

# Tree bark prediction along the bole through the support vector regression technique

Maria Diamantopoulou<sup>1,\*</sup>

<sup>1</sup>School of Forestry and Natural Environment, Aristotle University of Thessaloniki, 54125 Thessaloniki, Greece

**Abstract.** Tree bark plays a protective role by surrounding the wood of a tree like a cloak. Due to its chemical composition and the possibility of its use in various fields, such as pharmaceuticals, landscape architecture, etc., tree bark receives much attention having outstanding importance for industrial utilization and markets. Tree bark is considered a valuable forest product, along with the wood volume. Thus, the accurate prediction of the bark quantity that a tree can produce is of utmost importance for the sustainable management of the forests. For this reason, the knowledge of its quantities, further enables the accurate prediction of the plain wood volume that can be produced by the forest, as well. Because of the nonlinear nature of this biological variable, its accurate quantification is a very complicated problem. Artificial intelligent methods have shown the potential to reliably predict biological variables that are non-linear in nature. In this work, the support vector regression methodology, as a nonlinear nonparametric machine learning approach, is tested for the accurate prediction of the tree bark factor in every different height of the tree bole, through easily obtained measurements on trees.

## 1 Introduction

Pine trees (*Pinus brutia* Ten.) have important characteristics, as far as drought tolerant and fast growth are concerned. Furthermore, they are coniferous species which are native in the Mediterranean region [1]. They can also be found in planted areas, such as parks and urban forests in the same region. Pine trees belong to these species that can reach high bark thickness with ageing (Fig. 1), with mean value of about the 15% of its over-bark wood bole volume [2].



**Fig. 1.** Pine bark of about forty years of age tree.

Beyond to its known protective role in the survival of the tree, bark derived from different tree species can be utilized as a pollution bioindicator [3-5], as well. Due to bark chemical composition [6] it can be used as a raw material for the manufacture of pharmaceutical substances, as a soil conditioner along with groundcover product in garden architecture [7] and finally, as an alternative source of energy. These are the main reasons

why it is essential for forest management decisions to rely on the accurate prediction of the bark quantities that a forest can produce, even in the case that tree bark can be considered just as remnant. According to previous studies [8] inaccurate predictions in bark volume can significantly and negatively influence the profit that can be derived from a forest area.

Since the variation of bark thickness along the tree bole, is a reality, accurate bark factor ( $BF_i$ ), which represents a bark thickness ratio, values prediction through models, can significantly contribute to the accurate and reliable diameters bark percentage ( $b_{di}\%$ ) prediction for each diameter ( $d_i$ ), leading to the accurate prediction of the tree bole volume along with the tree bark volume of the trees. To this direction, in order for the bark volume of the trees to be accurately predicted, many different modeling methodologies have been applied till today, mostly including linear and non-linear regression models [9-14]. Lately, in the scientific area of forest research, attention is paid to the machine learning methodologies, that due to their nonparametric algorithms, are free of assumptions beset the regression modeling methodology [15]. Furthermore, there is no need for a specific model form to be pre-specified, as it is essential in regression modeling; instead, in the area of the artificial intelligence, the learning methodology applied shows the ability to learn from noisy and, frequently, incomplete data, such as the ground-truth forest data. Till now, some attempts have been made, in order for the tree bark quantities to efficiently be modeled through the usage of the machine learning strength. That is, [7] compared the performance of artificial neural networks with non-linear regression

\* Corresponding author: [author@e-mail.org](mailto:author@e-mail.org)

models, while [16] compared the performance of the support vector machine for regression with the same technique, in both works the comparisons made with the use of pine and black alder trees bark volume prediction, respectively. [17] used nonlinear autoregressive exogenous neural network technique (recurrent neural network) for estimating the double bark thickness of oak and Scots pine trees. All the above attempts concluded that the machine learning approach has the ability to adequately describe the patterns of the preliminary data in hand. Furthermore, from the comparison of the machine learning methodologies with more traditional approaches, such as the regression analysis, the first showed superior performances than the latter.

Following the necessity of reliable models production that can positively influence the accurate prediction of the tree bark quantities in the forest, and taking into account that the bark factor ( $BF$ ) is one of the most important variables that directly affect the bark volume configuration, this study aims to explore, test and finally provide an efficient modeling approach, from the new perspective of the machine learning modeling, hoping that the current approach has the potential to lead to the most accurate tree bark quantities prediction.

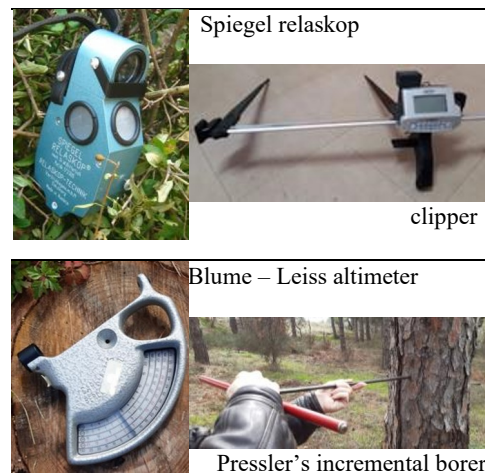
## 2 Materials and methods

### 2.1 Study area and data collection

In order for the bark factor ( $BF$ ) to be accurately predicted by an efficient model, a stem analysis dataset was created including several measurements, namely a total of 1474 data on fifty-seven standing pine trees from Seich-Sou urban forest of Thessaloniki, Greece, were taken. Urban forest of Thessaloniki lays in the area that is included by the geographic coordinates  $40^{\circ}37'33.0''N$  and  $23^{\circ}00'45.0''E$ . It is an almost purely planted pine forest, along with some other species that can be found in the forest, as well, such as trees of *Cupressus sempervirens* L., *Cupressus arizonica* Greene, *Quercus coccifera* L, etc. [18].

In an effort for the measurements to be taken from trees of all different site qualities and classes in order for all the possible variation of the trees attributes to be covered, systematic sampling was used. A GPS instrument was used for the sampled trees to be located on the ground.

Tree measurements included over bark diameter at stump height ( $d_{0.3}$ ) (0.3 meters height from ground), diameter at breast height ( $d_{1.3}$ ) (1.3 meters height from ground), over bark diameters ( $d_i$ ) at one-meter height intervals above breast height until the tree tip, double bark thickness ( $b_i$ ) in all relative heights ( $h_i$ ) where the over bark diameters were measured, and total height ( $tht$ ) of the sampled trees. The instruments used for obtaining the measurements on the tree boles were the caliper, the Spiegel relaskop, the Blume – Leiss altimeter, and the Pressler's incremental borer [2, 19] (Fig. 2).



**Fig. 2.** Instruments used for obtaining the measurements on the tree boles.

### 2.2 Comprehensive data base construction

Different values of the bark factor ( $BF_i$ ) were produced in different heights of the trees boles:

$$BF_i = d_{oi} / d_{ui} \quad (1)$$

where  $d_{oi}$  is the over-bark diameter at  $i$  meters from ground and  $d_{ui}$  is the under-bark diameter at the same bole height ( $h_i$ ).

The values of the under-bark diameters ( $d_{ui}$ ) were calculated as:

$$d_{ui} = d_{oi} - b_i \quad (2)$$

where  $b_i$  is the double bark thickness at the tree bole height ( $h_i$ ).

In order to predict the bark factor profile over the tree bole, the relative height ( $RH_i$ ) of the tree bole that the measurements were taken, is also calculated:

$$RH_i = h_{di} / tht \quad (3)$$

where  $h_{di}$  is the tree bole height ( $h_i$ ) where the diameter measurement ( $d_i$ ) was taken and  $tht$  is the total height of the tree bole.

The resulting values of the  $\varepsilon$ -SVR estimation and prediction system are essential for the calculation of the bark percentage of any diameter along the tree stem, while in this way, it is possible to accurately calculate the pure (without bark) tree stem volume for any tree stem section and the relative tree bark volume by applying the following system of equations [2]:

$$v_{ui} = v_{oi} / BF_i \quad (4)$$

$$v_{bi} = v_{oi} \cdot (1 - (1/BF_i)) \quad (5)$$

where,  $v_{ui}$  is the under-bark volume of the tree stem section  $i$ ,  $v_{oi}$  is the over-bark volume of the tree stem section  $i$ , and  $v_{bi}$  is the bark volume of the tree stem section  $i$ .

Having the quantities of equations (4) and (5), the total tree stem pure volume (under-bark) and the total

bark volume can be assessed by summing all the above portions of the stem.

Arithmetic mean values, maximum and minimum values and the standard deviation (Sd) of the basic variables used as the ground-truth information for the bark factor modeling configuration, are given in Table 1.

**Table 1.** Descriptive statistics of the field measurements and the variables used for the bark factor modeling configuration.

feature	mean	min	max	Sd
$d_{0.3}$ , cm	16.56	11.58	22.85	2.76
$d_{0.1.3}$ , cm	13.67	7.13	19.10	2.61
$b_i$ , cm	0.81	0.08	3.31	0.036
$tht$ , m	8.36	6.00	11.00	0.061
$BF$	1.10	1.03	1.40	0.002
$RH$	0.48	0.03	0.99	0.013

In order the predictive ability of the constructed model to be tested in new, never seen by the model in its construction phase, data, the available data set was divided randomly into two parts. That is, the fitting data set consists of the 90% of the total data in hand, and the test data set consists of the remaining 10%.

### 2.3 Support vector regression model (SVR) construction

Nowadays, artificial intelligence (AI) is examined in order for its effectiveness to be verified in environmental and more specifically, in forest modeling [20, 21]. As a subset of AI, the machine learning methodologies include the Support Vector Regression (SVR) technique [22]. The outstanding characteristic of this methodology is its generalization ability. Furthermore, its potential of learning from noisy or incomplete data by detecting inherent complex nonlinear relationships between output and input variables, is a challenging ability, that worths exploration in forest modeling research [16, 23-25], where such kind of data are encountered. Specifically, the  $\epsilon$ -SVR algorithm used in the present modeling effort, which is a supervised learning algorithm, attempts to find a function  $f(x)$  from these data pairs between inputs and output, so as the regression error of all training samples to be minimized by lying within an optimum width of the  $\epsilon$ -insensitive zone  $[-\epsilon, \epsilon]$ . In order for intricate patterns in the data to be captured by the algorithm, non-linear Radial Basis Function (RBF) kernels have been used. The RBF Kernel Support Vector regression was implemented in the scikit-learn libraries [26] and the Python programming language [27]. It is worth noting that a set of the optimum values of three meta-parameters guides to the successful learning of the algorithm. These meta-parameters are the  $\epsilon$  which controls the width of the  $\epsilon$ -insensitive zone, the gamma parameter ( $\gamma$ ) which is inversely proportional to the variance ( $\sigma$ ) and used for the RBF kernel width control, and the cost parameter ( $C$ ) which is a regularization parameter, known as the penalty parameter, used in order to specify the trade-off between mis-prediction against simplicity of the model [16]. Each one of them separately and interactively at the same time, controls

the adequate learning of the system under the kernel equation:

$$K(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2), \gamma > 0 \quad (6)$$

where  $x_1, x_2$  are two points or Support Vectors (SV) with an Euclidean ( $L_2$ -norm) distance between them ( $\|x_1 - x_2\|$ ),  $\gamma = (1 / 2\sigma^2)$  and  $\sigma^2$  is the variance.

The bark correction factor ( $BF_i$ ) was used as the output variable of the  $\epsilon$ -SVR modeling system. The over bark diameter at breast height ( $d_{0.1.3}$ ), the relative stem heights ( $h_i$ ) where the over bark diameters were measured, the total height ( $tht$ ) of the sampled trees, the ratio between the relative height and the total height of the trees and the ratio between the over- and under-bark diameters at breast height, that are easily obtained on the tree bole, were used as the input information to the modeling system.

### 2.4 Model evaluation criteria

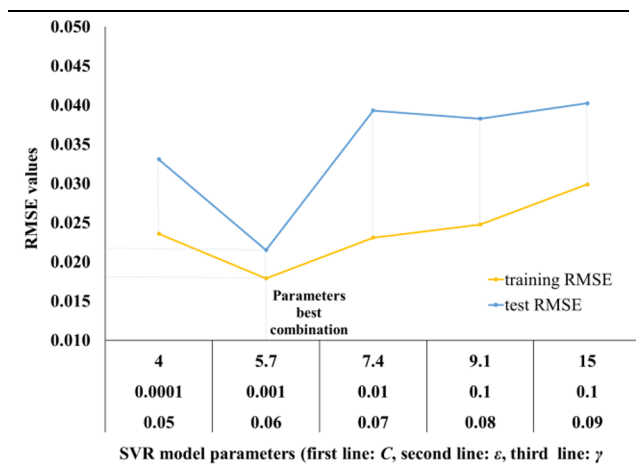
The indicators calculated for the training and the test data set were:

the correlation coefficient (R), the root mean square error (RMSE) and the % root mean square error (RMSE%), and the average absolute error (AAE), of the bark factor ( $BF_i$ ) used as the output feature in the modeling process. In addition, the measured bark factor ( $BF_i$ ) values were compared with the predicted by the constructed model values. For this reason, the 45-degree line and the paired t-test were used. As a final evaluation of the statistical behaviour of the  $\epsilon$ -SVR constructed model, its error distribution was examined, as well.

## 3 Results and discussion

The optimum combination of the meta-parameters values was selected using the tuning technique called grid-search methodology [26]. The exhaustive search was performed in the range of  $10^{-4}$  to 0.30, by 0.001 concerning the configuration of the ( $\epsilon$ ) parameter, in the range of 0.05 to 1.00, by 0.05 concerning the configuration of the ( $\gamma$ ) parameter, while the tested values of the ( $C$ ) ranged from 2 to 30, by 0.1, in order for the optimum combination that led the system to the minimum loss, to be selected.

Following the grid-search methodology, the  $\epsilon$ -SVR model produced the most accurate results, used the optimum combination of its meta-parameters with values  $C = 5.7$ ,  $\gamma = 0.06$ , and  $\epsilon = 0.001$  (Fig. 3).



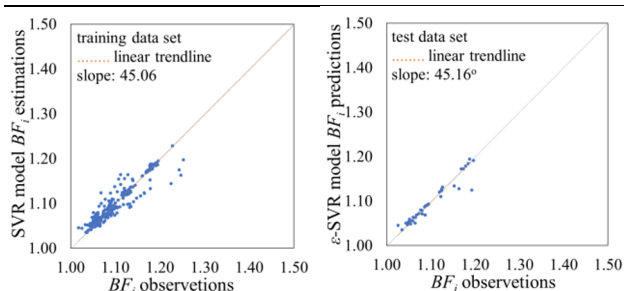
**Fig. 3.** SVR model meta-parameters optimum combination.

The evaluation statistics of the  $\epsilon$ -SVR model that will be used for the accurate estimations and predictions of the bark factor values along the tree bole, for the fitting and the testing data set, are given in Table 2.

**Table 2.** Evaluation statistics of the  $\epsilon$ -SVR model showed the best adaptation to the available data.

metrics	Data set	
	Training (90% of all data)	Test (10% of all data)
R	0.9374	0.9885
RMSE	0.0179	0.0216
RMSE%	1.63	1.96
AAE	0.00032	0.00046

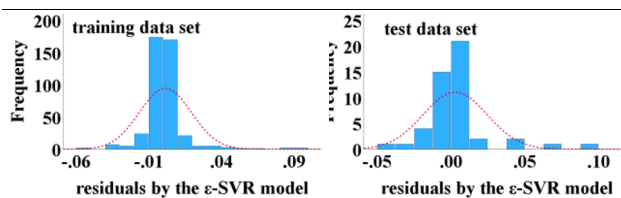
As it is shown (Table 2), the  $\epsilon$ -SVR manifested a remarkable adaptation not only to the training data set but to the “never seen by the model” test data set, as well. The estimations and predictions RMSE% which were 1.63% and 1.96% of the  $BF_i$  mean observed values, respectively, can be considered as sufficiently small error percentages  $BF$  values.



**Fig. 4.** 45-degree line for the training and the test data sets.

Furthermore, the proximity of each point to the 45-degree line indicated that the  $\epsilon$ -SVR constructed model, has the ability to adequately estimate and predict the bark factor values throughout the tree bole (Fig. 4).

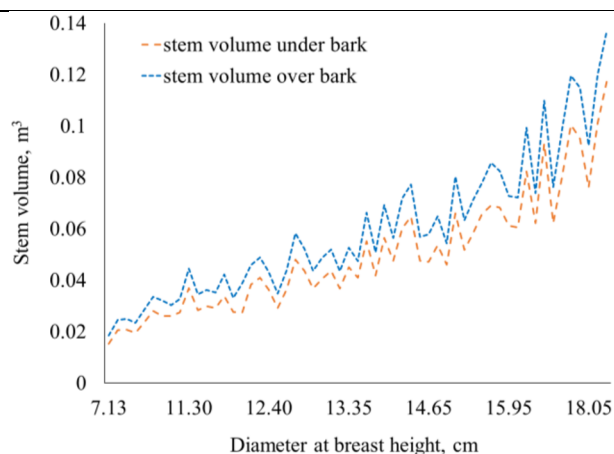
According to the t-tests applied under the significance of  $\alpha=0.05$ , to examine if there are any differences between the observed and a) the estimations and b) the predictions, which were both derived by the constructed  $\epsilon$ -SVR model, the p-values were equal to 0.100 and 0.371, respectively. The p-values were greater than 0.05, meaning that there are no significant differences, for both categories. Finally, the residual distribution of the  $\epsilon$ -SVR constructed model was also examined (Fig. 5).



**Fig. 5.** Residuals distribution.

It is obvious (Fig. 5) that both the estimations and predictions derived are reliable, with a pick of the residuals derived around zero, while there are a few larger residual values, meaning that the constructed system is a healthy one and shows the potential to accurately produce the bark factor values of pine trees.

Using the result by the  $\epsilon$ -SVR model and following the reasoning of the equations (4) and (5), the predicted pure total tree stem volume can be summarized to Figure 6, where the stem volume under-bark it is shown as it has been calculated availing the information derived by the constructed  $\epsilon$ -SVR model.



**Fig. 6.** Over bark and under bark volume as predicted using the results of the  $\epsilon$ -SVR bark factor model.

## 4 Discussion

In forestry, the production of reliable models that can accurately predicted the values of any tree attribute, is of vital importance, especially when there are difficulties in

the collection of the ground-truth data that almost always faced in the primary data collection in the field. Furthermore, data measured on biological organisms usually include high variability and belong to several kinds of non-normal distributions. For all the above reasons, the research in the modeling area to be focused in new and smart technologies that show the potential to effectively adapt to biological organisms' attributes without any prerequisites is a challenge. To this direction, machine learning is worth of exploration in the forestry research field.

On the other hand, nowadays, there is an increasing need followed by an equally increasing interest for wood and bark quantities that can be used not only for industrial utilization, but for energy, as well. Therefore, it is of significant importance to produce models that can effectively predict the quantities of wood and bark that a forest can produce. This knowledge is essential for the sustainable management of this natural resource, that is the forest. Since bark can reach about 15% of the wood volume, it is of utmost importance for the tree bark to accurately be predicted. The modelling technique that can reliably predict the tree bark quantities would significantly improve the forest management decisions in favor both the forest and the economic benefit from the use of this product. In this way, decisions related to the amount of biomass that can be safely removed from the forest will be safer, resulting to the avoidance of the forest ecosystem degradation over time.

The machine learning methodology used in this research effort was the  $\varepsilon$ -SVR technique, that uses the  $\varepsilon$ -insensitive loss function that ignores errors that are outside of the margin  $[-\varepsilon, \varepsilon]$ , which is known as the  $\varepsilon$ -insensitive tube. By solving the prediction problem at hand through the non-linear kernel radial basis function under optimum values of its meta-parameters using mapping of the nonlinearly separable available data into a linear separable future space, showed remarkable estimation and prediction abilities related to tree bark factor.

The pure volumes predicted with the use of the  $BF_i$ , where very close to the observed under-bark volumes of the pine trees, with mean absolute error equal to 0.0018  $m^3$  with values ranged from 0.000016  $m^3$  to 0.0042  $m^3$  and standard error of the mean equal to 0.00013  $m^3$  (Fig.7).

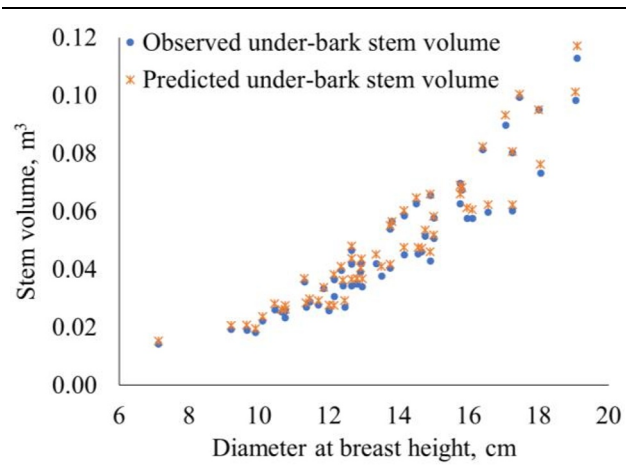


Fig. 7. Observed and predicted under-bark tree stem volume.

## 5 Concluding remarks

According to the described behaviour of the constructed  $\varepsilon$ -SVR model, the ability of the system to overcome outliers, noise, non-Gaussian distributions that frequently characterize the forest preliminary data measured in the field, it is clearly revealed.

Furthermore, the  $\varepsilon$ -SVR modeling technique showed remarkable adaptation to the non-linear ground-truth data, as well, producing effective and reliable bark factor prediction model. It was shown that the knowledge of bark factor values can lead to accurate estimations of bark quantities, further enables the accurate prediction of the plain wood volume that can be produced by the trees, as well.

Based on the results obtained, the use of the support vector regression technique, where the non-linear radial basis function kernels were embedded, is strongly recommended.

Finally, the results of this study can provide evidence along with a consistent basis for further research in the scientific area of forest modeling.

## References

1. A. Mauri, M. Di Leo, D. de Rigo, G. Caudullo, *European Atlas of Forest Tree Species* (Publ. Off. EU, Luxembourg, e0166b8+, 2016)
2. K. Matis, *Forest biometry: II. Dendrometry* (Pegasus, Thessaloniki, Greece, 2004)
3. K.E. Saarela, L. Harju, J. Rajander, J.O. Lill, S.J. Heselius, A. Lindroos, K. Mattsson, *Sci. Total Envir.* **343**, 1-3 (2005)
4. K.L. Mandiwana, T. Resane, N. Panichev, P. Ngobeni, *J. Hazard. Mater.* **137**, 2 (2006)
5. K. Steindor, B. Palowski, P. Góras, A. Nadgórska-Socha, *Pol. J. Environ. Stud.* **20**, 3 (2011)
6. Z. Pásztor, I.R. Mohácsiné, G. Gorbacheva, Z. Börcsök, *BioResources* **11**, 3 (2016)
7. M.J. Diamantopoulou, *Com. Elec. Agric* **48**, 3 (2005)
8. H.D. Marshall, G.E. Murphy, B. Lachenbruch, *For. Prod. J.* **56**, 11/12 (2006).
9. J. Laasasenaho, T. Melkas, S. Aldén, *For. Ecol. Manag.* **206**, 1-3 (2005)
10. M. Pompa-Garcia, J. Vega-Munoz, R., Sato-Gutierrez, G. Trincado, F. Cruz-Cobos, *Res. J. Forest.* **6**, 2 (2012)
11. C. Wehenkel, F. Cruz-Cobos, A. Carrillo, J. Lujan-Soto, *Scand. J. For. Res.* **27**, 6 (2012)
12. R. Bauer, A. Billard, F. Mothe, F. Longuetaud, M. Houballah, A. Bouvet, H. Cuny, A. Colin, F. Colin, *Ann. For. Sci.* **78**, 104 (2021)
13. B. Konôpka, J. Pajtik, V. Šebeň, K. Merganičová, *Plants*, **11**, 1148 (2022)

14. S.I. Yang, P.J. Radtke, *For. Ecol. Manag.* **503**, 119778 (2022)
15. S. Archontoulis, F.E. Miguez, *Agron. J.* **105**, 1–13 (2013)
16. M.J. Diamantopoulou, R. Özçelik, H. Yavuz, *Comput. Electron Agric.* **151**, (2018)
17. D. Cywicka, A. Jakóbk, J. Socha, D. Pasichnyk, A. Widlak, *PLoS ONE*, **17**, 11 (2022)
18. Management schedule of the urban forest of Thessaloniki, for the 2019-2028 period (Forest Service of Thessaloniki, Greece)
19. J.A. Kershaw Jr, M.J. Ducey, T.W. Beers, B. Husch, *Forest Mensuration* (Wiley-Blackwell, New York, USA, 2016)
20. Z. Liu, C. Peng, T. Work, J-N. Candau, A. DesRochers, D. Kneeshaw, *Environ. Rev.* **26**, 4 (2018)
21. M.J. Diamantopoulou, R. Özçelik, B. Koparan, O. Alkan, *Turk. J. Agric. For.* **47**, 2 (2023)
22. V.N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York USA, 2000)
23. J. Wu, W. Yao, S. Choi, T. Park, R.B. Myneni, *IEEE Geosci. Remote Sens. Lett.* **12**, 11 (2015)
24. D.H.B. Binoti, M.L.M. Binoti, H.G. Leite, A.V. Andrade, G.S. Nogueira, M.L. Romarco, C.G. Pitanguí, *Revista Árvore* **40**, 4 (2016)
25. P.J. García-Nieto, J. Martínez-Torres, M. Araújo-Fernández, C. Ordóñez-Galán, *Appl. Math. Modell.* **36**, 12, (2012)
26. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J Mach. Learn. Res.* **12**, 85 (2011)
27. Python Software Foundation, *Python* **3.9** (2022)