Machine learning modeling exploration for under-bark tree bole volume estimation*

Maria J. Diamantopoulou[1,*,†]

[1] *Aristotle University of Thessaloniki, University Campus 54124, Thessaloniki, Greece*

Abstract

This paper investigates the potential of utilizing both probabilistic and ensemble supervised machine learning modeling strategies to accurately estimate under-bark tree bole volume. For this purpose, primary measurement data from pine trees (*Pinus brutia* Ten.) in the Seich–Sou suburban forest of Thessaloniki, Greece, were used. The described analysis can offer a strong foundation for understanding the performance of both non-parametric modeling approaches. Specifically, the study employed the probabilistic Gaussian Process Regression (GPR) modeling methodology with an integrated radial basis function (RBF) kernel. Furthermore, based on its well-known ability to predict values for continuous variables, the ensemble learning technique chosen for investigation was Random Forest regression (RFr), which integrates the bootstrap aggregation methodology. A cross-validation procedure, combined with an exhaustive grid-search methodology, was employed to determine the optimal hyperparameter combination for each constructed model. Despite the challenge of identifying the optimal combination of numerous hyperparameters unique to each modeling approach, the results demonstrated that both methodologies, due to their flexibility, have significantly strong potential to provide reliable under-bark tree bole diameters and volume estimations. This contributes to the sustainable management of forest resources and highlights potential areas for further exploration and improvement.

Keywords

Gaussian Process Regression, Random Forest regression, pine trees

1. Introduction

Accurately predicting the total volume of trees is crucial for anticipating forest growth and productivity. To estimate the bole volume by section, sophisticated formulas derived from the methods developed by Huber, Smalian, and Newton are employed [1]. These techniques necessitate multiple measurements of bole diameters at specific heights, which can be difficult to obtain from standing trees.

Directly measuring the under-bark diameters of a tree bole several meters above the ground, which is necessary for calculating the true under-bark bole volume, is unfeasible, as these measurements can only be obtained from a felled tree. To avoid this destructive method, alternative indirect approaches are being explored. Traditionally, regression analysis has been used to estimate various forest attributes. However, the standard regression methodology encounters difficulties due to the need to meet multiple assumptions [2].

Lately, the emerging field of artificial intelligence (AI), including machine learning (ML) techniques have shown great potential providing accurate estimations and predictions of biological attributes, even when dealing with noisy data and non-normal distributions, which are common in primary forest measurements. Over the past two decades, there has been increasing interest in utilizing machine learning in forestry [3, 4], driven by its advanced computational capabilities.

* *Short Paper Proceedings, Volume I of the 11th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA 2024), Karlovasi, Samos, Greece, 17-20 October 2024.*
* Corresponding author.
† These authors contributed equally.
mdiamant@for.auth.gr
0000-0002-6003-1285
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3930/paper15.pdf

In line with this objective, the goal of this study is to accurately estimate and predict the under-bark tree bole volume of pine trees using field measurements that are easily obtainable. To achieve this, two distinct machine learning approaches were employed: the probabilistic Gaussian Process Regression (GPR) method, known for its effectiveness in handling noisy continuous data, and the Random Forest regression (RFr) technique, an ensemble learning algorithm that enhances overall performance by combining the insights of multiple models.

2. Material and Methods

The ground-truth data was collected from measurements on pine trees (*Pinus brutia*) within the Seich-Sou suburban forest of Thessaloniki, Greece. This forest, covering an area of 3,085.82 ha with an elevation range between 563 meters and 100 meters [5]. Systematic sampling was employed to ensure that all different site classes were represented. Tree measurements included over bark (d_{oh}) and under bark diameters (d_{uh}) at one-meter height intervals starting from 0.3 meters above the ground ($d_{o0.3}$, $d_{u0.3}$, $d_{o1.3}$, $d_{u1.3}$, ..., $d_{o9.3}$, $d_{u9.3}$), as well as the total height (h) of the sampled trees. Upon completion of the measurements, a sample size of n = 999 measurements was obtained.

The under bark bole volume (v_{ubole}) was calculated using the Smalian's cross-sectional equation [1]:

$$v_{ubole} = \sum_{i=1}^{k} \left[\frac{\pi}{4} \cdot \left(\frac{d_{ui}^2 + d_{(ui+1)}^2}{2} \right) \cdot l \right] + \frac{\pi}{12} \cdot d_{uk}^2 \cdot l_k, \quad (1)$$

where d_{ui}, $i=1,...k$ are the under bark diameters of the lower and upper stem's sections in m, l is the length of each section in m, in this case equal to one meter, and l_k is the length of the tree top, in m, with $l_k < l=1$.

The mean and the standard deviation (std) for the observed over and under bark tree diameters, the tree total height and the under bark calculated volumes, are given in Table 1.

Table 1
Summary statistics of the observed tree bole diameters, in centimeters, total height, in meters and under bark calculated volumes, in cubic meters

diam	mean	std	diam	mean	std	diam	mean	std	diam	mean	std
$d_{o0.3}$	16.57	2.76	$d_{o3.3}$	8.88	2.79	$d_{o6.3}$	3.90	2.03	$d_{o9.3}$	1.99	1.55
$d_{u0.3}$	14.03	2.42	$d_{u4.3}$	8.41	2.59	$d_{u6.3}$	3.64	1.98	$d_{u9.3}$	1.82	1.47
$d_{o1.3}$	13.67	2.61	$d_{o4.3}$	7.01	2.45	$d_{o7.3}$	3.20	1.59	h	8.17	1.33
$d_{u1.3}$	12.16	2.36	$d_{u4.3}$	6.65	2.32	$d_{u7.3}$	2.95	1.56	v_{ubole}	0.05	0.02
$d_{o2.3}$	11.28	2.62	$d_{o5.3}$	5.13	2.23	$d_{o8.3}$	2.67	1.48			
$d_{u2.3}$	10.32	2.40	$d_{u5.3}$	4.83	2.15	$d_{u8.3}$	2.44	1.44			

2.1. Machine learning modeling approaches

Using a probabilistic supervised machine learning method like Gaussian process regression (GPR) [6] for estimating under bark bole volume (v_{bole}) brings significant benefits. This approach incorporates prior knowledge through kernels and provides uncertainty measures for predictions. Furthermore, this approach works well on small datasets, and it is more efficient in low dimensional spaces, matching perfectly in the present case study. Generally, GPR is characterized by the mean and covariance of the prior Gaussian process, along with the kernel that defines the relationship between two observations. In this context, the kernel radial basis function (RBF) was employed [7]:

$$k(x_i, x_j) = \sigma_m^2 \cdot e^{\left(-\frac{\|x_i - x_j\|^2}{2 \cdot ls^2} \right)}, \quad (2)$$

where σ_m^2 is the signal variance that controls the overall variance of functions drown from the Gaussian process regression, *ls* is the length scale, determines how rapidly the correlation between two points diminishes as the distance between them increases, $\|x_i - x_j\|^2$ is the squared Euclidean distance between the x_i and x_j.

In the equation (2), both the hyperparameter *ls* (length scale) and σ_m^2 (signal variance) are critical to the quality of the resulting model and must be properly optimized. To achieve this, the tree samples were randomly divided into a fitting data set, comprising 70% of the total data, and a testing data set with the remaining 30%. Additionally, the fitting data sets were subjected to k-fold cross-validation with k=5, ensuring the constructed model's predictive ability is adequate. The same data division approach was applied to the Random Forest regression model construction, as well.

The second non-parametric approach chosen was the RFr, selected in part for its ability to bypass the assumptions inherent in standard regression modeling. This technique is recognized as a robust non-parametric, supervised machine learning algorithm, originally proposed by [8]. The concept behind this approach is that combining multiple models can better capture the true structure of the data. RFr employs multiple individual models, called decision trees, which are combined into a single model. The goal is to minimize both the variance and bias of the base model—the decision tree—as much as possible within the system.

The successful training of the RFr model significantly depends on fine-tuning its hyperparameters, particularly the number of decision trees (n_{dt}), known as learners, and the maximum depth (d_{max}) of these learners. These hyperparameters are crucial as they govern the complexity of the RFr model. The RFr training utilized the bootstrap aggregation algorithm, commonly known as bagging [8, 9].

Both the machine learning methodologies were implemented in the scikit-learn libraries [10] and the Python programming language [11].

2.2. Evaluation criteria

The evaluation criteria crucial for assessing the suitability of the machine learning models used in this study were as follows: a) root mean square error (*RMSE*), which calculates the square root of the average squared differences between estimated/predicted and observed values; b) the coefficient of determination (R^2), which reflects the proportion of variance in the dependent variable that can be explained by the independent variables; c) bias (*BIAS*), representing the mean difference between estimated/predicted and observed values; and d) relative sum of square errors (*RSSE*), which is the (%) ratio of the sum of squared errors (SSE) to the sum of the actual values of the under-bark bole volume values. High model performance is indicated by low *RMSE, BIAS*, and *RSSE* values, coupled with high R^2 values.

3. Results

Taking into account the difficulty faced in obtaining tree bole diameters in different heights, the variables used as input variables to the under bark volume machine learning systems with output variable the under bark bole volume (v_{ubole}) were the diameters located near the ground, therefore easy to be measured, which were the ($d_{o0.3}$), ($d_{u0.3}$), ($d_{o1.3}$), ($d_{u1.3}$) and the total height (h) of the trees. Moreover, these variables produce high correlation with the (v_{ubole}) values, contributing mostly to the (v_{ubole}) values configuration.

Employing both machine learning Gaussian process regression modeling, and Random Forest for regression modeling, the required hyperparameters were assessed using the grid-search methodology [12], which resulted to the optimal hyperparameters' values presenting in Table 2.

Table 2
Optimal hyperparameters values for both modeling approaches

Gaussian process regression (GPR)			Random Forest for regression (RFr)		
hyperparameters	range	optimal value	hyperparameters	range	optimal value
σ_m^2	0 - 1	0.05	n_{dt}	1 - 300	10
ls	1 - 5	1.1	d_{max}	1 - 10	7

The evaluation criteria for the constructed models are presented in Table 3. As indicated in the table, both models yield similar outcomes. However, the GPR model provides the most accurate and reliable results for both the fitting and testing datasets.

Table 3
Evaluation criteria for both the constructed (GPR) and (RFr) modeling approaches, for both fitting and testing data sets

	data	criteria			
models	set	*RMSE*	R^2	*BIAS*	*RSSE%*
GPR	fitting	0.0026	0.988	-0.00002	0.0141
	testing	0.0032	0.977	-0.00009	0.0233
RFr	fitting	0.0028	0.986	-0.00005	0.0163
	testing	0.0038	0.974	-0.00136	0.0319

The performance of both constructed models was further assessed through the 45-degree line plots.

4. Discussion

As a Bayesian regression technique, GPR modeling offers a probabilistic approach to inference, enabling the prediction of not just the expected value of a target variable but also the uncertainty associated with that prediction.

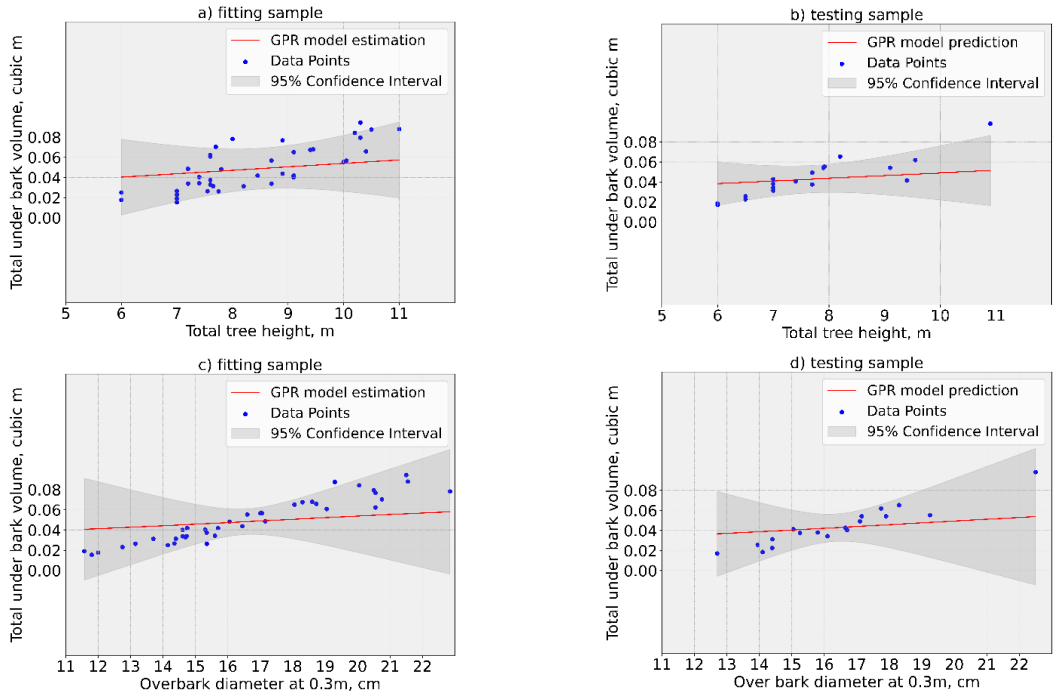

Figure 1: GPR model performance associated by its uncertainty

Offering a probabilistic prediction with a mean and variance provides a natural measure of uncertainty in the predictions. Indicatively, the uncertainty in the under bark bole volume predictions against the total tree height and the stump diameter (the tree bole diameter located at 0.3 m from ground) is shown in Figure 1. Similar plots under similar uncertainty could be produced for all predictors. This evaluation is particularly useful in forestry, where risk assessment is essential for the effective implementation of sustainable forest management.

The flexible structure of the Random Forest algorithm helps prevent the serious issue of overfitting and enables the system to handle real-world data, which often includes challenges such as high variance, outliers, and missing values. However, it's important to note that the further a predicted value is from the range of the fitting data, the less reliable that prediction will be.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. E. Avery, H. E. Burkhart, Forest Measurements, Mc Graw Hill, New York, NY, 2002.

[2] N. R. Draper, H. Smith, Applied regression analysis, 3rd ed., Wiley, New York NY, 1998. doi:10.1002/9781118625590.

[3] M. J. Diamantopoulou, R. Özçelik, H. Yavuz, Tree-bark volume prediction via machine learning: A case study based on black alder's tree-bark production, Comput Electron Agric 151(2018): 431-440. doi: 10.1016/j.compag.2018.06.039.

[4] S. S. Ghosh, U. Khati, S. Kumar, A. Bhattacharya, M. Lavalle, Gaussian process regression-based forest above ground biomass retrieval from simulated L-band NISAR data. Int J Appl Earth Obs Geoinf 118(2023) 103252. doi: 10.1016/j.jag.2023.103252.

[5] FILOTIS - Database for the Natural Environment of Greece. URL: https://filotis.itia.ntua.gr/biotopes/c/AT4011119/.

[6] CE. Rasmussen, CKI Williams, Gaussian Processes for Machine Learning, The MIT Press, Massachusetts, 2006.

[7] W. Chen, H. Wang, Q. H. Qin, Kernel Radial Basis Functions, in Computational Mechanics, Springer, Berlin, Heidelberg, 2007. doi: 10.1007/978-3-540-75999-7_147.

[8] L. Breiman, Random Forests, Machine Learning 45(2001): 5–32. doi: 10.1023/A:1010933404324.

[9] A. M. Prasad, L. R. Iverson, A. Liaw, Newer Classification and Regression Techniques: Bagging and Random Forests for Ecological Prediction, Ecosystems 9(2006): 181-199. doi: 10.1007/s10021-005-0054-1.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, et al., Scikit-learn: Machine Learning in Python, J Mach Learn Res 12(2011): 2825-2830. doi: 10.48550/arXiv.1201.0490.

[11] Python Software Foundation: Python Documentation, 2022. ULR: http://www.python.org/.

[12] S. M. LaValle, M. S. Branicky, S. R. Lindemann, (2004). On the relationship between classical grid search and probabilistic roadmaps, The International Journal of Robotics Research 23(2004): 673–692.