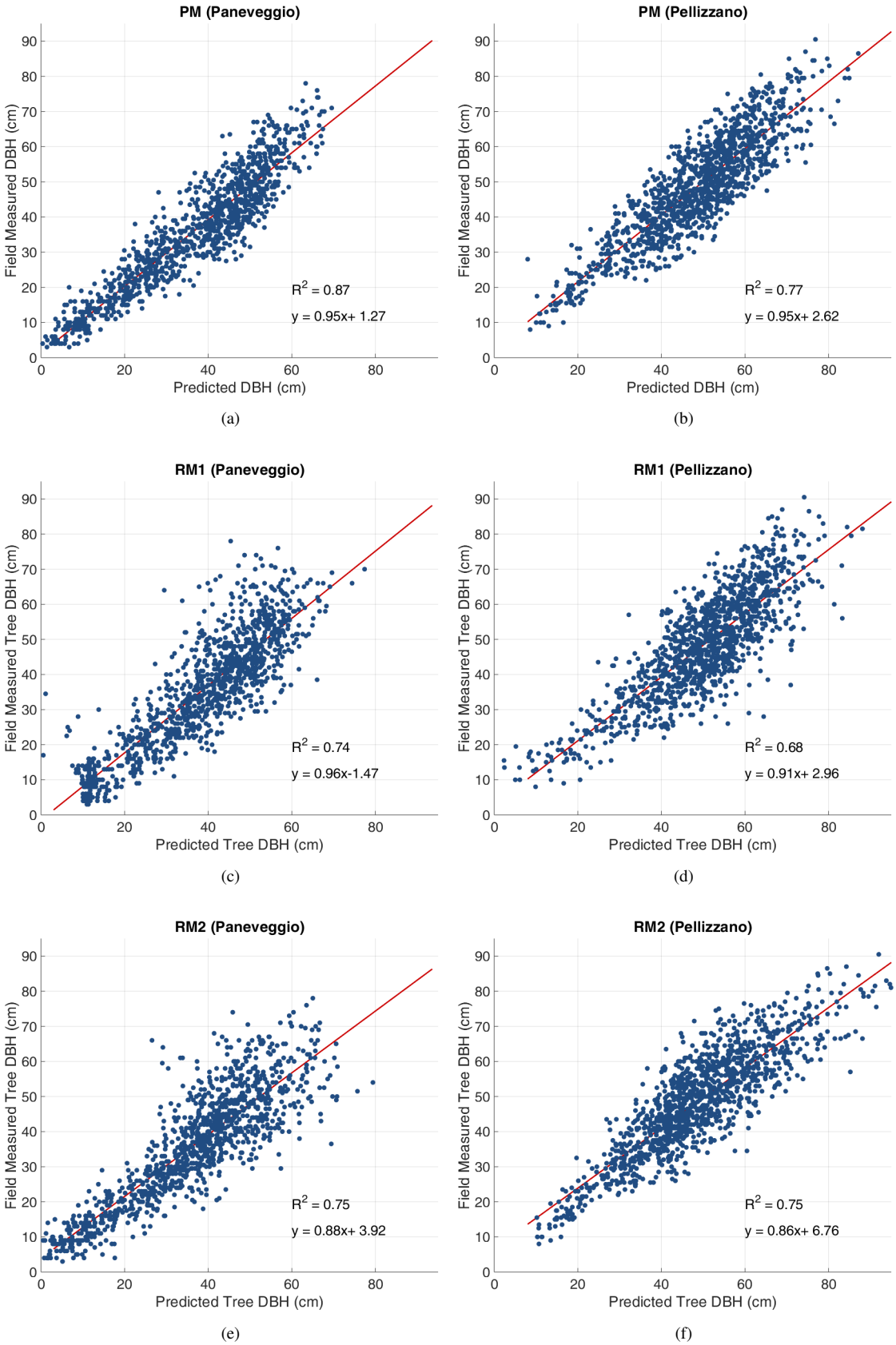


Figure 10. Field measured vs predicted by (a)-(b) the PM, (c)-(d) the RM1 and (e)-(f) the RM2 for Paneveggio and Pellizzano, respectively.

Table VII
ORDERED RANKING OF THE MOST DISCRIMINATIVE FEATURES BETWEEN ω_2 AND ω_3: (a) PANEVEGGIO, AND (b) PELLIZZANO.

Mature Growth-Models (Paneveggio)	
Maximum Altitude	A_{max}
Profile Curvature	φ
Wetness Index	w
Ratio of 2^{nd} and 1^{st} returns	p_2/p_1
Global Canopy Cover	GCC

(a)

Mature Growth-Models (Pellizzano)	
Aspect	γ
Southern Slope	S_{south}
Minimum Altitude	A_{min}
Plan Curvature	ϕ
Wetness Index	w

(b)

From the results obtained it turned out that the altitude plays a dominant role in the class separability of mature trees, i.e. A_{max} (Tab. VIIa) and A_{min} (Tab. VIIb). Indeed, the tree growth rate decreases when increasing the altitude because of the colder temperature, the increased exposure to wind, the shorter growing seasons and the reduced amount of soil nutrients [43]. The second variables selected in both datasets is the wetness index w which model the soil water drainage and thus is strongly correlated to the soil fertility, especially in mountainous and hilly terrains. This is confirmed by the selection of the profile curvature φ in the Paneveggio dataset and of the plan curvature ϕ In the Pellizzano dataset. Thus, these variables have a similar meaning since they both represent the curvature of the terrain from the hydrological and geomorphological point of view [34]. Finally, in Paneveggio the vertical density of the forest stand p_2/p_1 plays a fundamental role in the stem growth in terms of availability of water, whereas the horizontal forest density metrics GCC models the impact of forest density in terms of surrounding trees. In particular, trees characterized by low height values are more sensitive to the presence of taller neighbouring trees because of the light reduction effect. In contrast, in Pellizzano, the aspect γ and the southern slope S_{south} are selected, which provide an accurate representation of the terrain morphology. Differently from Paneveggio, none of the feature representing the forest density are selected. This is due to the fact that Pellizzano is an old-growth forest, where most of the trees are no more affected by the forest density but mainly by the terrain properties. This confirm the capabilities of the proposed data-driven automatic technique to obtain from the data information that is important for a physical understanding of the modelling processes.

VI. CONCLUSION

In this paper we have presented a data-driven method for the identification of clusters of trees belonging to different growth-model classes for the adaptive estimation of the individual DBH. The experimental results obtained on two LiDAR datasets demonstrate that the method is able to account for the environmental factors which can be computed from LiDAR data that influence the growth of the DBH. Due to the proper representation of the forest conditions, it accurately identifies clusters of trees belonging to different growth-model classes. This allows us to adapt the regression rule to the different classes and to select the best set of features for each growth-model class, thus improving the correlation between the predicted and the field measured DBH values. The PM allows a significant improvement with respect to the reference methods in the youngest mixed-aged forest characterized by different growth models (Paneveggio). As expected, the improvement is a smaller on the old-growth forest characterized by few young trees (Pellizzano). However, note that also in this case the RMSE obtained by the PM on trees characterized by large stems is significantly smaller than those yielded by reference methods, regardless of the density of the LiDAR data and the forest structure. Note that the PM is automatic and data-driven. Thus, it can be applied to different areas for adaptively identifying the specific growth-models to be used.

A growth-model classes analysis that determines which variables mainly discriminate growth-model classes of mature trees has been also presented. From the results obtained, it turned out that the altitude together with the water drainage and the wetness index plays a dominant role in the separability of the mature-tree classes. Thus, the altitude is fundamental to characterize the local forest environment and the tree growth condition, while the water drainage is directly linked to the soil fertility. Moreover, results demonstrate that for the young-growth forest dataset considered (Paneveggio) attributes modelling the forest density strongly influence the discrimination of mature classes. In contrast, in the old-growth forest dataset (Pellizzano) the topography has a dominant role in the separability of the mature classes.

By focusing the attention on the feature selection results obtained per growth-model class, the crown radius and the tree top height are always selected as features in input to the regression models regardless of the growth-model classes and the dataset. This result confirms the strong influence of the crown shape in the DBH estimation. Moreover, also the altitude is always selected by the feature selection algorithm due to the impact of the elevation on the forest environment. This is evident in the considered mountainous forest datasets. Although the precise characterization of the shape of the crown is important to obtain an accurate estimation of the DBH, features representing the forest density and the terrain morphology are always selected. In particular, the estimation results confirm that the terrain properties affect the stem expansion together with the forest density.

As future developments of this work, we aim to test the method on forest stands characterized by different ages and structures. Moreover, we plan to further analyze the growth-model classes for a better comprehension of the environmental factor which affect the growth of the trees. In this context, experiments will be extended to LiDAR dataset characterize by different laser point density and to forest having different environmental conditions.

VII. Acknowledgement

The authors would like to thank the "Dipartimento Risorse Forestali e Montane" of the Autonomous Province of Trento for providing the LiDAR data used in this study in the framework of the FORLIDAR project.

References

[1] R. A. Spriggs, D. A. Coomes, T. A. Jones, J. P. Caspersen, and M. C. Vanderwel, "An alternative approach to using lidar remote sensing data to predict stem diameter distributions across a temperate forest landscape," *Remote Sensing*, vol. 9, no. 9, p. 944, 2017.

[2] J. Vauhkonen and L. Mehtätalo, "Matching remotely sensed and field-measured tree size distributions," *Canadian Journal of Forest Research*, vol. 45, no. 3, pp. 353–363, 2014.

[3] Q. Xu, Z. Hou, M. Maltamo, and T. Tokola, "Calibration of area based diameter distribution with individual tree based diameter estimates using airborne laser scanning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, pp. 65–75, 2014.

[4] A. Persson, J. Holmgren, and U. Söderman, "Detecting and measuring individual trees using an airborne laser scanner," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 9, pp. 925–932, 2002.

[5] M. Heurich, "Automatic recognition and measurement of single trees based on data from airborne laser scanning over the richly structured natural forests of the bavarian forest national park," *Forest Ecology and Management*, vol. 255, no. 7, pp. 2416–2433, 2008.

[6] S. C. Popescu, "Estimating biomass of individual pine trees using airborne lidar," *Biomass and Bioenergy*, vol. 31, no. 9, pp. 646–655, 2007.

[7] J. Wu, W. Yao, S. Choi, T. Park, and R. B. Myneni, "A comparative study of predicting dbh and stem volume of individual trees in a temperate forest using airborne waveform lidar," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2267–2271, 2015.

[8] M. Maltamo, E. Næsset, and J. Vauhkonen, *Forestry applications of airborne laser scanning: concepts and case studies*. Springer Science & Business Media, 2014, vol. 27.

[9] M. Dalponte, L. Bruzzone, and D. Gianelle, "A system for the estimation of single-tree stem diameter and volume using multireturn lidar data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 7, pp. 2479–2490, 2011.

[10] X. Yu, J. Hyyppä, M. Vastaranta, M. Holopainen, and R. Viitala, "Predicting individual tree attributes from airborne laser point clouds based on the random forests technique," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 1, pp. 28–37, 2011.

[11] X. Yu, J. Hyyppä, M. Holopainen, and M. Vastaranta, "Comparison of area-based and individual tree-based methods for predicting plot-level forest attributes," *Remote Sensing*, vol. 2, no. 6, pp. 1481–1495, 2010.

[12] W. Yao, P. Krzystek, and M. Heurich, "Tree species classification and estimation of stem volume and dbh based on single tree extraction by exploiting airborne full-waveform lidar data," *Remote Sensing of Environment*, vol. 123, pp. 368 – 380, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425712001599

[13] C.-S. Lo and C. Lin, "Growth-competition-based stem diameter and volume modeling for tree-level forest inventory using airborne lidar data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 4, pp. 2216–2226, 2013.

[14] M. Vastaranta, V. Kankare, M. Holopainen, X. Yu, J. Hyyppä, and H. Hyyppä, "Combination of individual tree detection and area-based approach in imputation of forest variables using airborne laser data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 73–79, 2012.

[15] J. Vauhkonen, I. Korpela, M. Maltamo, and T. Tokola, "Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics," *Remote Sensing of Environment*, vol. 114, no. 6, pp. 1263–1276, 2010.

[16] Q. Ma, Y. Su, S. Tao, and Q. Guo, "Quantifying individual tree growth and tree competition using bi-temporal airborne laser scanning data: a case study in the sierra nevada mountains, california," *International Journal of Digital Earth*, pp. 1–19, 2017.

[17] A. Bucksch, R. Lindenbergh, M. Z. A. Rahman, and M. Menenti, "Breast height diameter estimation from high-density airborne lidar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 6, pp. 1056–1060, 2014.

[18] C. Salas, L. Ene, T. G. Gregoire, E. Næsset, and T. Gobakken, "Modelling tree diameter from airborne laser scanning derived variables: a comparison of spatial statistical models," *Remote Sensing of Environment*, vol. 114, no. 6, pp. 1277–1285, 2010.

[19] L. Zhang, H. Bi, P. Cheng, and C. J. Davis, "Modeling spatial variation in tree diameter–height relationships," *Forest Ecology and Management*, vol. 189, no. 1, pp. 317–329, 2004.

[20] L. Zhang, J. H. Gove, and L. S. Heath, "Spatial residual analysis of six modeling techniques," *Ecological Modelling*, vol. 186, no. 2, pp. 154–177, 2005.

[21] F. Mauro Gutiérrez, V. J. Monleon, H. Temesgen, and L. Á. Ruíz Fernández, "Analysis of spatial correlation in predictive models of forest variables that use lidar auxiliary information," *Canadian Journal of Forest Research*, no. ja, 2017.

[22] Y. Zhang, Y. Yao, X. Wang, Y. Liu, and S. Piao, "Mapping spatial distribution of forest age in china," *Earth and Space Science*, vol. 4, no. 3, pp. 108–116, 2017.

[23] W. Simonson, P. Ruiz-Benito, F. Valladares, and D. Coomes, "Modelling above-ground carbon dynamics using multi-temporal airborne lidar: insights from a mediterranean woodland," *Biogeosciences*, vol. 13, no. 4, p. 961, 2016.

[24] M. W. Palace, F. B. Sullivan, M. J. Ducey, R. N. Treuhaft, C. Herrick, J. Z. Shimbo, and J. Mota-E-Silva, "Estimating forest structure in a tropical forest using field measurements, a synthetic model and discrete return lidar data," *Remote Sensing of Environment*, vol. 161, pp. 1–11, 2015.

[25] C. Paris, D. Valduga, and L. Bruzzone, "A hierarchical approach to three-dimensional segmentation of lidar data at single-tree level in a multilayered forest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4190–4203, 2016.

[26] E. Kansa, "Multiquadricsa scattered data approximation scheme with applications to computational fluid-dynamicsii solutions to parabolic, hyperbolic and elliptic partial differential equations," *Computers & Mathematics with Applications*, vol. 19, no. 8, pp. 147 – 161, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/089812219090271K

[27] E. Naesset, "Estimating timber volume of forest stands using airborne laser scanner data," *Remote Sensing of Environment*, vol. 61, no. 2, pp. 246–253, 1997.

[28] M. García, D. Riaño, E. Chuvieco, and F. M. Danson, "Estimating biomass carbon stocks for a mediterranean forest in central spain using lidar height and intensity data," *Remote Sensing of Environment*, vol. 114, no. 4, pp. 816–830, 2010.

[29] Q. Chen, P. Gong, D. Baldocchi, and Y. Q. Tian, "Estimating basal area and stem volume for individual trees from lidar data," *Photogrammetric Engineering & Remote Sensing*, vol. 73, no. 12, pp. 1355–1365, 2007.

[30] H. O. Ørka, E. Næsset, and O. M. Bollandsås, "Effects of different sensors and leaf-on and leaf-off canopy conditions on echo distributions and individual tree properties derived from airborne laser scanning," *Remote Sensing of Environment*, vol. 114, no. 7, pp. 1445–1461, 2010.

[31] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and lidar remote sensing data for classification of complex forest areas," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 5, pp. 1416–1427, 2008.

[32] J. E. Means, S. A. Acker, D. J. Harding, J. B. Blair, M. A. Lefsky, W. B. Cohen, M. E. Harmon, and W. A. McKee, "Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the western cascades of oregon," *Remote Sensing of Environment*, vol. 67, no. 3, pp. 298–308, 1999.

[33] A. T. Hudak, J. S. Evans, M. J. Falkowski, N. L. Crookston, P. E. Gessler, P. Morgan, and A. Smith, "Predicting plot basal area and tree density in mixed-conifer forest from lidar and advanced land imager (ali) data," *Proceedings of the 26th Canadian Symposium on Remote Sensing*, 2005.

[34] I. D. Moore, P. Gessler, G. Nielsen, and G. Peterson, "Soil attribute prediction using terrain analysis," *Soil Science Society of America Journal*, vol. 57, no. 2, pp. 443–452, 1993.

[35] G. Jordan, "Morphometric analysis and tectonic interpretation of digital terrain data: a case study," *Earth Surface Processes and Landforms*, vol. 28, no. 8, pp. 807–822, 2003.

[36] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[37] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[38] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 911–916.

[39] L. Pasolli, C. Notarnicola, and L. Bruzzone, "Multi-objective parameter optimization in support vector regression: General formulation and application to the retrieval of soil moisture from remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 5, pp. 1495–1508, 2012.

[40] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[41] M. Nilsson, "Estimation of tree heights and stand volume using an airborne lidar system," *Remote Sensing of Environment*, vol. 56, no. 1, pp. 1–7, 1996.

[42] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[43] D. A. Coomes and R. B. Allen, "Effects of size, competition and altitude on tree growth," *Journal of Ecology*, vol. 95, no. 5, pp. 1084–1097, 2007.