Journal of Computational Science 58 (2022) 101517

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Above-ground biomass estimation from LiDAR data using random forest algorithms☆

Leyre Torre-Tojal [a,*], Aitor Bastarrika [a], Ana Boyano [b], Jose Manuel Lopez-Guede [c,e], Manuel Graña [d,e]

[a] Department of Mining and Metallurgical Engineering and Materials Science, Faculty of Engineering, University of the Basque Country (UPV/EHU), Nieves Cano 12, 01006 Vitoria-Gasteiz, Spain
[b] Mechanical Engineering Department, Faculty of Engineering of Vitoria-Gasteiz, University of the Basque Country, UPV/EHU, Nieves Cano 12, 01006 Vitoria-Gasteiz, Spain
[c] Department of Systems Engineering and Automatic Control, Faculty of Engineering, University of the Basque Country (UPV/EHU), Nieves Cano 12, 01006 Vitoria-Gasteiz, Spain
[d] Department of Computer Science and Artificial Intelligence, Faculty of Computer Science, University of the Basque Country (UPV/EHU), Paseo Manuel De Lardizabal, 1, 20018 Donostia-San Sebastian, Spain
[e] Computational Intelligence Group, University of the Basque Country (UPV/EHU), Spain

ARTICLE INFO

Keywords:
LiDAR
Biomass
Regression
Random forest

ABSTRACT

Random forest (RF) models were developed to estimate the biomass for the *Pinus radiata* species in a region of the Basque Autonomous Community where this species has high cover, using the National Forest Inventory, allometric equations and low-density discrete LiDAR data. This article explores the tuning for RF hyperparameters, obtaining two models with an R^2 higher than 0.7 using 2-fold cross-validation. The models selected were applied in Orozko, a municipality with more than 5000 ha of this species, where the model predicts a biomass of 1.06–1.08 Mton, which is between 16–18 % higher than the biomass predicted by the Basque Government.

1. Introduction

The United Nations framework convention on climate change, which is the basis for the Kyoto Protocol, recognises in its first paragraph that changes observed in the Earth's climate and their adverse effects have become a shared concern of all humanity, since part of these changes have been attributable to human activities [1]. Sustainable forestry development can greatly help to mitigate climate change in the long term, because on the one hand it prevents the introduction of new carbon into its active cycle and, on the other, it supplies goods and services to society. An important part of the carbon cycle is related to the amount of carbon that is retained in the biomass, which will later be exchanged naturally with the atmosphere. For this reason, estimation of the above-ground biomass, especially forest biomass, has gained great interest among the scientific community.

Traditionally, and still today, biomass estimates have been made mainly using both direct and indirect methods. Direct or destructive methods involve cutting individual trees for subsequent weighing of the stems, branches and leaves directly, subsequently determining their dry weight [2]. Conversely, indirect methods advocate not destroying material by relying on data already inventoried. The data commonly acquired refers to the diameter of the trunk at chest height (at 1.3 m) and the height of the trees located in the sample plots, with biomass then being estimated from this data using allometric equations [3].

One of the main drawbacks of forest inventories is that their creation is a process that involves major investment in time and money - all the more so, the larger the area to be inventoried. To alleviate the cost, Remote Sensing (RS) techniques have been used because of their ability to obtain data quickly and accurately over large areas [4]. Light detection and ranging (LiDAR) is an alternative methodology that can be used

☆ The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: leyre.torre@ehu.es (L. Torre-Tojal).

https://doi.org/10.1016/j.jocs.2021.101517
Received 25 April 2021; Received in revised form 30 October 2021; Accepted 24 November 2021
Available online 29 November 2021

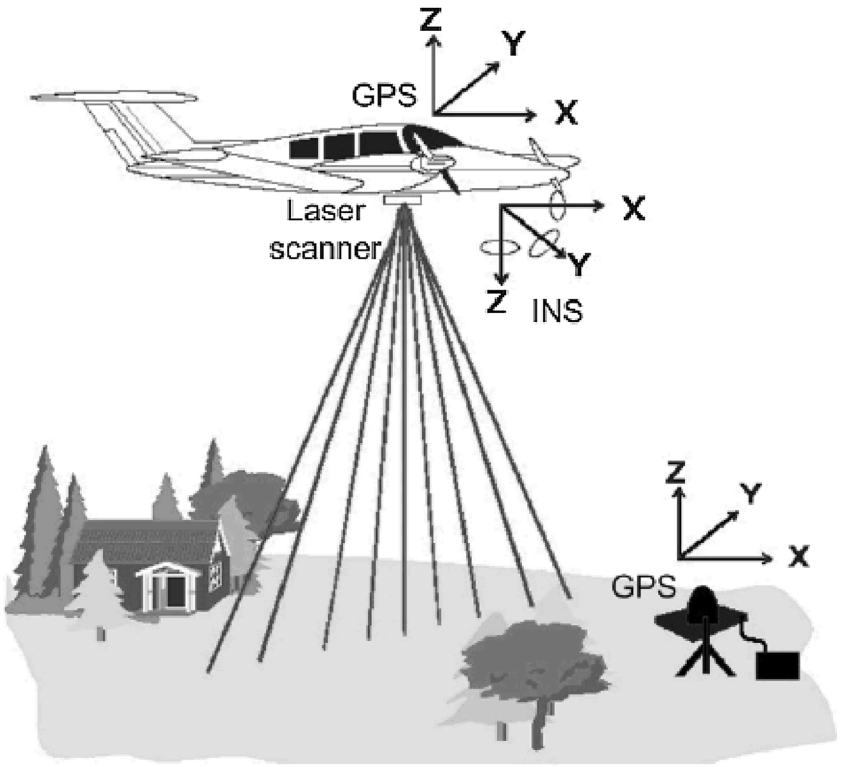


Fig. 1. Simplified conceptual diagram of the LiDAR system (source: Alternative Methodologies for LiDAR System Calibration, Ki In Bang).

to avoid the disadvantages mentioned previously, and has become an effective and accurate tool for characterising the forest canopy in large areas [5], even being able to discriminate between tree species [6].

Diverse approaches have been considered to estimate forest biomass using LiDAR data [7–11], classified depending on footprint size and the object being studied (plot size or individual tree). To this end, time series of LiDAR measurements have already been used to monitor tree growth [12]. Other studies have combined LiDAR data with other imaging sources, such as hyperspectral images [13,14] or multispectral satellite images [15] or, lately, Unmanned Aerial Vehicle UAV-LiDAR data [16].

Different approaches from the methodological point of view have been taken using a range of techniques, the most popular option being the Multiple Linear Regression (MLR) [17–19]. Some disadvantages of this technique, such as the difficulty in capturing complex and non-linear relationships or multicollinearity problems, have led to the need to experiment with other methods such as Machine Learning techniques, which have achieved promising results [20,21]. Among these methods, Random Forest (RF) is a robust Machine Learning algorithm which is able to capture these complex relationships in order to predict forest characteristics accurately, as several studies in different regions have shown [22,23].

The purpose of this paper is to explore Random Forest regression for above-ground biomass estimation of *P. Radiata* species, by carrying out an in-depth sensitivity analysis of Random Forest hyperparameters using public data. On hand, the ground truth biomass was computed from the Fourth National Forest Inventory (NFI4) dendometric measures (diameter at breast height and height), while on the other we exploited the low density (0.5 points/ m^2) LiDAR dataset contained in the cartographic National Plan of Aerial Orthophotography (PNOA), a product that covers the whole area of Spain.

The remaining sections of the paper are organised as follows: Section 2 explains the basis of the LiDAR technology; Section 3 describes the dataset used in this research; Section 4 explains the methodology applied in the study; Section 5 presents the results obtained from the experiments and formulation of the best model and validation of the model developed in a different area, while Section 6 includes the discussion of the results. Finally, Section 7 explains the conclusions and proposals for future work.

2. LiDAR background

LiDAR (Light Detection and Ranging) technology constitutes an active system for massive remote position measurement, based on a laser scanning sensor (infrared spectral region) that emits pulses and registers returns against the surface. The LiDAR measurement system is based on the response time obtained for each pulse, from its emission to its reception by the sensor, once reflected against a surface. Knowing the speed of light and the time elapsed, the distance at which the object that has generated the return is located is then established immediately, and so if the pointing angle of the laser is known at the time of measurement, it will be possible to obtain the X, Y, Z coordinates of the objects reflected.

An airborne LiDAR system can be installed in different platforms, such as satellites, aircrafts, helicopters or drones, while terrestrial LiDAR installations can be static, when mounted on a tripod, or dynamic, if mounted on ground vehicles. Even if the technological principles are the same [24], there are differences regarding data capture and processing steps, and also regarding applications of the data between static and dynamical systems. When the LiDAR sensor is mounted on a dynamic platform (aircraft, car, etc.), a Global Positioning System (GPS) and an Inertial Measurement Unit (IMU) are needed to integrate the information, in order to obtain accurate positioning of the data (Fig. 1). Distance measurement can be carried out according to two main technologies: either in full waveform mode or by recording discrete returns. In the first case, the sensor digitizes the entire return signal for a time interval. In the second, the sensor registers the different echoes of the beam as the laser beam partially intercepts different objects in its path, if they exist. Besides, recently several innovative LiDAR sensors have emerged, enabling new applications to be used in Earth sciences [25].

LiDAR ability to characterise the vertical forest structure makes it a

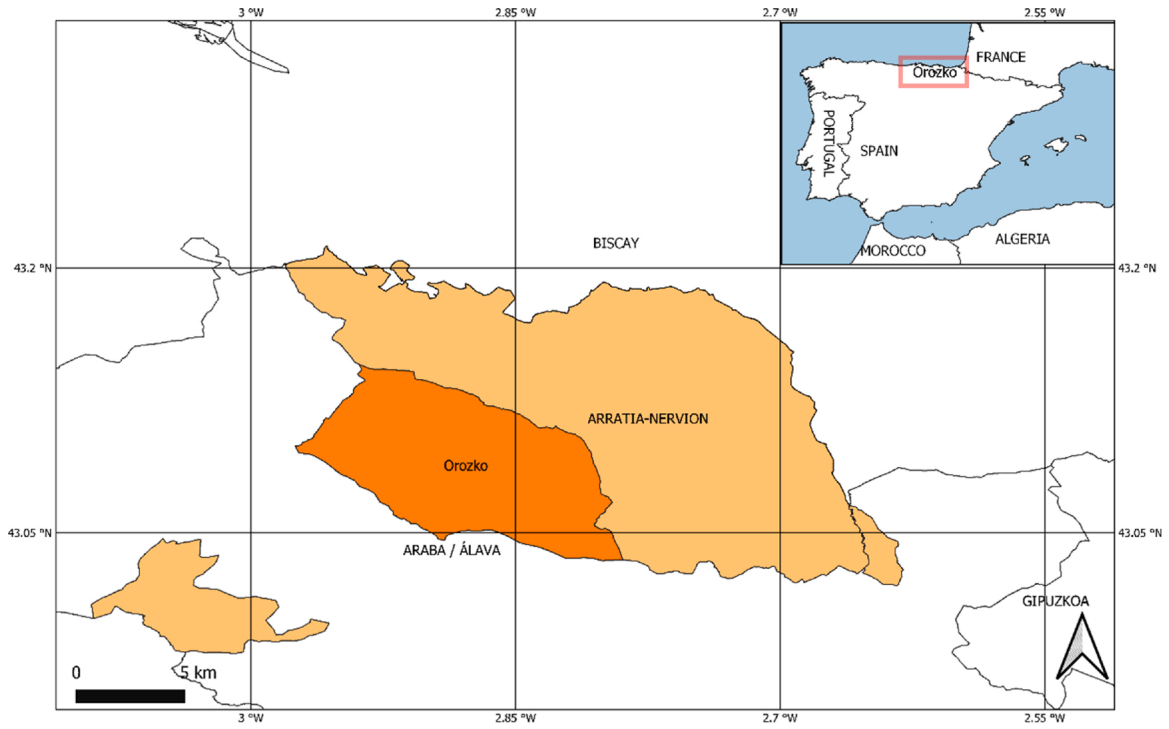


Fig. 2. Location of the study area, namely Arratia-Nervión (Biscay, Spain) and the test site, Orozko.

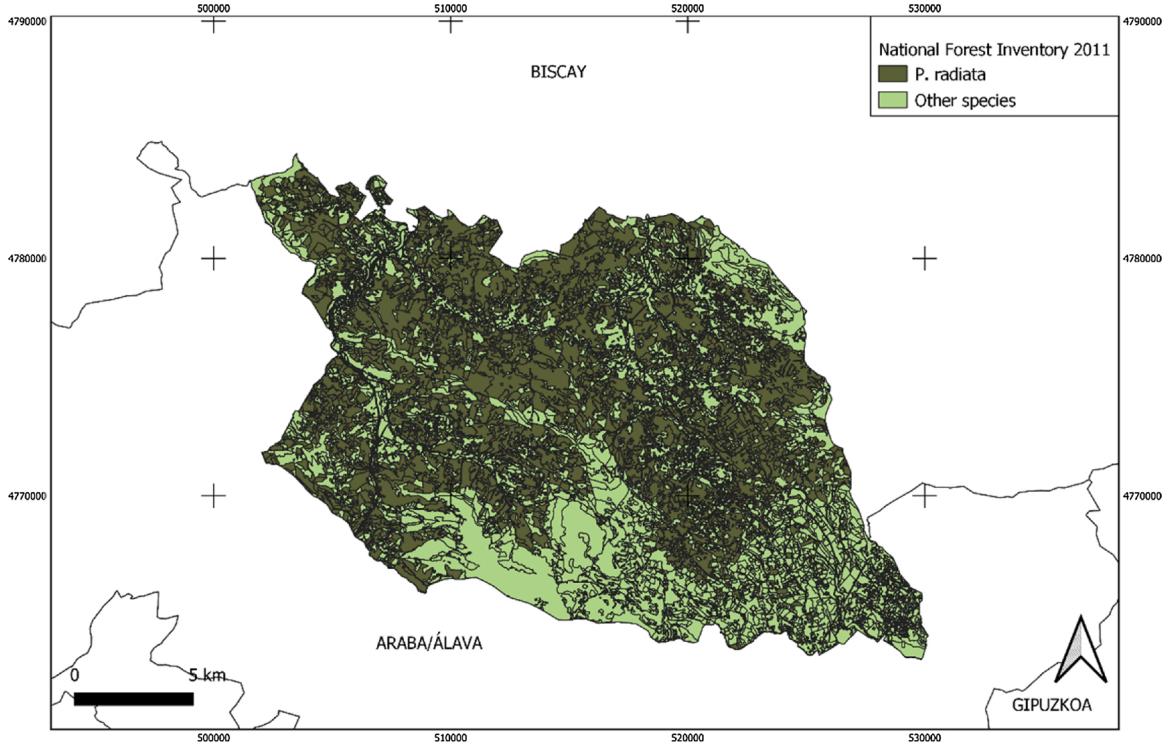


Fig. 3. *P. radiata* distribution (dark green colour) in the Arratia-Nervión region (ETRS89 UTM zone 30 North reference system) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

break-through technology in forestry applications, providing accurate estimations for essential forest structural characteristics such as canopy heights, basal area, stand volume and above-ground biomass [26]. The integration of LiDAR data with other datasets with global coverage [27] represents an opportunity to estimate forest characteristics in very large areas while taking temporal and spatial dynamics into consideration.

3. Materials

3.1. Study area

The study area covers the Arratia-Nervión region, located in the Historical Territory of Bizkaia, in the northern part of Spain (Fig. 2).

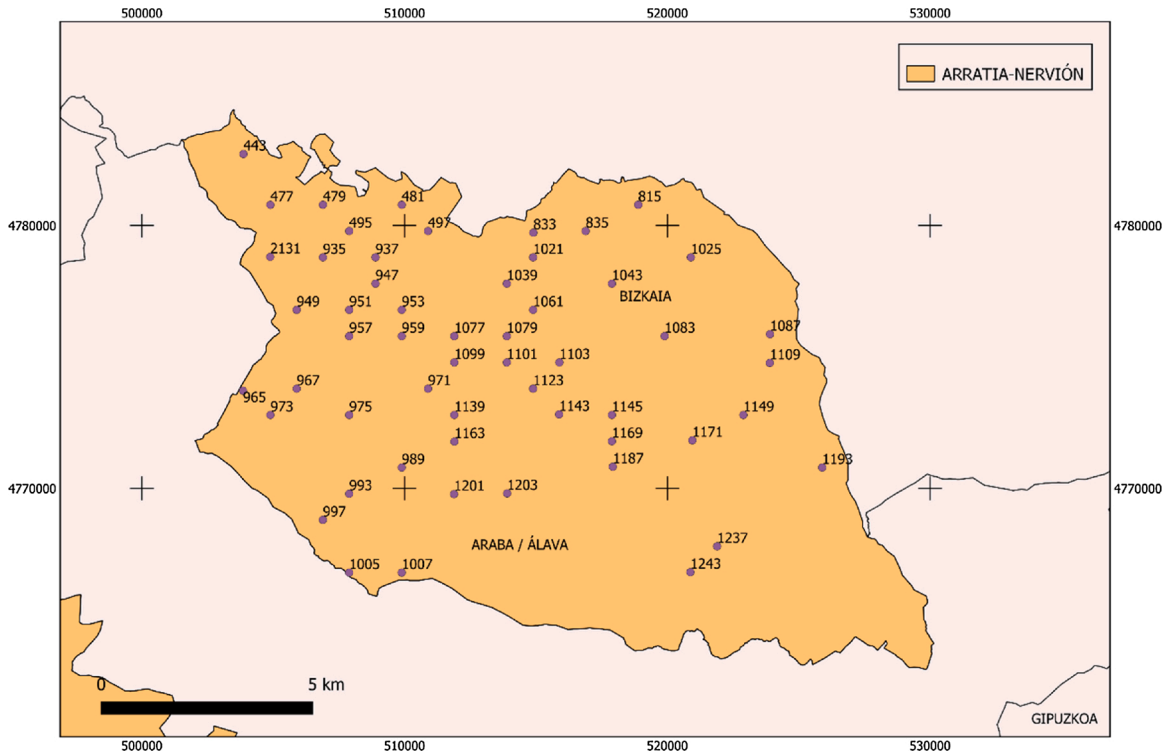


Fig. 4. Sample plot distribution of the NFI 4 in Arratia-Nervión (ETRS89 UTM zone 30 North reference system).

With an average altitude of 465 m, the slopes of the area have an average slope of 18.6 °, finding areas of high slopes between 30 and 45 °, distributed almost throughout the total area of the region (400 km^2). Average annual rainfall is estimated at about 1200 mm with monthly highs being recorded in November and December, while the lows are in September and October. The rest of the year, rainfall remains regular, except for storms of cyclonic origin during the summer months.

Regarding vegetation, the progressive replacement of native species (holm oak, beech, scrub and meadows) by conifers and other fast-growing species has led to pine forests becoming the main tree formation in the region. Pine forests of *P. radiata* are the most important species in the Basque Country, representing over 32 % of the forested area and occupying 125,000 ha. According to the NFI4, 16,260 ha out of 28,065 ha of forest areas in Arratia-Nervión belong to the *P. radiata* tree species, accounting for almost 60 % of the tree specimens in the region (Fig. 3).

One of the municipalities of the region – Orozko - which will be used as the test site, is set in a mountainous landscape dominated by varied and extensive forest masses. This area has a surface area of 102.3 km^2, with 5,000 ha being occupied by the *P. radiata* and constituting the largest area in the Basque Country occupied by this species.

3.2. Data

3.2.1. Fourth National Forest Inventory (NFI4)

The dendrometric data (such as tree height and diameter), collected for the species analysed, will allow biomass to be obtained in the reference plots in the research by applying allometric equations. The data for ground truth are provided by the Fourth National Forest Inventory (NFI 4) of the Basque Autonomous Community, and was gathered between January 17 and June 15, 2011. The NFI 4 is the fourth phase of a ten-year statistic to recognise and estimate values and indicators for a major part of these productive, protective, ecological and recreational functions.

The vertices of the UTM cartography kilometre grid in the ED50 reference system, which are within the areas classified as wooded, were adopted as sampling plots (Fig. 4). This distribution of the sample allows the latter to be distributed in the strata with proportional fixation, systematic establishment of random starting, and a sampling intensity generally of one plot per square kilometre. For strictly budgetary reasons, one out of every two parcels were measured in this fourth edition of the inventory.

Nested plot methodology was used to measure the trees in the plots, as defined by the Nature Conservation Institute [28], where each plot generates four circular plots of variable radius 5, 10, 15, and 25 m, representing a maximum area of approximately 0.2 ha in the case of the biggest radius. The use of plots of variable radius for field data collection implies the need to use expansion factors to be able to ascertain the results per unit area [29]. The expansion factor refers to the coefficient that, multiplied appropriately, converts the concentric plots into estimates of the "real" plots of the maximum radius being considered, and. is calculated according to the following expression (Eq. (1)):

$$f = \frac{10,000}{\pi R^2} \tag{1}$$

Where R is the radius of the plot.

From a total of 118 plots placed in the area of interest, 67 were chosen for the study, with the fact of having more than 80 % occupation of *P. radiata* being the primary inclusion criterion. Several concordance checks were subsequently undertaken based on this initial sample, as the forest inventory field work and LiDAR data acquisition are subject to a 10-month delay.

Plot coherence was verified and any possible anomalies detected, such as forest treatment experienced in the zone or any geolocation discrepancies between the NFN4 and corresponding LiDAR data. Orthophotos obtained from the frames of the digital photogrammetric flight with a 25 cm pixel resolution were used for such purpose, these being gathered between July 23 and August 8, 2012. Silvicultural treatment was applied in various plots, and so the population of pines had decreased significantly. In other cases, classification mistakes in the point cloud were detected, and four plots were discarded because of having hardly any trees. Finally, 12 plots from the original sample were

Table 1
LiDAR flight parameters.

Scan angle	60°
Pulse Repetition Rate (PRR)	100 kHz
Scan frequency	70 kHz
Beam divergence	<0.5 mrad
Average speed of the aircraft	67 m/s
Point density	0.5 points/ m^2
Average altitude above ground level	1500 m
Spatial localization of the points accuracy	<10 cm

removed, obtaining a final sample comprising 55 plots.

3.2.2. LiDAR data

LiDAR data in the point cloud was obtained from the LiDAR flight of the Basque Autonomous Community undertaken between July 12 and August 28, 2012, using a Lite Mapper 6800 Airborne Laser. The configuration parameters are described in Table 1:

The reference system used was the European Terrestrial Reference System 89 (ETRS89) and the coordinate system was UTM for the 30th time zone. The dataset was divided into sheets of 2 × 2 km in extension, classified into eight types: Unclassified, Ground, Low Vegetation, Medium Vegetation, High Vegetation, Building, Low Points and Reserved. Data is publicly available at: ftp://ftp.geo.euskadi.eus/lidar/LIDAR_2012_ETRS89/LAS/. 134 sheets were downloaded and processed in total, containing more than 400 million points, with an altimetric range from 45 m to 1482.66 m.

4. Methods

4.1. Ground biomass estimation

As mentioned in Section 3.2.1., allometric equations were applied to obtain a ground truth biomass, with most of the allometric equations having related the value of biomass measured in the field (generally by destructive methods) to easily obtainable variables, such as diameters and heights. Several authors have compiled numerous allometric equations for different species across very diverse areas of the world, such as Europe, Australia and America [30–32]. The allometric equations for biomass estimation developed by Montero et al., have been widely use in Spain [2].

In this study, volume per tree was calculated using an allometric model developed by the HAZI Foundation due to geographical proximity to our study area [33]. The model was based on the destructive sample of 732 *P. radiata* specimens extracted from locations distributed across the Basque region, between the years 1990 and 2001. The model showed a difference of 5% between estimated values and those obtained using the destructive method. The volume over bark (VOB) for every tree was estimated using diameter (d) and height (h), in accordance with Eq. (2):

$$\mathrm{VOB}\ (\mathrm{dm}^3) = 0.0006785 \bullet \mathrm{d}^{1.86004}\ \mathrm{h}^{1.01378} \quad (2)$$

To obtain the biomass, a correction factor of 4% of the volume was added, because the tree branches and the thinnest part of the tree trunk were not taken into account in the field measurements, due to the wood production processes. This quantity was assumed to be fixed per species and independent of any characteristics of the forest stand. Finally, these biomass values were extrapolated to a hectare extension, using expansion factors detailed in section 3.2.1.

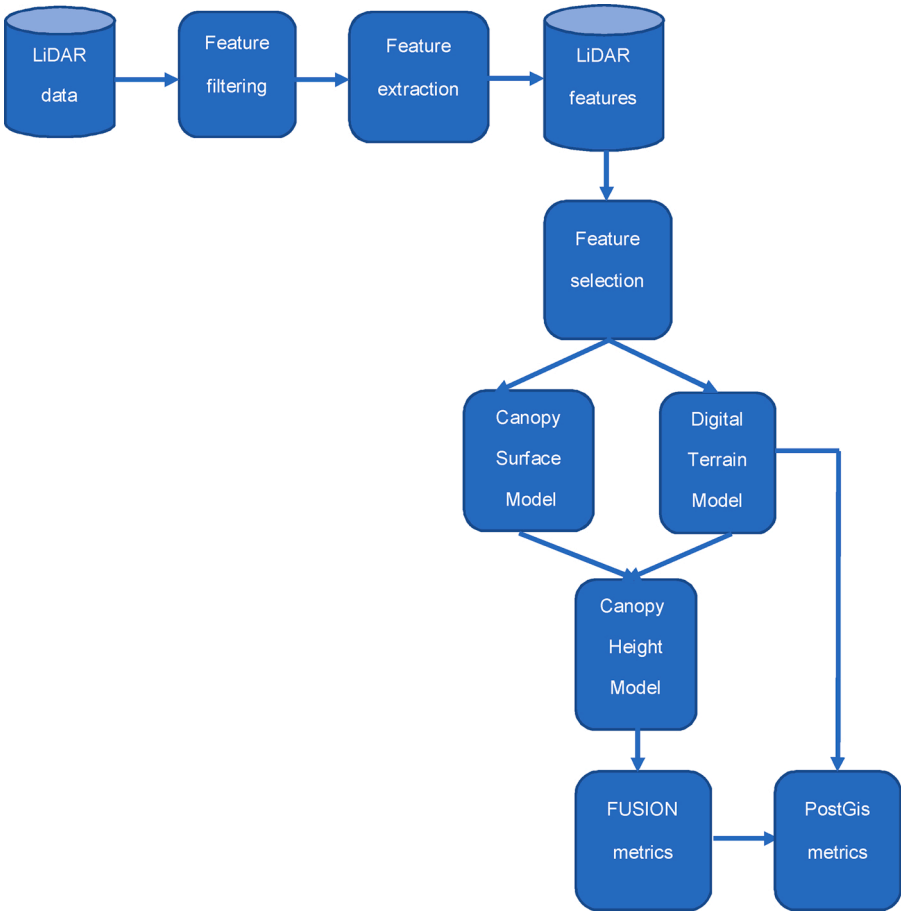


Fig. 5. LiDAR Data processing pipeline in the study.

4.2. Lidar data pre-processing

LiDAR point clouds were handled using the FUSION/LDV (LiDAR data viewer) suite (http://forsys.cfr.washington.edu/fusion/fusionlatest.html) free open-source software. First, we selected the LiDAR returns which spatially coincide with the 25 m radius circular sample plots inventoried in the NFI4. Next, we applied an algorithm to detect any possible erroneous returns, in the course of which only 0.06 % of the total number of returns were removed. This algorithm divides the altimetric range of the plot into a set of altimetric layers of constant thickness, and each point is assigned to the corresponding layer. This will detect which layers have dots and which ones appear empty, with those surrounded by empty layers and far from the area where contiguous, non-empty layers are concentrated being considered as anomalous points.

Overall LiDAR data processing is shown in Fig.5:

Even if the percentage of erroneous returns is small, erroneous return readings may introduce critical errors in later calculations. We then extracted the Digital Terrain Models (DTM) and Digital Surface Models (DSM) of each sample plot from the filtered LiDAR data. The DTM represents the bare ground surface with objects ignored, whereas the DSM provides the top of the canopy in forested areas, called Canopy Surface Model (CSM). In the first case, only the returns classified as ground were used, in the second one, the last returns were used to compute the CSM. The Canopy Height Model (CHM) was established as referring to the subtraction of these two models, resulting in a normalized point cloud, which provides tree heights directly from the model.

Once the DTM, CSM, and CHM had been obtained, a collection of variables was then computed for each sample plot from the set of all returns above a 2 m threshold, in order to avoid shrubs [34], including height distributions, canopy density metrics, and descriptive statistics as detailed in Appendix A.

In addition to those 55 variables obtained from the FUSION/LDV software, another 10 metrics related to the canopy point density for each sample plot were computed in the PostGis environment (https://postgis.net). According to literature, canopy density metrics (such as mean height, dominant height, mean diameter, basal area and timber volume) have proved to be useful in estimating forest parameters [35]. These density metrics were computed based on subdivision of the point clouds into 10 vertical equal layers distributed between the lower and upper limits. The lower limit was set as 2 m, in order to avoid shrubs, while we set the upper limit at the 95th percentile of height distribution. This upper limit was chosen instead of the maximum height due to the stability demonstrated in previous studies [36,37]. Finally, we calculated the proportion of points from the total number of points contained above each layer, obtaining 10 new density variables denoted as {tr10, tr20, tr30,…,tr100}, described in Appendix A. Overall, 66 variables were finally extracted from each of the 55 plots selected.

4.3. Biomass estimation: variable identification and modelling

A wide variety of statistical approaches has been applied in literature regarding empirical modelling of the biomass [38,39], multivariate linear regression being the most widely-used method [40]; however, recently, researchers have been increasingly experimenting with sophisticated machine learning regression for biomass estimation like Random Forest (RF), Support Vector Machine (SVM) or Artificial Neural Network (ANN) [20,21,41]. Among them, the RF approach has gained acceptance in forestry applications due to its robustness and modelling flexibility in predicting values associated with new unknown samples and is considered to be one of the most accurate general-purpose learning techniques available [42]. This technique is fast and easy to implement, and makes highly accurate predictions even with highly-correlated variables, which is the case when modelling biomass data.

By definition, RF is a supervised non-parametric ensemble classification/regression technique, which selects randomly a subset of explanatory variables totally randomly and builds the tree up to a certain point. This method first grows several decision trees using CART methodology and later combines the predictions from all the trees to produce the ensemble response [43]. Random Forest may be used both for classification (the output of the random forest is the class selected by most trees) or regression (the mean or average prediction of the individual trees is returned). By ensuring the forest grows up to a user-defined number of trees, the algorithm creates trees that have high variance and low bias [44].

Table 2
Description and value range for the RF hyperparameters being considered. 'MSE' and' MSA' correspond to Mean Standard Error and Mean Absolute Error, respectively.

Hyperparameter	Description	Default value
n_estimators	The number of trees in the forest.	100
max_features	The number of variables to consider.	Auto, max_features = n_features.
max_depth	The maximum depth of the tree.	None, nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
min_samples_leaf	The minimum number of samples required to be at a leaf node.	1
min_samples_split	The minimum number of samples required to split an internal node.	2
max_leaf_nodes	Best nodes are defined as relative reduction in impurity.	None, unlimited number of leaf nodes.
criterion	The function to measure the quality of a split.	MSE
Bootstrap	Whether bootstrap samples are used when building trees.	True

Table 2 describes the parameters and default values established by Random Forest Regression Algorithm implementation in scikit learn (sklearn.ensemble.RandomForestRegressor), a very well-known Python module that integrates a wide range of state-of-the-art machine learning algorithms [45]. The first two parameters to adjust when using the random forest regressor methods are the *n_estimators* and the *max_features* parameters (Table 2). The first one defines the number of trees in the forest. Increasing the number of trees usually allows for a more reliable result, although the trees will stop significantly improving beyond a critical number, increasing the computational cost. The latter refers to the size of the random subsets of features to be considered when splitting a node: The lower this is, the greater the reduction in variance, but also the greater the increase in bias. There are some parameters used to control the depth of the tree (and therefore to reduce overfitting) such as maximum depth (*max_depth*), the longest path between root node and leaf node, and defining the minimum samples required to be a leaf node (*min_samples_leaf*), the minimum number of samples required to split an internal node (*min_samples_split*) and the maximum number of leaf nodes (*max_leaf_nodes*). The criterion parameter defines the function used to measure the quality of the split. Finally, the bootstrap parameter activates a sub-sample size used to build each tree (the default option is True); otherwise, the whole dataset is used to build each tree if the bootstrap option is False.).

4.3.1. Variable reduction

As occurs with many other applications with a biological background, the number of available samples is lower than the number of variables, being able to assess the importance of each variable is crucial in reducing the high dimensionality, in order to improve the accuracy of the model and prevent overfitting. Automated model selection methods,

Table 3
Description and value range for the RF hyperparameters being considered. 'MSE' and' MSA' correspond to Mean Standard Error and Mean Absolute Error, respectively.

Hyperparameter	Range [min, max] or [options]	Interval
n_estimators (n_e)	[1, 500]	1
max_features (m_f)	[1, Number of selected features]	1
max_depth (m_d)	[1, 100]	1
min_samples_leaf (m_s_l)	[1,Number of samples]	1
min_samples_split (m_s_s)	[2, Number of samples]	1
max_leaf_nodes (m_l_n)	[1, Number of samples]	1
criterion (c)	[MSE, MAE]	
Bootstrap (b)	[True, False]	

such as backward or forward stepwise regression, are classical solutions to this problem, but are generally based on major assumptions about the functional form of the model or the distribution of residuals. Many researchers use RF to screen variables and reduce dimensionality [46].

In this study, the importance of each feature was given by the Gini importance or Mean Decrease in Impurity (MDI), which calculates the sum over the number of splits (across all trees) that include the variable, proportionally to the number of samples it splits [47]. The higher the mean MDI, the greater the importance of the variable in the model. A large number of trees (2000) were used during the variable reduction step in order to assess the importance of each variable.

Gini Importance was used to sort the variables (from greatest importance to least) and assess model performance by taking consecutive larger variable subsets. We started assessing the model using the most important variable (f1), followed by the two most important ones (f1, f2), the three most important (f1, f2, f3), and so on. Each model was assessed by cross-validation, preventing overfitting by partitioning the data into two sets (2 CV). Using a higher number of folds was not considered, this being deemed appropriate because of the low number of available training samples (N = 55). For its part, the shuffle option was activated in order to generate a user-defined number of independent training/test dataset splits, and samples were first shuffled and then split into a pair of training and test sets.

4.3.2. Hyperparameter tuning

Selection of the best hyperparameter configuration for machine learning models has a direct impact on how the model will perform, with the process pursued with a view to designing the ideal model architecture being known as hyperparameter tuning. Manual optimization of the hyperparameters is not always satisfactory when dealing with a large number of hyperparameters, resulting in computationally expensive or complex models. A more automated procedure to tune hyperparameters is the hyperparameter optimization (HPO). This process reduces the human effort required, in turn improving performance and reproducibility of the models [48].

Grid search and Randomized search are the two popular methods for hyperparameter optimization of models. While Grid Search is an exhaustive search over of all combinations of values of all the specified hyperparameters and their values and calculating the performance for each combination, the randomized search is more time-efficient as it selects a user-defined random combination of hyperparameter values to train the model and score it. In this study, both approaches are explored. On the one hand, an exhaustive search was undertaken but only with one hyperparameter each time, setting the other parameters as default, while on the other, the randomized search implemented in scikit-learn was employed (simultaneously searching all the hyperparameters without fixing any default value), making 100,000 combinations (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html). Both approaches use the variables' range defined in Table 3.

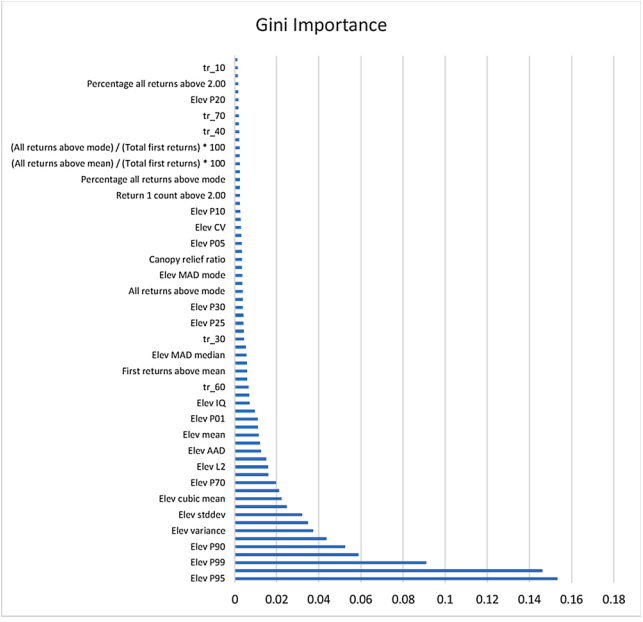


Fig. 6. GINI Importance of the variables extracted from LiDAR clouds.

4.3.3. Model assessment and validation

The k-fold cross-validation procedure is used to estimate the expected performance of machine learning models when making predictions on data not used during the training process. This procedure makes it possible to compare models for model selection and hyperparameter tuning without the need for a separate validation set. A model is trained using k−1 of the folds as training data, and the resulting model is then validated on the remaining part of the data that is used as a test set to compute performance metrics or measures such as accuracy. The performance measure reported by *k*-fold cross-validation will then be the average of the values computed in the loop. A two-fold cross-validation is used in our study, and the coefficient of determination (R^2) and Root Mean Square Error (RMSE) statistics are used to evaluate model performance. The R^2 score represents the proportion of the variability explained by the independent variables in the model, which provides an indication of goodness of fit and therefore a measure of how well information associated with unseen samples is likely to be predicted by the model, via the proportion of explained variance [49]. This coefficient provides values from 0 to 1, with 1 being the best possible score. The performance measure reported by 2-fold cross-validation will be the average and standard deviation of the R^2 computed in the loop, while the RMSE represents the standard deviation of the residuals (prediction errors). Residuals measure of how far data points are from the regression line, while RMSE measure how spread out these residuals are [50].

Once the parameters of the RF have been tuned and models obtained, they have then been validated. The validation process involves a comparison between the biomass predictions and observations over a set of measures that are different to those used in the model adjustment [51]. In our study, this validation was achieved by comparing the biomass predicted in the Orozko municipality to an estimation by the HAZI Foundation that is considered official by the Basque Government, while the biomass estimations provided by the HAZI foundation for this municipality were used as the ground truth biomass. For its estimation, HAZI used the NFI4 data gathered in 2011 and LiDAR data from the 2012 flight, using a linear regression model approach to estimate wood volume from the mean height of the LiDAR points 4 m above ground as the single regressor of the model. As explained in section 4.1, a correction factor of 4% of the volume was calculated and added to obtain the biomass.

The forestry map contained in the NFI4 was employed to compute the biomass estimated in Orozko municipality using RF models. This

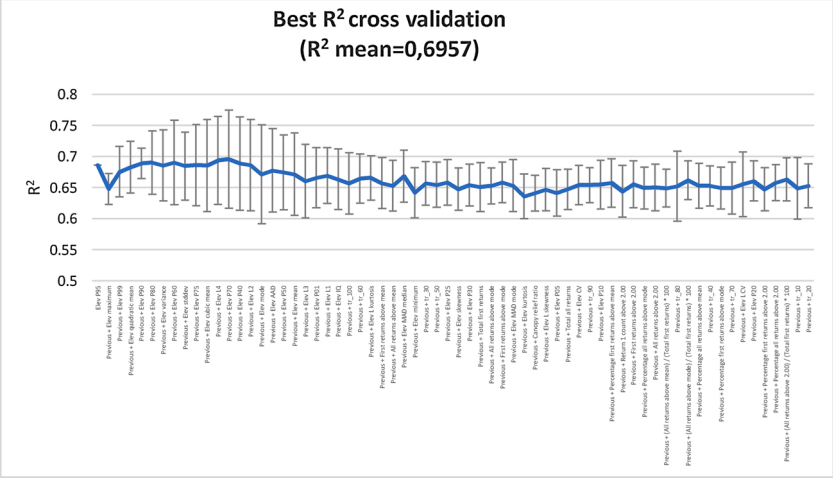


Fig. 7. R^2 mean of the two folds and standard deviation shown as error bar. The X-axis is read from left to right, with the first column being the most important variable according to the GINI Importance Index.

map was put together via photointerpretation of the 25 cm pixel aerial orthophotography of the Basque Government, and the minimum size of the tessera (homogeneous vegetation enclosure) is 0.10 ha, in line with the size of the NFI field inventory plots with radius 25 m. Each polygon includes a fraction cover for each forest specie, and so the final estimated biomass must take into account the percentage of occupation of the species in each polygon.

For each polygon of *P. radiata* species selected on this map, LiDAR variables described in Annex A were computed (as done using the plots to train the model), and this data is then used to predict the ln of the Biomass (ton/ha). The final biomass for each polygon (in ton units) is obtained by applying Eq. (3):

$$Biomass_{polygon(ton)} = e^{RF\_predicted\_biomass\left(\frac{ton}{ha}\right)} * polygon_{area\,(ha)} * fraction\_cover_{(unitless)} \quad (3)$$

5. Results

5.1. Variable importance

The most important variables, according to the GINI Importance Index, are those related to height, while density metrics computed in addition to the FUSION output showed less importance (Fig. 6). The variable with the maximum GINI IImportance Index is Elevation 95th percentile (*Elev P95*), which is a suitable variable l for modelling the height of the trees, followed by the maximum value of elevation (*Elev_maximum*) and 99th percentile (*Elev_P99*). Elevation-related variables follow until the 24th position, where the first density-related metrics start to appear (tr_100). Once variables have been ordered according to GINI importance, the results obtained from 2-fold cross-validation by inserting the variable in a forward direction ordered according to GINI importance index are shown in Fig. 7. In this figure, the mean of R^2 is shown as the main line of the graphic and the standard deviation of R^2 above and below the line as error bars. The maximum mean R^2 value (mean R^2 of 0.696) is obtained using the first 13 tree height related variables that are included *(Elev P95, Elev maximum, Elev P99, Elev quadratic mean, Elev P90, Elev P80, Elev variance, Elev P60, Elev stddev, Elev P75, Elev cubic mean, Elev L4 and Elev P70* variables), from which the mean R^2 starts decreasing (Fig. 6). Variables derived from density metrics were therefore not significant enough to be included in the random forest model. At this stage, a limited dataset containing only the selected 13 features were created in accordance with hyperparameter tuning, and this variable selection will enable a simple model to be obtained that may interpret, reduce overfitting and ultimately reduce the computational cost of training the model.

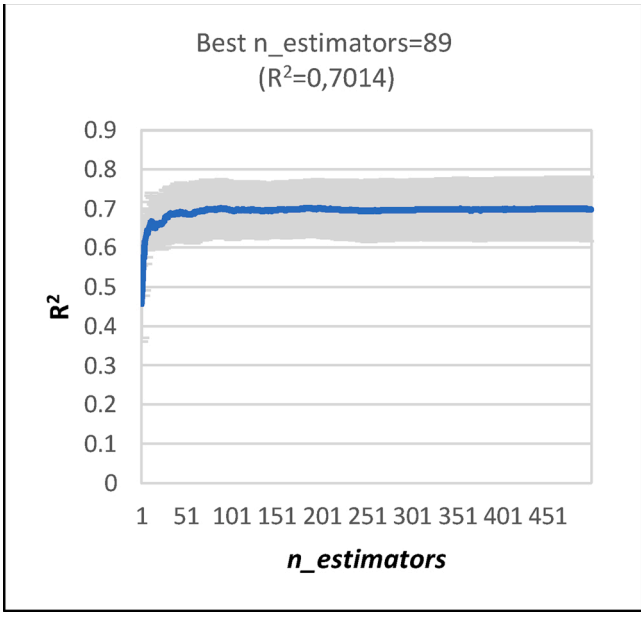


Fig. 8. Hyperparameter tuning results for *n_estimators*. The title of the plot gives the best value and R^2 score found.

5.2. Hyperparameter tunning

Figs. 8–12show the results of the exhaustive hyperparameter search for each hyperparameter across the ranges of the variables defined in Table 2. Note that the parameters not searched are defined by their default values as given in Table 2. The blue line indicates the mean determination coefficient of the 2-fold cross-validation for each variable value, while the standard deviation is drawn as an error bar. The title of each figure shows the parameter value with the highest mean R^2 value and score obtained for the ranged variable.

The *n_estimators* parameter evidenced a logarithmic curve increase in the approach to the asymptote close to the 70 trees, obtaining the best mean R^2 (0.7014) when its value is 89 (Fig. 8). A larger number of trees does not mean greater performance of the model, which increases the computational cost, and it is debatable whether the number of trees (*n_estimators* parameter) should simply be set to the largest computationally manageable value or whether a smaller value may in some cases be better. Authors have recently argued in favour of setting the number

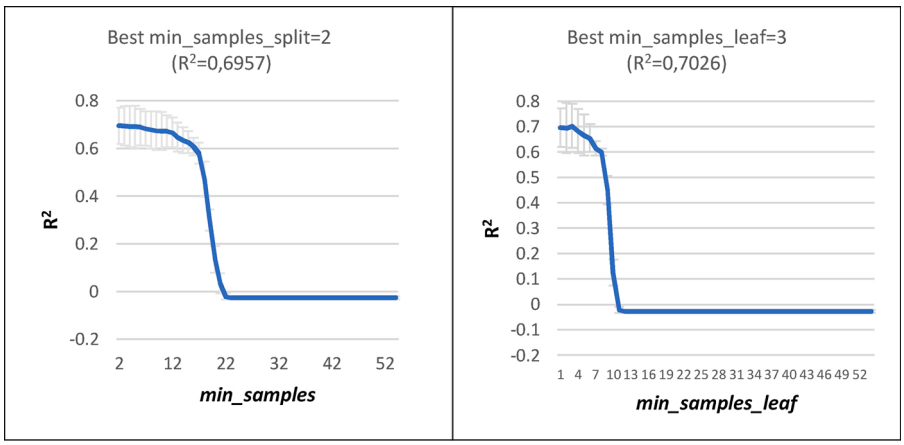


Fig. 9. Hyperparameter tuning results for *min_samples* and *min_samples_leaf*. The title of each plot gives the best value and R^2 score found.

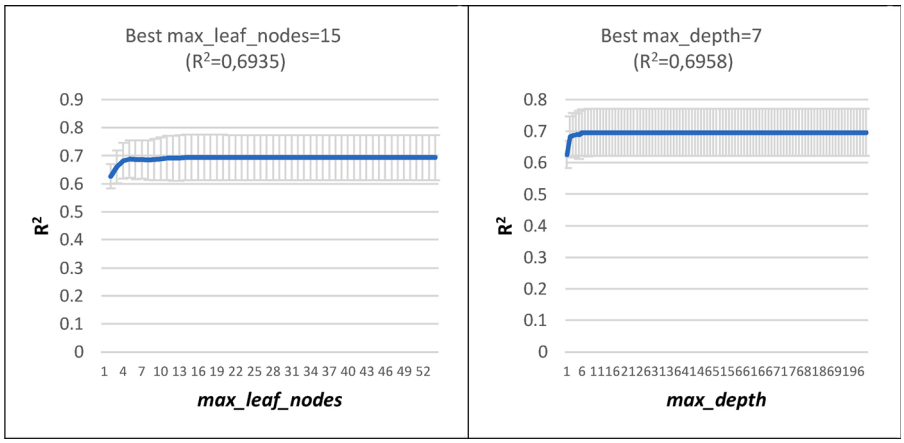


Fig. 10. Hyperparameter tuning results for *max_leaf_nodes* and *max_depth*.The title of each plot gives the best value and R^2 score found.

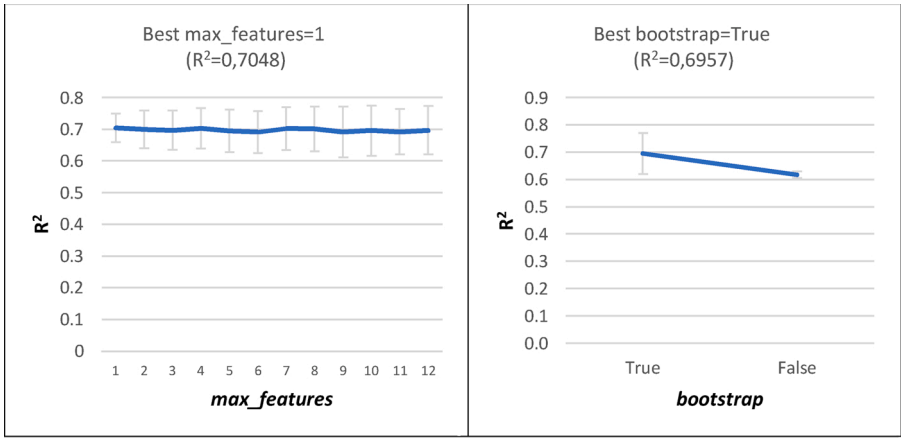


Fig. 11. Hyperparameter tuning results for *max_features* and *bootstrap*. The title of each plot gives the best value and R^2 score found.

of trees to a computationally feasible large number, depending on convergence properties of the performance measure desired, instead of tuning its parameter [52].

min_samples_split and *min_samples_leaf* hyperparameters are related, and both guarantee a minimum number of samples that need to be split, the first in internal nodes (with further splits) and the second in a leaf or external node (without any further splits). In fact, the *min_samples_leaf* guarantees a minimum number of samples in every leaf, no matter the value of min_samples_split. Both hyperparameters, therefore, follow the same distribution, evidencing higher performance when values are low (best values are 2 and 3, respectively for *min_samples_split* and *min_samples_leaf*) (Fig. 9), while a low number of minimum samples implies a deeper tree. The *max_leaf nodes* and *max_depth* also control the depth of the tree, and share a similar logarithmic distribution, reaching maximum performance when *max_leaf_nodes* is 15 and *max_depth* is 7, and maintaining performance when their value increases (Fig. 10).

The *max_features* parameter is one of the most critical parameters of RF models, as it affects the generalization error of the individual trees,

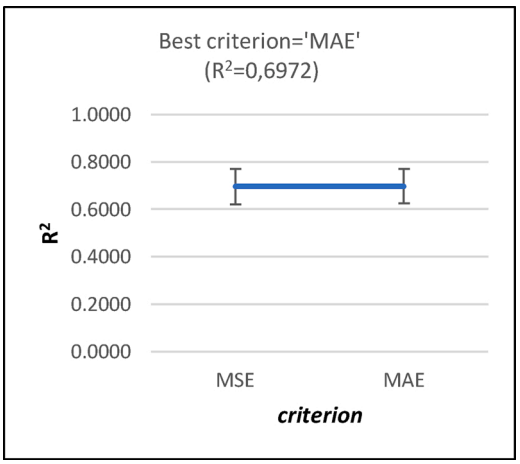


Fig. 12. Hyperparameter tuning results for *criterion.* The title of each plot gives the best value and R^2 score found.

Table 4
Values for the hyperparameters in the models developed and performance attained.

	Model	Model 1	Model 2
Parameters	*n_estimators*	89	85
	max_features	1	2
	max_depth	7	66
	min_samples_leaf	3	2
	min_samples_split	2	3
	max_leaf_nodes	15	29
	criterion	'MAE'	'MSE'
	Bootstrap	True	True
Cross-validation results	R^2 mean	0.708	0.726
	R^2 Std	0.055	0.086
	RMSE mean	0.289	0.276
	RMSE std	0.018	0.036

and hence the correlation between them [44]. Subsampling features is one of the essential properties of random forests and improves their performance by decorrelating the trees. A lower value of the *max_features* parameter allows the power of the individual trees to be decreased, but also injects a higher degree of randomness, thus improving final performance. The hyperparameter tuning for this parameter (Fig. 11) is quite stable across the range of the variable, with a maximum mean R^2 when the value is 1. Note especially that the standard deviation for the *max_features* = 1 is the lowest of all ranges. The benefit of random forests lies in creating a large variety of trees by sampling not only features, but also observations. For its part, the *bootstrap* parameter is a Boolean one: if it is True (default value), the observations are sampled with a replacement, and if it is False, the entire same training dataset is used to build each tree. Fig. 11 shows that the bootstrap = true option performs better ($R^2 = 0.6957$) than when it is not used ($R^2 = 0.6181$).

Finally, the criterion parameter that measures the quality of a split had an effect on the behaviour of the random forest regressor, and both criteria (Mean Squared Error –MSE- and Mean Absolute Error (MAE) performed equally (0.6972 vs 0.6957) (Fig. 12).

5.3. Model development

The hyperparameter tuning resulted in two RF models (Table 4). Model 1 was trained using the best hyperparameter derived from the exhaustive individual search (those obtained from the individual hyperparameter sensitivity analyses shown in Fig. 8 to 12), with the other parameters remaining fixed to the default value. Model 2 was obtained with the hyperparameters found using the randomized search with 100,000 iterations, in accordance with the variable ranges described in Table 3.

According to both hyperparameter tuning strategies, similar parameter values were obtained for *n_estimators (89 vs 85), max_features (1 vs 2)*, *min_samples_leaf (3 vs 2), min_samples_split (2 vs 3) and bootstrap* (both True) for Model 1 and Model 2, respectively. Greater differences were detected for *max_leaf_nodes* (15 vs 29) and especially for *max_depth* (7 vs 66) that forces the trees in Model 2 to be deeper than the ones in Model 1. The cross-validation resulted in a higher mean R^2 (0.726 vs 0.721) and lower RMSE (0.276 vs 0.289) for Model 2, albeit with higher standard deviation for both metrics, which suggests much better performance in one of the two folds when using Model 2 (Table 4). Although both models evidenced a reasonable R^2 (> 0.7) an F-statistic analysis was computed in order to statistically compare their ability to explain the variance in the dependent variable. According to this analysis, the null hypothesis that Model 1 is equal to Model 2 is thereby rejected (p-value < 0.05), and so we can state that Model 2 performs better than Model 1.

Table 5
Biomass estimation in the Orozko municipality compared to HAZI biomass figures.

	Biomass (ton)	Biomass HAZI (ton)	Difference (%)
Model 1	1063620.7	915851.6	16.13
Model 2	1084776.3	915851.6	18.44

5.4. Application of the RF models selected

Both models were applied to the NFI4 polygons digitized over the orthophotography and labelled as *P. radiata* in the Orozko municipality (detailed in 4.3.3) in order to compute the total biomass and compare it to those estimated by HAZI and considered official by the Basque Government. Both our models predicted a similar quantity of biomass (1.06 Mton for Model 1 and 1.08 Mton for Model 2), with only around a 20 K ton difference between them (2 %) and overestimated the official biomass quantity in a range between 16 % and 18 % (Table 5). Distribution of the RF biomass prediction was almost identical in the case of both selected models, with more than half of the area being focused on between 0 and 500 tons of biomass (Fig. 13).

The spatial distribution of the biomass predictions obtained for Orozko, via application of the model developed using the randomized search with 100,000 iterations is represented in Fig. 13.

6. Discussion

In this study a Random Forest model was developed in order to predict the biomass for *P. radiata* in a region of the Basque Autonomous Community (North of Spain) where this species has high cover. The *P. radiata* species continues to be one of the most coveted species in the wood sector, due to its physical-mechanical characteristics, although this has occurred later as a result of the need to establish an estimate of carbon sequestration of forests due to its relationship with climate change. *P. radiata* represents 60 of the species in Arratia-Nervión, where it is clearly predominant, with this region being very suitable for developing this study.

Accurate forest biomass estimation is the foundation of timber industry and forest management, and is the key to ecological research and the basis for a range of fields including forest productivity, energy flux, carbon and nitrogen cycling, nutrient cycling and forest dynamics. In addition, estimation of the amount of carbon stored in forests is a key challenge in understanding the global carbon cycle [53]. The usefulness

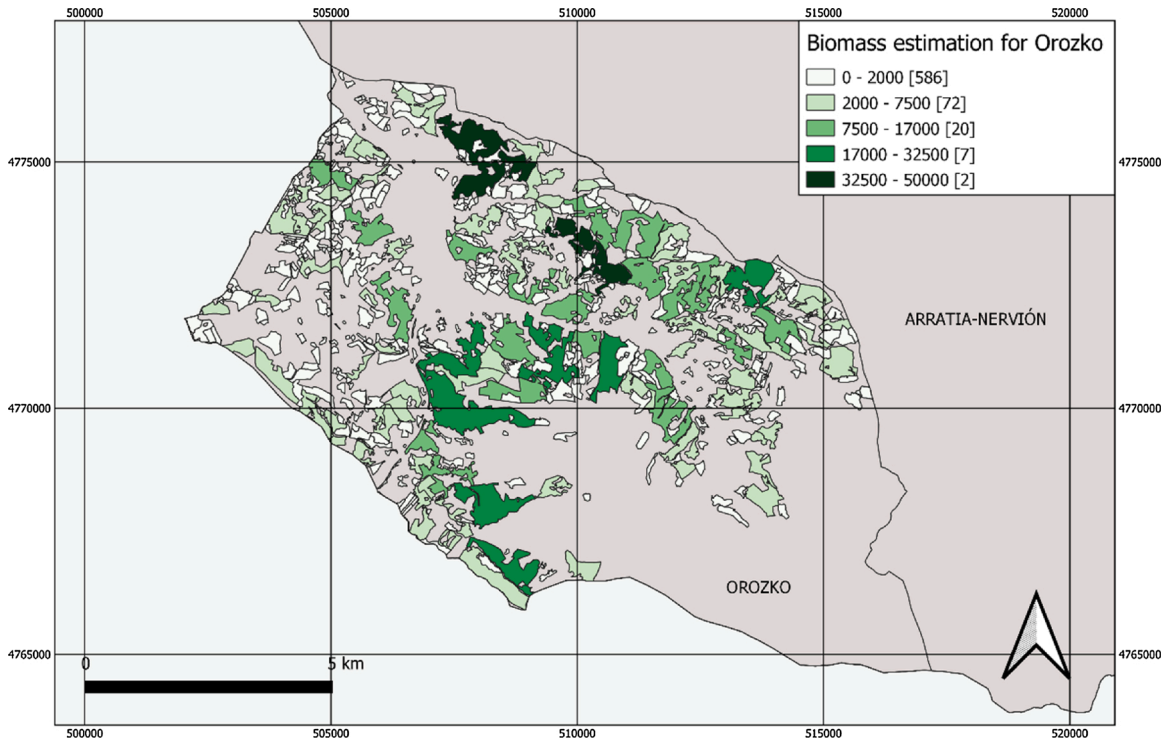


Fig. 13. Biomass estimations for the Orozko municipality (Arratia-Nervión) via application of the model developed using the randomized search with 100,000 iterations. The number of polygons included in each category are shown in brackets.

of LiDAR data in estimating forest characteristics has been largely proven [54], insofar as biomass has been estimated with very good results in different regions and with different species [55–57]. Although multivariate regression is the most popular approach, the complex relationships between forest variables are not always well captured by the models [58]. In order to overcome this disadvantage, and because the relevant literature reports better accuracy than linear regression techniques for biomass estimation [59], the potential of the RF estimation technique was thus the one researched in this study.

The dependent variable in this study was obtained using allometric equations fit by the National Forest Inventory statistics. Instead of using the allometric equations developed by Montero et al. [2] for the entire Spanish territory, those published by the HAZI foundation were used [33]. These equations were developed using destructive methods for 732 *P. radiata* specimens extracted from locations distributed across the Basque region, including the study area. Taking into account that Montero et al. took their sample entirely in Lugo (Galicia) for the species subject to study and that they used a smaller number of specimens (38), the allometric equation used in this study would seem to be more appropriate., on-site On-site allometric relationships can have a major influence in developing efficient and accurate biomass estimation models using remotely sensed data - even if regional scale allometry is suitable for biomass estimation across large areas, local estimation can ensure that biomass model accuracy reflects knowledge of forest biomass [21]. One of the main difficulties of this work was to obtain a model with the small number of samples available in our study to train and validate the models (N = 55), although this is a common situation in other reviewed works on the topic [19]. In forest applications, as in other fields such as medicine, data collection involves a major investment in terms of time and money, and this circumstance affected some decisions taken while the RF models were being developed. For example, only 2 folds were used in the cross-validation exercise, and so 27–28 samples were available for training / testing. We would also like to note the importance of using the shuffle option when splitting the samples into two folds, which is not the default option of the scikit-learn library employed in this work when creating the folds.

The selection of the most important variables and cross-validation over these 55 forest plot samples enabled the independent variables (originally 66) to be reduced to a subset of 13 variables, maintaining a high R^2. In contrast with previous work [40,55], the variables selected in our study derived exclusively from the height variable, while the density metrics (both computed by Fusion and the metrics added ad-hoc) were not found to be useful for the model. This simplification allows for a simpler computation of the variables so as to obtain a model that is more easily interpretable. Both hyperparameter search strategies used in this study (exhaustive search for each parameter and a random search with 100,000 iterations) produced similar parameter values. For example, using more than 89 trees was not shown to improve the result, while the computation effort increased. For their part, the minimum number of samples required to split an internal node or a leaf node obtained a higher score when using low values (less than 4), inducing more deep trees – the main difference between the two models developed was in fact the maximum depth of the tree. The randomized exploration of hyperparameter space was able to experimentally find the best combination of hyperparameters, with a higher R^2 (0.726) than the individual exhaustive search model (0.708), but also with a higher standard deviation, with its being noted that the results for both CVs were more heterogeneous. Even if both models shared similar statistics, the first model was able to produce fewer residual errors (RMSE mean 0.276 vs 0.289) and the difference was statistically significant. Both models, while using different parameter settings, provided similar outputs (only 20 K ton difference) in one of the municipalities within Arratia-Nervión, a region with a high density of the species subject to study. Compared to the estimations made by the HAZI Foundation, our models overestimated 16–18 % of the biomass. It is not easy to put these results in context with other studies, because most of them reported the root mean square error (RMSE) of the model, rather than biomass ton differences obtained in extensive test areas when applying different methodologies. However, the difference reported is in line with RMSE values provided by other authors [19,55].

If we compare the results obtained using RF techniques with those obtained applying Linear Model (LM) approaches, we can find other authors who have obtained very high values (>0.9) for R^2 using RF, better than those obtained in our study, although they have reported that this algorithm tends to overfit the data, while LMs had better predictive power and generalization features [12]. Gleason et al. [21], compared different machine learning approaches, such as RF, Support Vector Regression (SVR) and Cubist regression trees with LME techniques, in order to estimate biomass. Specifically in the case of the coniferous species, they obtained the best R^2 values using the LME approach (0.53), even if all the results were actually worse than those obtained in our study. In doing so, they used a specific tree delineation algorithm [60], which may have had an influence on the latter biomass estimation.

Low-density discrete LiDAR data was used in this study, with a point density of 0.5 points/m^2. This data was acquired under the National Plan of Aerial Orthophotography - LIDAR project designed mainly to produce the Digital Terrain Model of Spanish Territory, although it has also shown potential for accurate biomass estimation of *P. radiata*. In addition, the data is freely available, making it possible to extend such biomass estimation models to other regions, and also make a temporary estimation of biomass using data from the second LiDAR coverage, which started in 2015. A great variation in density values has been observed in literature, ranging from a point density similar to the one used in this study, to a high dense LiDAR point cloud, with more than thousand pts/m^2 gathered using an UAV. Theoretically, a higher point density should result in better biomass estimation, although this is not always the case in literature. For example, Ahmed et al., used a similar LiDAR point density (0.7 pts/m^2) in a study carried out in Canada, within an area containing conifer forest [20]. In their study they compared MLR to RF approaches in order to estimate canopy structure for mature, young and mature /young (combined) stands, and obtained values ranging from 0.59−0.72 for canopy cover; and 0.792−0.82 for canopy height. Also in Canada, albeit in the boreal zone, Matasci et al. estimated biomass values in the Canadian boreal forest, with a mean density of 2.8 returns per m^2 using the RF algorithm and obtaining a lower R^2 than those obtained in this study for R^2 ($R^2 = 0.52$) values of [15], due to the variability of the size of the trees and diversity in terms of forest types. Similarly, Dalla et al. reported a lower R^2 value (R^2 mean 0.47 and standard deviation of 0.187) using a UAV-Lidar very high point density (1500–2500 pt s/m^2) data on 17 ha of crop, livestock and seminal forest plantations of *Eucalyptus benthamii* by also applying the RF algorithm in southern Brazil, with a final very high point density of [22]. Conversely, Luo et al. estimated the biomass in China for young, middle-aged and near-mature forests, comparing the full-waveform derived metrics and traditional discrete-return metrics airborne LiDAR data with a pulse density significantly greater than those used in our study (4.1 pulses/m^2). In doing so, they obtained better results when modelling Random Forest using full-waveform derived metrics (R2 between 0.81−0.84) than for discrete-return metrics ($R^2 = 0.8$), albeit still higher than in our study [57,61]. The results obtained by González-Ferrreiro et al. [19] for *P. radiata* in Galicia (Spain), were very similar to ours ($R^2 = 0.74$, RMSE = 40.469 ton/ha) in an area also located in the north of the peninsula, although their point density was relatively high (8 points/m^2). In this case, MLR techniques were used to estimate biomass.

7. Conclusions

A Random Forest (RF) model was tuned to estimate the biomass for the *P. Radiata* species in a region of the Basque Country, Arratia-Nervión, where this species has high cover, using the National Forest Inventory, allometric equations and low-density discrete LiDAR data. Taking into account that this LiDAR data focuses on producing altimetric information, close agreement is achieved when applying cross-validation with 2 folds ($R^2 > 0.7$). The randomized search method used when finding optimal hyperparameters performed statistically better than that obtained by individual exhaustive hyperparameter searching, with the main difference between the models being the maximum depth that affected the depth of the trees involved in the RF algorithm. The models were applied in a municipality with more than 5000 ha of this species, and where a biomass of 1.06–1.08 Mton was predicted - a value between 16–18 % higher than the biomass predicted by the Basque Government.

For future research efforts, the combination of Airborne LiDAR data with data acquired by sensors orbiting in satellites such as those included in the Copernicus programme (Synthetic Aperture Radar Sentinel -1 and Optical sensor Sentinel-2 and) will be explored in order to obtain a more robust biomass estimation. Data acquired by NASA's Global Ecosystem Dynamics (GEDI) sensor from the International Space Station will also be explored in order to estimate evolution of the aboveground biomass in the Basque Country.

Authorship statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the *Hong Kong Journal of Occupational Therapy*.

CRediT authorship contribution statement

Leyre Torre-Tojal: Conceptualization, Software, Formal analysis, Writing - original draft. **Aitor Bastarrika:** Conceptualization, Software, Formal analysis, Writing - original draft, Supervision, Funding acquisition. **Ana Boyano:** Conceptualization, Formal analysis, Writing - original draft, Supervision. **Jose Manuel Lopez-Guede:** Conceptualization, Visualization, Writing - original draft. **Manuel Graña:** Conceptualization, Formal analysis, Writing - original draft, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work reported in this paper was partially supported by FEDER funds for the MINECO project TIN2017-85827-P and project KK-202000044 of the Elkartek 2020 funding program of the Basque Government. Additional support comes from grant IT1284-19 of the Basque Autonomous Community.

Appendix A

Table A1

Table A1
Metrics collection obtained with FUSION and PostGis (66 variables).

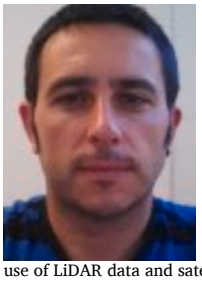
Variable	Description	Variable	Description
count	Number of returns above the minimum height	ccr	canopy relief ratio: ((mean - min)/(max – min)
densitytotal	total returns used for calculating cover	eqm	Elevation quadratic mean
densityabove	Returns above height break	ecm	Elevation cubic mean
densitycell	Density of returns used for calculating cover	r1count,…, r9count	Count of return 1,…,9 points above the minimum height
min	Minimum value for cell	rothercount	Count of other returns above the minimum height
max	Maximum value for cell	allcover	(all returns above cover height (h))/(total returns)
mean	mean value for cell	afcover	(all returns above cover h)/(total first returns)
mode	modal value for cell	allcount	number of returns above cover h
stddev	standard deviation of cell values	allabovemean	(all returns above mean h)/(total returns)
variance	variance of cell values	allabovemode	(all returns above h mode)/(total returns)
cv	coefficient of variation for cell	afabovemean	(all returns above mean h)/(total first returns)
cover	cover estimate for cell	afabovemode	(all returns above h mode)/(total first returns)
abovemean	proportion of first (or all) returns above the mean	fcountmean	number of first returns above mean h
abovemode	proportion of first (or all) returns above the mode	fcountmode	number of first returns above h mode
skewness	skewness computed for cell	allcountmean	number of returns above mean h
kurtosis	kurtosis computed for cell	allcountmode	number of returns above h mode
AAD	average absolute deviation from mean for the cell	totalfirst	total number of 1 st returns
p01,…,p99	1 st,…,99th percentile value for cell	totalall	total number of returns
iq	75th percentile minus 25th percentile for cell	tr10 to tr100	proportion of points from the total number of points contained above each layer

References

[1] Intergovernmental Panel of Climate Change, Climate Change 2001; Syntesis Report, 2001.
[2] G. Montero, R. Ruiz-Peinado, M. Muñoz, Producción de biomasa y fijación de CO2 por los bosques españoles, Monografías INIA, Serie Forestal, Madrid, 2005.
[3] E. Canga, I. Dieguez-Aranda, E. Afif-Khouri, A. Camara-Obregon, Above-ground biomass equations for Pinus Radiata d.dOn in Asturias, Forest Systems (INIA) 22 (3) (2013) 408–415, https://doi.org/10.5424/fs/2013223-04143.
[4] K. Johansen, S. Phinn, C. Witte, Mapping of riparian zone attributes using discrete return LiDAR, QuickBird and SPOT-5 imagery: assessing accuracy and costs, Remote Sens. Environ. 114 (11) (2010) 2679–2691.
[5] E. Næsset, Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data, Remote Sens. Environ. 80 (2002) 88–99, https://doi.org/10.1016/S0034-4257(01)00290-5.
[6] Y. Shi, T. Wang, A.K. Skidmore, M. Heurich, Important LiDAR metrics for discriminating forest tree species in Central Europe, J. Photogramm. Remote Sens. 137 (2018) 163–174, https://doi.org/10.1016/j.isprsjprs.2018.02.002.
[7] G. Shao, G. Shao, J. Gallion, M.R. Saunders, J.R. Frankenberger, F. Songlin, Improving Lidar-based aboveground biomass estimation of temperate hardwood forests with varying site productivity, Remote Sens. Environ. 204 (2018) 872–882, https://doi.org/10.1016/j.rse.2017.09.011.
[8] G. Vaglio Laurin, N. Puletti, Q. Chen, P. Corona, D. Papale, R. Valentini, Above ground biomass and tree species richness estimation with airborne Lidar in tropical Ghana forests, Int. J. Appl. Earth Obs. Geoinf. 52 (2016) 371–379, https://doi.org/10.1016/j.jag.2016.07.008.
[9] S. Magnussen, T. Nord-Larsen, T. Riis-Nielsen, Lidar supported estimators of wood volume and aboveground biomass from the Danish National Forest Inventory (2012–2016), Remote Sens. Environ. 211 (2018) 146–153, https://doi.org/10.1016/j.rse.2018.04.015.
[10] L.T. Ene, T. Gobakken, H. Andersen, E. Næsset, B.D. Cook, D.C. Morton, C. Babcock, R. Nelson, Large-area hybrid estimation of aboveground biomass in interior alaska using airborne laser scanning data, Remote Sens. Environ. 204 (2018) 741–755, https://doi.org/10.1016/j.rse.2017.09.027.
[11] S. Nie, C. Wang, H. Zeng, X. Xi, G. Li, Above-ground biomass estimation using airborne discrete-return and full-waveform LiDAR data in a coniferous forest, Ecol. Indic. 78 (2017) 221–228, https://doi.org/10.1016/j.ecolind.2017.02.045.
[12] K. Zhao, J. Suarez, M. Garcia, T. Hu, C. Wang, A. Londo, Utility of multitemporal lidar for forest and carbon monitoring: tree growth, biomass dynamics, and carbon flux, Remote Sens. Environ. 204 (2018) 883–897, https://doi.org/10.1016/j.rse.2017.09.007.
[13] J. Eitel, B. Höfle, L. Vierling, A. Abellán, G. Asner, J. Deems, K. Vierling, Beyond 3-d: the new spectrum of lidar applications for earth and ecological sciences, Remote Sens. Environ. 186 (2016) 372–392, https://doi.org/10.1016/j.rse.2016.08.018.
[14] S. Reutebuch, R. McGaughey, H. Andersen, W. Carson, Accuracy of a high-resolution lidar terrain model under a conifer forest canopy, Can. J. Remote. Sens. 29 (5) (2003) 527–535, https://doi.org/10.5589/m03-022.
[15] G. Matasci, T. Hermosilla, M. Wulder, J. White, N. Coops, G. Hobart, H. Zald, Large-area mapping of canadian boreal forest cover, height,biomass and other structural attributes using landsat composites and lidar plots, Remote Sens. Environ. 209 (2018) 90–106, https://doi.org/10.1016/j.rse.2017.12.020.
[16] M.B. Teixeira da Costa, et al., Beyond trees: mapping total aboveground biomass density in the Brazilian savanna using high-density UAV-lidar data, For. Ecol. Manage. 491 (2021) 119–155, https://doi.org/10.1016/j.foreco.2021.119155.
[17] E. Næsset, Estimating above-ground biomass in young forests with airborne laser scanning, Int. J. Remote 32 (2) (2011) 473.
[18] S.A. Hall, I.C. Burke, D.O. Box, M.R. Kaufmann, J.M. Stoker, Estimating stand structure using discrete-return lidar: an example from low density, fire prone ponderosa pine forests, For. Ecol. Manage. 208 (1–3) (2005) 189–209, https://doi.org/10.1016/j.foreco.2004.12.001, 2011.
[19] E. González-Ferreiro, U. Aranda, D. Miranda, Estimation of stand variables in Pinus radiata D. Don plantations using different LiDAR pulse densities, Forestry 85 (2) (2012), https://doi.org/10.1093/forestry/cps002.
[20] O. Ahmed, S. Franklin, M. Wulder, J. White, Characterizing stand-level forest canopy cover and height using landsat time series, samples of airborne lidar, and the random forest algorithm, Isprs J. Photogramm. Remote. Sens. 101 (2015) 89–101, https://doi.org/10.1016/j.isprsjprs.2014.11.007.
[21] C. Gleason, J. Im, Forest biomass estimation from airborne LiDAR data using machine learning approaches, Remote Sens. Environ. 125 (2012) 80–91, https://doi.org/10.1016/j.rse.2012.07.006.
[22] A.P. Dalla Corte, et al., Forest inventory with high-density UAV-Lidar: machine learning approaches for predicting individual tree attributes, Comput. Electron. Agric. 179 (2020) 105815, https://doi.org/10.1016/j.compag.2020.105815.
[23] K.E. Anderson, et al., Estimating vegetation biomass and cover across large plots in shrub and grass dominated drylands using terrestrial LiDAR and machine learning, Ecol. Indic. 84 (2018) 793–802, https://doi.org/10.1016/j.ecolind.2017.09.034.
[24] E.P. Baltsavias, Airborne laser scanning: basic relations and formulas, J. Photogramm. Remote. Sens. 54 (1999) 199–214.
[25] B. Lohani, S. Ghosh, Airborne LiDAR technology: a review of data collection and processing systems, Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. 87 (2017) 567–579, https://doi.org/10.1007/s40010-017-0435-9.
[26] R.O. Dubayah, J.B. Drake, Lidar remote sensing for forestry, J. For. 98 (6) (2000) 44–46, https://doi.org/10.1093/jof/98.6.44.
[27] R. Dubayah, J.B. Blair, S. Goetz, L. Fatoyinbo, M. Hansen, S. Healey, M. Hofton, G. Hurtt, J. Kellner, S. Luthcke, The Global Ecosystem Dynamics Investigation: High-Resolution Laser Ranging of the Earth's Forests and Topography, Sci. Remote Sens. 1 (2020), 100002, https://doi.org/10.1016/j.srs.2020.100002.
[28] I.C.O.N.A. Segundo Inventario Forestal Nacional. Explicaciones y Métodos, Spain, 1986–1995, Madrid, 1990.
[29] F. Bravo, M. Del Río, V. Pando, R. San Martín, G. Montero, I.Cañellas Ordoñez, El diseño de las parcelas del inventario forestal nacional y la estimación de variables dasométricas, fundación general de la universidad, 2002.
[30] D. Zianis, S.M. Seura, Biomass and Stem Volume Equations for Tree Species in Europe, Finnish Society of Forest Science, Finnish Forest Research Institute, 2005.
[31] M.T. Ter-Mikaelian, M.D. Korzukhin, Biomass equations for sixty-five North American tree species, For. Ecol. Manage. 97 (1) (1997) 1–24.
[32] D. Eamus, W. Burrows, K. McGuinness, Review of Allometric Relationships for Estimating Woody Biomass for Queensland, the Northern Territory and Western Australia, Australian Greenhouse Office, 2000.
[33] Ecuaciones de cubicación para el Pino radiata en el País Vasco, IKT / HAZI, Arkaute, 2004.
[34] M. Nilsson, Estimation of tree heights and stand volume using an airborne lidar system, Remote Sens. Environ. 56 (1996) 1–7, https://doi.org/10.1016/0034-4257(95)00224-3.
[35] E. Næsset, Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data, Remote Sens. Environ. 80 (2002) 88–99, https://doi.org/10.1016/S0034-4257(01)00290-5.
[36] E. Næsset, Effects of different flying altitudes on biophysical stand properties estimated from canopy height and density measured with a small-footprint

airborne scanning laser, Remote Sens. Environ. 91 (2004) 243–255, https://doi.org/10.1016/j.rse.2004.03.009.
[37] E. Næsset, t. Gobakken, Estimating forest growth using canopy metrics derived from airborne laser scanner data, Remote Sens. Environ. 96 (2005) 453–465, https://doi.org/10.1016/j.rse.2005.04.001.
[38] T. Gobakken, E. Næsset, R. Nelson, O.M. Bollandsås, T.G. Gregoire, G. Ståhl, S. Holm, H.O. Ørka, R. Astrup, Estimating biomass in Hedmark County, Norway using National Forest inventory field plots and airborne laser scanning, Remote Sens. Environ. 123 (0) (2012) 443–456, https://doi.org/10.1016/j.rse.2012.01.025.
[39] G. Goldbergs, S. Levick, M. Lawes, A. Edwards, Hierarchical integration of individual tree and area-based approaches for savanna biomass uncertainty estimation from airborne lidar, Remote Sens. Environ. 205 (2018) 141–150, https://doi.org/10.1016/j.rse.2017.11.010.
[40] L. Torre-Tojal, B. A.Bastarrika, J.M. Barrett, Sanchez Espeso, J.M. Lopez-Guede, M. Graña, Prediction of aboveground biomass from low-density LiDAR data: validation over P. Radiata data from a region north of Spain, Forests 10 (9) (2019) 819, https://doi.org/10.3390/f10090819.
[41] S. Nandy, R. Singh, S. Ghosh, T. Watham, S.P.S. Kushwaha, A.S. Kumar, V. K. Dadhwal, Neural network-based modelling for forest biomass assessment, Carbon Manag. 8 (4) (2017), https://doi.org/10.1080/17583004.2017.1357402.
[42] G.érard Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[43] L. Breiman, J.H. Friedman, R.A. Olshen, C.I. Stone, Classification and Regression Trees, Wadsworth, Belmont, Calif, 1984, https://doi.org/10.1201/9781315139470.
[44] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[46] H. Hong, G. Xiaoling, H. Yu, Variable selection using mean decrease accuracy and mean decrease Gini based on random Forest, 7th IEEE International Conference on Software Engineering and Service Science (ICSESS) (2016) 219–224, https://doi.org/10.1109/ICSESS.2016.7883053.
[47] S. Nembrini, I.R. König, M.N. Wright, The revival of the Gini importance? Bioinformatics 34 (21) (2018) 3711–3718, https://doi.org/10.1093/bioinformatics/bty373.
[48] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (20) (2020) 295–316, https://doi.org/10.1016/j.neucom.2020.07.061.
[49] R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye, Probabilidad y Estadística Para Ingeniería y Ciencias, novena ed., Pearson Educational, 2012.
[50] J.F. Kenney, E.S. Keeping, "Root Mean Square." §4.15 in Mathematics of Statistics, Pt. 1, 3rd ed., Van Nostrand, Princeton, NJ, 1962, pp. 59–60.
[51] E.J. Rykiel, Testing ecological models: the meaning of validation, Ecol Model 90 (1996) 229–244, https://doi.org/10.1016/0304-3800(95)00152-2.
[52] P. Probst, A.L. Boulesteix, To tune or not to tune the number of trees in Random Forest, J. Mach. Learn. Res. 18 (2018) 1–18.
[53] Dandan Xu, Haobin Wang, Weixin Xu, Zhaoqing Luan, Xia Xu, LiDAR applications to estimate forest biomass at IndividualTree scale: opportunities, challenges and future perspectives, Forests 12 (2021) 550, https://doi.org/10.3390/f12050550.
[54] R. Nelson, How did we get here? An early history of forestry lidar1, Can. J. Remote. Sens. 39 (sup1) (2013) S6–S17, https://doi.org/10.5589/m13-011.
[55] M. Maltamo, K. Eerikainen, J. Pitkanen, J. Hyyppa, M. Vehmas, Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions, Remote Sens. Environ. 90 (2004) 319–330, https://doi.org/10.1016/j.rse.2004.01.006.
[56] E. Næsset, T. Gobakken, Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser, Remote Sens. Environ. 112 (6) (2008) 3079–3090, https://doi.org/10.1016/j.rse.2008.03.004.
[57] S. Condés, D. Riaño, El uso del escáner láser aerotransportado para la estimación de la biomasa foliar del pinus sylvestris L. En Canencia (madrid), Cuadernos De La Sociedad Española De Ciencias Forestales. 19 (2005) 63–70.
[58] G. Chen, G.J. Hay, B. St-Onge, GEOBIA framework to estimate forest parameters from lidar transects, Quickbird imagery and machine learning: a case study in Quebec, Canada, Int. J. Appl. Earth Obs. Geoinf. 15 (2012) 28–37, https://doi.org/10.1016/j.jag.2011.05.010.
[59] S.L. Powell, W.B. Cohen, S.P. Healey, R.E. Kennedy, G.G. Moisen, K.B. Pierce, J. L. Ohmann, Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches, Remote Sens. Environ. 114 (2010) 1053–1068, https://doi.org/10.1016/j.rse.2009.12.018.
[60] C.J. Gleason, J. Im, A fusion approach for tree crown delineation from LiDAR data, Photogramm. Eng. Remote Sensing 78 (2012) 679–692, https://doi.org/10.14358/PERS.78.7.679.
[61] S. Luo, et al., Estimating Forest aboveground biomass using small-footprint full-waveform airborne LiDAR data, Int. J. Appl. Earth Obs. Geoinf. 83 (2019) 101922, https://doi.org/10.1016/j.jag.2019.101922.

Leyre Torre Tojal finished her studies as Technical Surveying Engineer in the UPV/EHU in 2002, two years later she obtained her Cartographic Engineering, Geodesics and Photogrammetry degree in the Polytechnic University of Valencia and finally, her Phd degree in 2016 in the University of Cantabria. She is associate professor, belonging to the Department of Mining Engineering and Metallurgy and Materials Science, at the School of Engineering of Vitoria-Gasteiz, UPV/EHU since 2007. Her investigation line is related to the estimation of the biomass using LiDAR data, and from this work 2 indexed publications has been achieved. She has participated in various R&D projects, as result, various indexed papers has been published.



Aitor Bastarrika holds a Surveying Engineering bachelor's by the University of the Basque Country (UPV/EHU) and a Geodesic and Cartographic Engineering Master's degree at the University of Alcala (Madrid), where he also completed his doctorate in the improvement of Burned Area Mapping algorithms. He currently holds a position as lecturer in geospatial techniques and Remote Sensing and Geographic Information Systems in the School of Engineering of Vitoria-Gasteiz, Basque Country, Spain. He is also responsible for the Master in Geoinformatics and Geospatial Analysis of the UPV / EHU and his research has focused on the designing of semi-automatic algorithms to map burned areas using satellite images and the use of LiDAR data and satellite images to extract diverse environmental information.

Ana Boyano studied degree in Industrial Engineering (2001) and finished her PhD in Mechanical Engineering (2016), in the University of the Basque Country, UPV / EHU. She is lecturer and researcher in the Department of Mechanical Engineering at the School of Engineering of Vitoria-Gasteiz, UPV/EHU since 2006. She also has 5 years of previuos work experience as a researcher and aeronautical test engineer in a technology center and has been participating in R&D projects at a regional and national, and international level since 2002. Her research field is characterization of composite materials under fracture and application of new materials in renewable energies. She is the author of a dozen indexed publications.

Jose Manuel Lopez-Guede received the M.Sc. degree in 1999 and the pH.D. degree in 2012, both in Computer Engineering at the University of the Basque Country (UPV/EHU), Spain. He got 3 investigation grants, and from 2000 to 2004 he worked at an Industrial Informatics company. Since 2004 he is working at University of the Basque Country (UPV/EHU), Spain. His current position is Assoc. Prof. with the Systems Engineering and Automatic Control Department at the Faculty of Engineering Vitoria-Gasteiz, Spain. His research interests are robotics and computational intelligence techniques applied to different areas as Energy, Robotics, etc.

Manuel Graña Romay received the M.Sc. and pH.D. degrees from Universidad del Pais Vasco (UPV/EHU), Donostia, Spain, in 1982 and 1989, respectively, both in computer science. His current position is a Full Profesor (Catedrático de Universidad) with the Computer Science and Artificial Intelligence Department of the Universidad del Pais Vasco (UPV/EHU). He is the head of the Computational Intelligence Group (Grupo de Inteligencia Computational). The research works in the group cover applications of computational intelligence to linked multicomponent robotic systems, reinforcement learning, medical image in the neurosciences, multimodal human computer interaction, remote sensing image processing, content based image retrieval, lattice computing, semantic modelling, data processing, classification, and data mining. He has been advisor for over 30 PhD Thesis, co-author of more than 150 journal papers. He is associated editor of Neurocomputing, Information Fusion, Computational Intelligence and Neurosciences, Intelligent Decision Technologies.