Tree Height-Diameter Allometry Based on TLS Point Clouds and Machine Learning Regression

Fernando J. Aguilar[a], Abderrahim Nemmaoui[a], Manuel A. Aguilar[a] & Alberto Peñalver[b]

[a]Department of Engineering, University of Almería, Ctra. de Sacramento s/n, La Cañada de San Urbano, 04120 Almería, Spain; faguilar@ual.es, an932@ual.es, maguilar@ual.es. [b]Faculty of Technical Education for Development, Santiago de Guayaquil Catholic University, Av. Carlos Julio Arosamena, Guayaquil 090615, Ecuador; alberto.penalver01@cu.ucsg.edu.ec.

Abstract

Most of the allometric models used to estimate tree aboveground biomass are based on tree diameter at breast height (DBH). However, it is difficult, if not impossible, to measure DBH from airborne remote sensors in order to increase the efficiency of forest monitoring programs. In this sense, it is common to draw upon traditional least squares linear regression models to relate DBH and others dendrometric variables that can be measured from airborne sensors such as tree height (h) and crown diameter (CD). This study explores the usefulness of ensemble-type supervised machine learning regression algorithms, such as random forest regression (RFR), categorical boosting (CatBoost), gradient boosting (GBoost) or AdaBoost regression (AdaBoost), as an alternative to linear regression (LR) for modelling the allometric relationships DBH = Φ(h) and DBH = Ψ(h, CD). The original dataset was made up of 2272 teak trees (*Tectona grandis* Linn. F.) belonging to three different plantations located in Ecuador. All teak trees were digitally reconstructed from terrestrial laser scanning point clouds. The results showed that allometric models involving both h and CD to estimate DBH performed better than those based solely on h. Furthermore, boosting machine learning regression algorithms (CatBoost and GBoost) outperformed RFR (bagging) and LR (traditional linear regression) models both in terms of goodness-of-fit (R^2) and stability (variations in training and testing samples).

Keywords: terrestrial laser scanning, machine learning regression, allometric models, teak plantations, forest inventory

1 Introduction

Forest ecosystems cover about 30% of our planet, contain 80% of the Earth's biomass and account for 75% of the gross primary productivity of the terrestrial biosphere [1].

In this way, they account for 50% of the annual carbon flux between the atmosphere and the Earth's land surface [2], thus contributing to fix atmospheric carbon up to rates of about 30% of the fossil fuel emissions [3]. In other words, they are extremely important for our planet, and one of the reasons why accurate models of forest vegetation are essential for the development of a sustainable bio-economy based on renewable resources [4]. This sequestration of atmospheric carbon by forests has turned into a major strategy for the United Nations Framework Convention on Climate Change within the context of Reducing Emissions from Deforestation and forest Degradation (REDD) in order to help mitigate greenhouse gas emissions on forest-rich developing countries [5].

Despite the increasingly need for forest monitoring, characterization of forest at tree level has been limited to traditional methods based on field inventory and aerial photography interpretation. However, field inventories are labor-intensive, time-consuming, and limited by spatial accessibility, while optical aerial photography does not directly provide 3D forest information [6, 7].

Remote Sensing (RS) technology can help to solve the aforementioned drawback as we have witnessed an exponential increase in RS datasets derived from different sources (satellites, aircrafts, UAV (Unmanned Aerial Vehicle)) at different resolutions based on different sensors (hyperspectral and multispectral cameras, LIDAR and SAR sensors, etc.) during the last decade. In fact, RS provides an exceptional source of data and powerful tools for monitoring forests dynamics at a variety of spatial and temporal resolutions [8, 9]. That is the case of terrestrial laser scanning (TLS), an efficient and non-destructive measurement method that is becoming a new paradigm based on a tree-centric approach to deal with 3D forest reconstruction at plot scale [10].

At the same time, parallel developments in Information Technology (IT) allow for the storage of very large datasets and their efficient processing. It has driven the development of many libraries and packages that implement supervised machine-learning algorithms to investigate phenomena by automatically creating regression (and classification) models from labeled datasets in a very efficient way. It makes it possible to use machine-learning methods in datasets derived from RS with the aim of increasing the level of automaticity in the extraction of valuable information [11, 12].

Current allometric models to estimate forest dry above-ground biomass (AGB) rely on stem diameter (diameter at breast height; DBH) and tree height (h) as key inputs [13, 14]. However, DBH cannot be measured directly from airborne or spaceborne sensors, which are the most suitable RS technologies to carry out wall-to-wall forest inventories by determining tree height [9]. In this regards, classical linear regression techniques, after a previous logarithmic transformation of the variables for adjustment, are usually applied to model allometric relationships between DBH (dependent variable) and predictor variables (h and crown diameter (CD)/crown area (CA)) [14–16]. Note that linear models are easily adjustable and as such usually offer satisfactory precision for allometric modeling of trees in natural forests [17].

Taking into account the above-mentioned background, this study uses TLS data collected at tree level in three teak plantations located in the Coastal Region of Ecuador (tropical dry forest) to compare the performance of several supervised machine-learning regression methods with respect to traditional linear regression for modelling the

local allometric relationships between DBH and h and CD. The underlying hypothesis is that learning-based models could outperform the results provided by traditional linear regression in the case of highly non-linear relationships found in tree allometry. These locally calibrated machine-learning based models could be used to improve forest AGB and carbon estimation, especially in large-scale inventories where only h and CD can be estimated from airborne or spaceborne sensors.

2 Material and Methods

2.1 Study area

The work area is located in the Coastal Region of Ecuador (Figure 1), comprising up to 58 planted teak (*Tectona grandis* Linn. F.) reference plots belonging to three different plantations: Morondava, El Tecal and Allteak.

The study site presents a rainfall ranging from 600 to 1600 mm (from south to north), with an average temperature of 24.4 ◦C. It belongs to the so-called Tropical Dry Forest, which is characterized by a very typical unimodal rainfall regime with a rainy period in the first quarter of the year and a marked drought during the rest of the year.

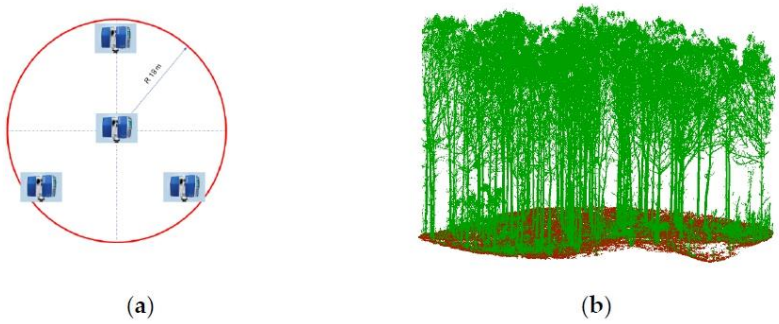
Fig. 1. Situation map of the three teak plantations (Morondava, El Tecal and Allteak) located in the province of Guayas, Ecuador.

2.2 Field data

A TLS field campaign was carried out in November 2018 (leaf-off conditions) over 58 reference plots of 18 m radius of even-aged teak trees belonging to three plantations located in the Coastal Region of Ecuador (Figure 1). Some key features of the reference plots can be found in [7].

A very dense and accurate TLS point cloud was obtained within each reference plot through a field survey carried out with a FARO Focus 3D X-330 TLS instrument, collecting x, y, and z coordinates together with high-resolution RGB images. The TLS was set to acquire data with medium resolution and quality (1/5 resolution and 4 quality) for potentially obtaining up to 28.2 million pulses per scan along a range from 0.6 to 330 m and a vertical and horizontal field of view of 300 and 360◦, respectively.

Considering that multi-scan data are more accurate to extract stem diameter and tree height than single scans [18], four scanning positions were set up within a radius of 18 m from the center of each reference plot to configure a scan pattern with a central scan and the rest located around, drawing an equilateral triangle (Figure 2a).



(a) (b)

Fig. 2. Configuration of scans positions within a reference plot and TLS derived point cloud. a) Circular reference plots showing the four TLS scans pattern. b) Automatically segmented TLS point cloud showing ground (brown) and vegetation (green) classified points.

The FARO Scene© 7.1 software was used to co-register the four scans within each reference plot, thus producing a single and colored 3D point cloud from using nine artificial targets (15 cm diameter spheres) conveniently distributed over the reference plot to ensure that at least three spheres were visible from every two consecutive scan positions.

A scattered bare earth points were obtained from automatically segmenting TLS ground points by applying the octree search algorithm implemented in the open-source software 3D Forest [19] (Figure 2b). These bare earth points allowed to build an accurate 20 cm grid spacing digital terrain model (DTM) as ground reference to compute the normalized heights of every point belonging to the TLS point cloud. 3D Forest software was also used to carry out the segmentation of forest vegetation into individual trees by applying an algorithm based on point clusters whose details can be found in [19]. Finally, an additional manual editing was needed to correct segmentation errors.

3791 teak trees were extracted from the 58 reference plots previously described. After a data filtering process to remove trees with DBH < 5 cm and/or CD less than 1 m (underdeveloped trees), 2272 trees were chosen as the dataset to develop the regression models (DBH = Φ(h) and DBH = Ψ(h, CD)) tested in this study.

A MATLAB code software called Tree_geometry was developed for the automatic extraction of the variables of interest DBH and h [16] from the point clouds corresponding to segmented trees, while CD was manually measured for each tree in the digital environment provided by Fusion/LDV software [20]. A complete description about the methods followed to obtain the dataset used in this study can be found in [16].

2.3 Allometric models

Six allometric models were tested to fit the DBH estimation from h and h + CD predictor variables; one based on traditional linear regression and the remaining five focused on supervised machine-learning algorithms.

The linear regression model used in this study was based on the widely known potential form (e.g. [14, 16]). After taking logarithms to linearize the potential expression, we obtain the following equations:

$$DBH = e^{(\alpha+\beta \ln(h))} e^{\varepsilon} = e^{(\alpha+\beta \ln(h))} e^{\frac{\sigma^2}{2}} \quad (1)$$

$$DBH = e^{(\alpha+\beta \ln(h.CD))} e^{\varepsilon} = e^{(\alpha+\beta \ln(h.CD))} e^{\frac{\sigma^2}{2}} \quad (2)$$

Where DBH is expressed in centimeters, and h and CD in meters. α and β are model coefficients, and ε is an error term, which we assume to be normally distributed with zero mean and standard deviation σ [13]. Under these conditions, the mean of e^{ε} can be approximated by $e^{\frac{\sigma^2}{2}}$, which can be understood as a correction factor applied to back-transform predicted values and remove bias from the log-transformed data.

Regarding supervised machine-learning methods, this study has focused on testing tree-based regression learners such as individual tree-based models (Decision Tree Regression: DTR) and some derive ensemble algorithms grouped in bagging techniques (Random Forest Regression: RFR) and boosting techniques (AdaBoost Regression: AdaBoost; Gradient Boosting Regression: GBoost; Categorical Boosting Regression: CatBoost). The optimal combination of hyperparameters for each machine-learning model was determined using grid search with cross-validation [21].

The validation of the tested allometric models relied on the so-called true validation method, meaning that the data used to train the model were never used for its validation. In this sense, the testing set for validation consisted of 20% of the 2272 available trees, leaving the remaining 80% as a set for training and computing the regression model. This procedure was repeated 100 times, splitting the original data between the training and the testing sets by using random sampling. It allowed studying the stability of the tested regression models against changes in the training samples.

The entire procedure mentioned above was coded in Python 3.8 with the support of the scikit-learn and catboost libraries.

3 Results and Discussion

Table 1 shows the statistics of goodness-of-fit (R^2) for the six tested allometric models in the case of only including h as explanatory variable for estimating DBH. Specifically, it represents the mean value, the standard deviation and the range of R^2 for the 100 repetitions performed, pointing out that individual tree-based models like DTR perform significantly worse ($p<0.05$) than linear regression or ensemble machine-learning regression algorithms. In fact, small changes in the learning sample can cause dramatic changes in the built tree derive from individual tree-based models, and so the estimated results can be unstable and inaccurate. This is the reason why most recent studies have adopted bagging and boosting ensemble algorithms [21, 22].

Traditional linear regression turned out to be very competitive, providing results statistically similar to those yielded by sophisticated ensemble boosting algorithms, while, surprisingly, RFR worked significantly worse than boosting or linear regression methods, showing a high variability in prediction when varying training samples. Note that ensemble learning is a branch of machine learning in which learning tasks are completed by building and combining multiple learners. In the case of bagging methods, such as RFR, they apply bootstrap samples randomly generated from the original dataset to train tree models and then aggregates the ensembles to obtain final predictions by majority voting. In this sense, the RFR algorithm usually improves predictions by decreasing the variance and avoiding overfitting, which is more recommended when developing models that include several explanatory variables (multivariate models).

Table 1. Statistics of R^2 for bivariate allometric models DBH = Φ(h). Mean values with different superscript letters in a column are significantly different ($p<0.05$) (two-sample t statistic).

Allometric model	*Mean R^2 (%)*	*Stand. deviation R^2 (%)*	*R^2 range (min/max. %)*
GBoost	87.21[a]	1.02	84.92 – 89.65
CatBoost	87.08[a]	1.06	84.80 – 89.43
LR	86.87[a]	1.08	83.17 – 89.32
AdaBoost	86.59[a]	1.26	83.16 – 89.80
RFR	82.50[b]	1.44	79.04 – 86.42
DTR	78.35[c]	2.23	72.93 – 83.76

Table 2. Statistics of R^2 for multivariate allometric models DBH = Ψ(h, CD). Mean values with different superscript letters in a column are significantly different ($p<0.05$) (two-sample t statistic).

Allometric model	*Mean R^2 (%)*	*Stand. deviation R^2 (%)*	*R^2 range (min/max. %)*
GBoost	90.16[a]	0.91	87.52 – 92.32
CatBoost	90.15[a]	0.93	88.00 – 91.98

AdaBoost	88.73[ab]	1.02	85.75 – 91.46
RFR	88.67[ab]	1.04	85.23 – 91.13
LR	87.81[b]	1.13	84.33 – 90.70
DTR	81.22[c]	1.67	76.31 – 86.01

The statistics of goodness-of-fit corresponding to the multivariate allometric models, in which DBH depends on h and CD, are shown in Table 2. First, it should be noted that the prediction results were clearly better than those provided by the bivariate allometric models presented in Table 1, especially in the case of machine-learning methods. Except linear regression, they also showed lower variability in R^2 when varying training samples, which points to a greater stability of the machine learning models tested in the case of multivariate regression than in the bivariate.

Quite the opposite occurred with traditional linear regression, where the inclusion of CD variable slightly improved the mean value of R^2, but also increased its standard deviation. This result indicates that machine-learning regression methods are able to identify complex relationships between covariates not found using conventional regression-based approaches. In this regards, GBoost has been qualified as the primary method for learning problems when dealing with noisy data and complex non-linear dependencies [23].

GBoost and CatBoost boosting regression algorithms performed significantly better ($p<0.05$) than traditional linear regression and DTR, also showing high stability to the variation of training samples. These similar results between GBoost and CatBoost were expected as CatBoost is a member of the family of gradient boosting decision tree machine-learning ensemble techniques.

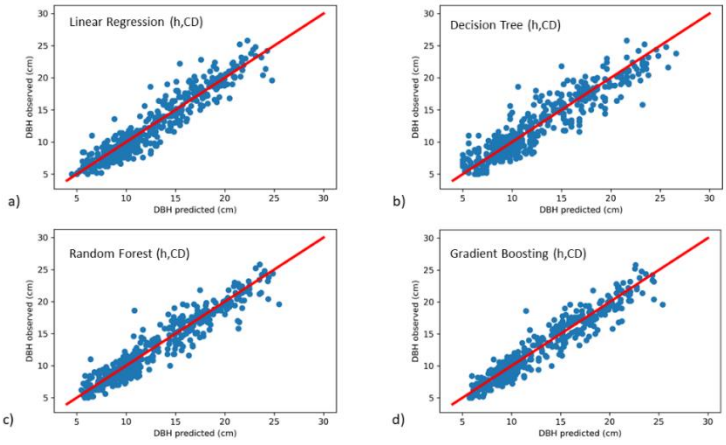


Fig. 3. Plots of predicted/observed values for DBH given by four allometric models in which the explanatory variables are h and CD.

AdaBoost and RFR were statistically situated between the very good results offered by GBoost and CatBoost and the good results offered by linear regression, providing predictions not significantly different from those provided by linear regression. In this way, boosting methods, such as GBoost, CatBoost and AdaBoost, are qualified as sequential ensemble algorithms that converts weak learners to strong learners by paying the most attention to the samples with the highest prediction errors, so increasing their weights in the next iteration and improving prediction accuracy by decreasing bias [23]. In any case, the bias of the multivariate models tested in this study was very low, as can be seen in Figure 3.

4 Conclusions

In this study, we tested several supervised machine-learning algorithms to model height-diameter allometry in teak plantations. The results obtained were compared with those provided by traditional linear regression, checking both bivariate models (DBH = Φ(h)) and multivariate models (DBH = Ψ(h, CD)). In this way, the allometric models that involved both h and CD to estimate DBH performed better than those based solely on h. Furthermore, boosting machine-learning regression algorithms (CatBoost and GBoost) significantly outperformed ($p<0.05$) individual tree-based model (DTR) and traditional linear regression model (LR), both in terms of goodness-of-fit (R2) and stability of regression models against changes in training samples. Random forest regression (ensemble bagging based algorithm) was statistically situated between the very good results offered by GBoost and CatBoost and the good results offered by linear regression, not achieving to significantly improve the predictions provided by LR.

The results obtained in this work demonstrate the great potential of supervised machine-learning regression methods to model complex nonlinear allometric relationships between DBH and two variables, such as tree height and crown diameter, which can be remotely sensed from spaceborne or airborne sensors. Without a doubt, it is a great step to facilitate the swift upscaling of plot-based field forest inventories to the immediate geographic area by applying remote sensing methods.

Acknowledgments

The research work reported here was funded by the following projects: 1) "*Enabling interdisciplinary COllaboration to Foster Mediterranean foREST sustainable management and Socio-ECOnomic valuation (ECO2-FOREST)*" (Proyecto Retos Junta de Andalucía, Spain, grant P18-RT-2327). 2) "*Evaluación de tecnologías de detección remota para la estimación de biomasa de teca en la Región Costa de Ecuador*" (Research and Development System of the Catholic University of Santiago de Guayaquil, Ecuador). This work also takes part of the general research lines promoted by the Agrifood Campus of International Excellence ceiA3, Spain (http://www.ceia3.es/en).

References

1. Pan, Y., Birdsey, R.A., Phillips, O.L., Jackson, R.B.: The structure, distribution, and biomass of the world's forests. Annu. Rev. Ecol. Evol. Syst. 44, 593–622 (2013). https://doi.org/10.1146/annurev-ecolsys-110512-135914.
2. Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M.A., Baldocchi, D., Bonan, G.B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K.W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F.I., Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. Science (80-.). 329, 834–838 (2010). https://doi.org/10.1126/Science.1184984.
3. Pan, Y., Birdsey, R.A., Fang, J., Houghton, R., Kauppi, P.E., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., Ciais, P., Jackson, R.B., Pacala, S.W., McGuire, A.D., Piao, S., Rautiainen, A., Sitch, S., Hayes, D.: A large and persistent carbon sink in the world's forests. Science (80-.). 333, 988–993 (2011). https://doi.org/10.1126/science.1201609.
4. Lindberg, E., Holmgren, J.: Individual Tree Crown Methods for 3D Data from Remote Sensing. Curr. For. Reports. 3, 19–31 (2017). https://doi.org/10.1007/s40725-017-0051-6.
5. Houghton, R.A., Nassikas, A.A.: Negative emissions from stopping deforestation and forest degradation, globally. Glob. Chang. Biol. 24, 350–359 (2018). https://doi.org/10.1111/gcb.13876.
6. Li, W., Guo, Q., Jakubowski, M.K., Kelly, M.: A New Method for Segmenting Individual Trees from the Lidar Point Cloud. Photogramm. Eng. Remote Sens. 78, 75–84 (2012). https://doi.org/10.14358/PERS.78.1.75.
7. Aguilar, F.J., Rivas, J.R., Nemmaoui, A., Peñalver, A., Aguilar, M.A.: UAV-Based Digital Terrain Model Generation under Leaf-Off Conditions to Support Teak Plantations Inventories in Tropical Dry Forests. A Case of the Coastal Region of Ecuador. Sensors. 19, 1934 (2019). https://doi.org/10.3390/s19081934.
8. Gómez, C., Alejandro, P., Hermosilla, T., Montes, F., Pascual, C., Ruiz, L.Á., Álvarez-Taboada, F., Tanase, M.A., Valbuena, R.: Remote sensing for the Spanish forests in the 21stcentury: A review of advances, needs, and opportunities. For. Syst. 28, 2171–9292 (2019). https://doi.org/10.5424/fs/2019281-14221.
9. White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., Tompalski, P.: Remote Sensing Technologies for Enhancing Forest Inventories: A Review. Can. J. Remote Sens. 42, 619–641 (2016). https://doi.org/10.1080/07038992.2016.1207484.
10. Liang, X., Kankare, V., Hyyppä, J., Wang, Y., Kukko, A., Haggrén, H., Yu, X., Kaartinen, H., Jaakkola, A., Guan, F., Holopainen, M., Vastaranta, M.: Terrestrial laser scanning in forest inventories. ISPRS J. Photogramm. Remote Sens. 115, 63–77 (2016). https://doi.org/10.1016/J.ISPRSJPRS.2016.01.006.
11. Gleason, C.J., Im, J.: Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sens. Environ. 125, 80–91 (2012). https://doi.org/10.1016/j.rse.2012.07.006.
12. Li, Y., Li, C., Li, M., Liu, Z.: Influence of Variable Selection and Forest Type on Forest Aboveground Biomass Estimation Using Machine Learning Algorithms. Forests. 10,

1073 (2019). https://doi.org/10.3390/f10121073.
13. Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M.S., Delitti, W.B.C., Duque, A., Eid, T., Fearnside, P.M., Goodman, R.C., Henry, M., Martínez-Yrízar, A., Mugasha, W.A., Muller-Landau, H.C., Mencuccini, M., Nelson, B.W., Ngomanda, A., Nogueira, E.M., Ortiz-Malavassi, E., Pélissier, R., Ploton, P., Ryan, C.M., Saldarriaga, J.G., Vieilledent, G.: Improved allometric models to estimate the aboveground biomass of tropical trees. Glob. Chang. Biol. 20, 3177–3190 (2014). https://doi.org/10.1111/gcb.12629.
14. Jucker, T., Caspersen, J., Chave, J., Antin, C., Barbier, N., Bongers, F., Dalponte, M., van Ewijk, K.Y., Forrester, D.I., Haeni, M., Higgins, S.I., Holdaway, R.J., Iida, Y., Lorimer, C., Marshall, P.L., Momo, S., Moncrieff, G.R., Ploton, P., Poorter, L., Rahman, K.A., Schlund, M., Sonké, B., Sterck, F.J., Trugman, A.T., Usoltsev, V.A., Vanderwel, M.C., Waldner, P., Wedeux, B.M.M., Wirth, C., Wöll, H., Woods, M., Xiang, W., Zimmermann, N.E., Coomes, D.A.: Allometric equations for integrating remote sensing imagery into forest monitoring programmes. Glob. Chang. Biol. 23, 177–190 (2017). https://doi.org/10.1111/gcb.13388.
15. Dalponte, M., Coomes, D.A.: Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. Methods Ecol. Evol. 7, 1236–1245 (2016). https://doi.org/10.1111/2041-210X.12575.
16. Aguilar, F.J., Nemmaoui, A., Peñalver, A., Rivas, J.R., Aguilar, M.A.: Developing allometric equations for teak plantations located in the coastal region of ecuador from terrestrial laser scanning data. Forests. 10, 1050 (2019). https://doi.org/10.3390/F10121050.
17. McRoberts, R.E., Westfall, J.A.: Propagating uncertainty through individual tree volume model predictions to large-area volume estimates. Ann. For. Sci. 73, 625–633 (2016). https://doi.org/10.1007/s13595-015-0473-x.
18. Pueschel, P., Newnham, G., Rock, G., Udelhoven, T., Werner, W., Hill, J.: The influence of scan mode and circle fitting on tree stem detection, stem diameter and volume extraction from terrestrial laser scans. ISPRS J. Photogramm. Remote Sens. 77, 44–56 (2013). https://doi.org/10.1016/j.isprsjprs.2012.12.001.
19. Trochta, J., Krůček, M., Vrška, T., Král, K.: 3D Forest: An application for descriptions of three-dimensional forest structures using terrestrial LiDAR. PLoS One. 12, e0176871 (2017). https://doi.org/10.1371/journal.pone.0176871.
20. FUSION/LDV LIDAR analysis and visualization software, http://forsys.cfr.washington.edu/fusion/fusion_overview.html, last accessed 2021/04/22.
21. Zhang, Y., Ma, J., Liang, S., Li, X., Li, M.: An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products. Remote Sens. 12, 4015 (2020). https://doi.org/10.3390/rs12244015.
22. Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., Sun, Y.: Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass. Forests. 12, 216 (2021). https://doi.org/10.3390/f12020216.
23. Hancock, J.T., Khoshgoftaar, T.M.: CatBoost for big data: an interdisciplinary review. J. Big Data. 7, (2020). https://doi.org/10.1186/s40537-020-00369-8.