k-NEAREST NEIGHBOR AND LINEAR REGRESSION IN THE PREDICTION OF THE ARTIFICIAL FORM FACTOR

Deivison Venicio Souza[1*], Júlio Cesar Nievola[2], Ana Paula Dalla Corte[3], Carlos Roberto Sanquetta[3]

[1]Federal University of Pará, Faculty of Forestry Engineering, Altamira, Pará, Brasil. E-mail: deivisonvs@ufpa.br*
[2]Pontifical Catholic University of Paraná, Graduate Program in Computer Science, Curitiba, Paraná, Brasil. E-mail: nievola@ppgia.pucpr.br
[3] Federal University of Paraná, Department of Forestry Sciences, Curitiba, Paraná, Brasil. E-mail: anapaulacorte@gmail.com; carlossanquetta@gmail.com

Received for publication on: 30/03/2019 – Accepted for publication on: 07/06/2019

Resumo

k-vizinhos mais próximos e regressão linear na predição do fator de forma artificial. A proposta deste estudo foi testar se a abordagem não-paramétrica, conhecida como *k*-Nearest-Neighbor (*k*-NN), poderia melhorar as estimativas do fator de forma artificial ($f_{1,3}$) individual de árvores do híbrido *Eucalyptus urophylla x Eucalyptus grandis*, em comparação ao método de Mínimos Quadrados Ordinário. Foram selecionadas e derrubadas 149 árvores-amostras e medidos ao longo do fuste os diâmetros a 10% ($d_{0,1}$), 30% ($d_{0,3}$), 50% ($d_{0,5}$) e 70% ($d_{0,7}$) da altura do fuste comercial e, posteriormente, a cada 2m. Modelos matemáticos reconhecidos na literatura para predição do fator de forma foram ajustados para comparação. O hiperparâmetro *k* de ajuste ótimo para o estimador *k*-NN foi obtido através de repetidas validações cruzadas. Os dados de treinamento do modelo de regressão *k*-NN foram idênticos aos utilizados no ajuste dos modelos de regressão linear. A maior parte dos modelos de regressão linear múltipla apresentou problemas de colinearidade ou multicolinearidade. O uso da covariável $(d_{0.3}.d_{0.7})/d_{1.3}^2$ e *k* = 15 possibilitou a construção de modelos *k*-NN com melhor capacidade de generalização. O potencial do estimador *k*-NN para predizer o fator de forma artificial e, por conseguinte, obter estimativas menos viesadas dos volumes individuais de árvores foi admitido e, considerado superior ao uso da regressão linear e fatores de forma médios. A abordagem *k*-NN pode ser considerada mais genérica para predizer o fator de forma de árvores, e seu uso pode ser aconselhado quando modelos de regressão linear clássicos, ou outros métodos mais simples, não mostrarem bons resultados.

Palavras-chave: Eucalyptus, aprendizado de máquina, vizinho mais próximo, regressão linear, fator de forma 0,7

Abstract

The proposal of this study was to test whether the performance of the nonparametric approach *k*-Nearest Neighbor (*k*-NN), would improve estimates of individual artificial form factor ($f_{1.3}$) of trees of the hybrid *Eucalyptus urophylla x Eucalyptus grandis* compared to the Ordinary Least Squares method. A total of 149 sample-trees were selected, felled, and diameter was measured along the trunk at 10% ($d_{0.1}$), 30% ($d_{0.3}$), 50% ($d_{0.5}$) and 70% ($d_{0.7}$) of commercial height and posteriorly at 2m intervals. Mathematical models recognized in the literature for predicting the form factor were adjusted for comparison. The hyperparameter *k* of optimum adjustment for the *k*-NN estimator was obtained by repeated cross-validation. The training data of the *k*-NN regression model were identical to those used in the adjustment of the linear regression models since most multiple linear regression models present problems of collinearity or multicollinearity. The use of the covariate $(d_{0.3}.d_{0.7})/d_{1.3}^2$ and *k* = 15 made it possible to construct *k*-NN models with better generalization capacity. The potential of the *k*-NN estimator to predict the artificial form factor and thus to obtain less biased estimates of individual tree volumes was demonstrated and considered to be superior to the use of linear regression and average form factors. The *k*-NN approach can be considered more generic for prediction of the tree form factor, and its use is recommended when classical linear regression models or other simpler methods do not yield good results.

Keywords: Eucalyptus, machine learning, nearest neighbor, linear regression, form factor 0.7

INTRODUCTION

In 2016, the total area of planted trees in Brazil increased by 0.5% in relation to the year 2015, totaling 7.84 million hectares (ha). The genus *Eucalyptus* occupies 5.7 million hectares of planted trees with a 2.4% annual growth rate during the last five years. In this period, the state of Mato Grosso do Sul had the largest expansion of *Eucalyptus* culture registering an increase of 400,000 ha, with an average annual growth rate of 1.3%. Among the Brazilian states, the state of Pará was 10th in terms of area of planted Eucalyptus in 2016, covering 133,996 ha (IBÁ, 2017).

The accurate estimation of wood volume is among the most important pieces of information about a forest plantation, which among other factors is essential for proper forest management. The use of volume equations, taper functions and the average form factor of the plantation stand are the most common ways of quantifying forest production (DRESCHER *et al.*, 2001, PACHECO *et al.*, 2017). The form factor is used to reduce the error caused by the action of equalizing the volume of the tree to the volume of a perfect cylinder. Therefore, the volume can be estimated by the equation: v_i = g.h.f, where v_i = volume of the i-th tree (m^3), g = sectional area at 1.30m from the ground (m^2), h = tree height (m), and *f* = tree form factor (ADEKUNLE *et al.*, 2013).

The form factor expresses the ratio of diameter and height to the volume of a tree (EASTAUGH, 2014). However, obtaining the volume through the basic characteristics of the individual stems, such as diameter and height, is widely recommended as a function of practicality, but this has potential to cause errors that result from variation in the shape of the trunk of the trees. (SOCHA; KULEJ, 2007).

Another important aspect of management of forest plantations is the generic use of the form factor. In the state of Pará, for example, the prediction of the volume of standing wood in planted stands has traditionally been conducted using the form factor 0.7, disregarding variation due to species, age, silvicultural characteristics, and other factors that affect the shape of the trunk. This 0.7 factor was developed by the FAO mission for native forest species, based on forest surveys carried out in Amazonia between 1956-1961. Therefore, the generic use of the form factor 0.7 for different species may imply biased estimates of the actual volumes.

Parametric approaches such as linear regression have been the main focus of many modeling studies of the form factor of trees in forests of unequal ages and of the same age (e.g., DRESCHER *et al.*, 2001; DRESCHER *et al.*, 2010; BOTH *et al.*, 2011; ADEKUNLE *et al.*, 2013; CUNHA NETO *et al.*, 2016, TENZIN *et al.*, 2017). However, nonparametric approaches, such as artificial neural networks, have been shown to be appropriate in modeling forest attributes (e.g., DIAMANTOPOULOU; MILIOS, 2010; MARTINS *et al.*, 2016, SANQUETTA *et al.*, 2018). In this context, recent initiatives have revealed the potential of the Nearest Neighbor (*k*-NN) method in the prediction of biometric variables, such as volume, biomass and carbon stock in trees (SANQUETTA *et al.*, 2013, SANQUETTA *et al.*, 2015, SOUZA *et al.*, 2019).

The proposal of this study was to test whether the performance of the nonparametric approach *k*-Nearest-Neighbor (*k*-NN) would improve estimates of individual artificial form factor ($f_{1.3}$) from trees of the hybrid *Eucalyptus urophylla* x *Eucalyptus grandis* compared to the Ordinary Least Squares method. Thus, the main contributions of the study were: i) the construction and use of parametric and nonparametric models to predict the artificial form factor ($f_{1.3}$); and ii) the comparison between the estimated volumes using different approaches to obtain the form factor (linear regression, *k*-NN, medium *ff* and *ff* 0.7) and the actual volume obtained by rigorous cubing. Finally, the hypothesis that was tested was if the *k*-NN approach has the potential to present better precision in the estimation of the artificial form factor and, consequently, reduce errors in the estimation of the individual volume inherent to the variation in the shape of the tree trunk.

MATERIALS AND METHODS

Study area and data collection.

The study was performed in a plantation stand of eight-year-old hybrid *E. urophylla x E. grandis*, with a spacing of 3m x 3m, density of 1,089 trees.ha^{-1} and with effective planting of 11.33 ha. The plantation stand belongs to Ipiaçava Indústria e Comércio de Madeiras LTDA, which is headquartered in Pacajá City, in the southwest of the state of Pará on the BR-230 (Transamazon highway), in the Southwest mesoregion with the following geographic coordinates: latitude N -03°38'59,20473" and longitude E -50°57' 19,45084" (Datum WGS 84).

A total of 149 trees were felled and their diameters were measured at 10% (d0.1), 30% (d0.3), 50% (d0.5) and 70% (d0.7) along the commercial height of the trunk. The trunk volume was obtained by Huber's equation, wherein the volume in m^3 of the i-th trunk section (v_i) was calculated following the expression $v_i = g_m . l$, where g_m = sectional area, in m^2 in the middle of the i-th section, and l = length of the i-th section in meters. The volume of the top was obtained assuming the volume of the cone.

Artificial form factor ($f_{1.3}$)

The artificial form factor ($f_{1.3}$) of i-th tree was obtained dividing the actual volume (method of Huber) by the volume of a perfect cylinder, which was obtained and based on the total height and transverse section (1.30 from the ground) (Eq.1) (TENZIN *et al.*, 2017; PACHECO *et al.*, 2017). In general, after the factor ($f_{1.3}$) calculated for each tree, an average value can be used to obtain estimates for the volume of the standing tree. Thus the volume of the *i*-th tree (m^3) can be calculated using the formula in Eq.2, where: v_i = volume of the i-th tree (m^3), diameter 1.30m from the ground; g_i = sectional area of the i-th tree, measured 1.30 from the ground (m^2), h_i = height of i-th tree; (m), and *f* = form factor (ADEKUNLE *et al.*, 2013; TENZIN *et al.*, 2017).

$$f_{1,3} = \frac{v_{Huber}}{v_{cylinder}} \quad \text{(Eq.1)}$$

$$\hat{v} = g_i\left(\frac{\pi.d_{1,3}^2}{4}\right).h_i.f_{1,3} \quad \text{(Eq.2)}$$

Linear regression (parametric modeling)

Successful mathematical models used in the prediction of the artificial form factor $f_{1,3}$ of other species in planted stands (see DRESCHER *et al.*, 2001; DRESCHER *et al.*, 2010; BOTH *et al.*, 2011; CUNHA NETO *et al.*, 2016) were chosen for modeling employing classical linear regression using ordinary least squares (OLS). In general, the predictive variables in the models reflect relationships between diameters measured at different positions relative to the total height of the tree, and with the total height of the tree. Ten artificial form factor models ($f_{1.3}$) were adjusted. The last two models of $f_{1.3}$ were proposed through experimentation, due to high linear correlation with the response variable and by searching for simpler linear models than those described in the literature (Table 1). Relative diameter models Hohenadl ($d_{0.i}$) were also evaluated since they are required for $f_{1.3}$ models. The original data set was divided into training (75%) and testing (25%) data by means of random stratified sampling, based on tree diameters.

Tabela 1. Modelos matemáticos de fator de forma artificial ($f_{1,3}$) e diâmetros Hohenadl ($d_{0,i}$).
Table 1. Mathematical models of artificial form factor ($f_{1,3}$) and Hohenadl diameters ($d_{0,i}$).

M	Artificial Form Factor ($f_{1,3}$)	Author
FF1	$\ln(f_{1.3}) = \beta_0 + \beta_1\ln(d_{0.5}/d_{1.3}^2) + \beta_2\ln(d_{0.1}/d_{1.3}^2) + \varepsilon_i$	Drescher et al. (2010)
FF2	$\ln(f_{1.3}) = \beta_0 + \beta_1\ln(d_{0.5}/d_{1.3}^2) + \beta_2\ln(d_{0.1}/d_{1.3}^2) + \beta_3\ln[1/(d_{1.3}.h)] + \varepsilon_i$	
FF3	$f_{1.3} = \beta_0 + \beta_1(d_{0.5}/d_{1.3}^2) + \varepsilon_i$	
FF4	$f_{1.3} = \beta_0 + \beta_1(d_{0.5}/d_{1.3}^2) + \beta_2[(d_{0.3}.d_{0.7})/d_{1.3}^2] + \varepsilon_i$	
FF5	$f_{1.3} = \beta_0 + \beta_1(d_{0.3}/d_{1.3}^2) + \beta_2(d_{0.3}.d_{0.5})/d_{1.3}^2 + \beta_3(d_{0.5}/d_{1.3})^2 + \varepsilon_i$	Drescher et al. (2001)
FF6	$f_{1.3} = \beta_0 + \beta_1(d_{0.3}/d_{1.3}^2) + \beta_2(1/d_{1.3}^2) + \beta_3.d_{0.5} + \varepsilon_i$	Cunha Neto et al. (2016)
FF7	$f_{1.3} = \beta_1(h/d_{1.3}) + \beta_2(1/h) + \beta_3.d_{0.7} + \varepsilon_i$	Both et al. (2011)
FF8	$f_{1.3} = \beta_1(d_{0.1}/d_{1.3}) + \beta_2[(d_{0.3}.d_{0.5})/d_{1.3}] + \varepsilon_i$	
FF9	$f_{1.3} = \beta_1\ln(d_{0.5}/d_{1.3}^2) + \beta_2\ln(d_{0.3}.d_{0.5})/d_{1.3} + \varepsilon_i$	Proposed
FF10	$f_{1.3} = \beta_0 + \beta_1.h + \beta_2\ln[(d_{0.3}.d_{0.7})/d_{1.3}^2] + \varepsilon_i$	
	Relative Hohenadl Diameters ($d_{0,i}$)	
DH1	$\ln(d_{0.i}) = \beta_0 + \beta_1 d_{1.3} + \beta_2(h.d_{1.3}) + \varepsilon_i$	Both et al. (2011)
DH2	$\ln(d_{0.i}) = \beta_0 + \beta_1\ln(1/d_{1.3}) + \varepsilon_i$	
DH3	$\ln(d_{0.i}) = \beta_1\ln(1/d_{1.3}) + \varepsilon_i$	
DH4	$\ln(d_{0.i}) = \beta_1\ln(1/d_{1.3}) + \beta_2(1/h) + \varepsilon_i$	
DH5	$\ln(d_{0.i}) = \beta_0 + \beta_1\ln(1/d_{1.3}) + \beta_2(1/h) + \varepsilon_i$	

Where: M = model; $f_{1.3}$ = artificial form factor; ln = neperian logarithm; $d_{1.3}$ = diameter measured at 1.30 m from the ground; $d_{0.1}$; $d_{0.3}$; $d_{0.5}$; $d_{0.7}$ = Hohenadl relative diameters at 10%, 30%, 50% and 70%, respectively, of commercial height of the tree trunk; h = height of the commercial trunk; $d_{0.i}$ = diameters relative to 30% or 50%, respectively, of commercial height of the tree trunk; β_0, β_1, β_2, β_3 = regression coefficients; and εi = random error term; εi = N (0, σ^2).

The original data set was divided into training (75%) and test (25%) data using stratified random sampling based on tree diameters. The regression models were adjusted in the training set using the ordinary least squares method and evaluated in the testing set. The training data of the *k*-NN regression model were identical to those used in adjusting the linear regression models. The training data of the linear regression model were identical to those used in adjusting the *k*-NN regression models.

The following adjustment quality statistics were used to choose the best linear regression model: PRESS statistic, R-squared, standard error of the estimate in percentage (Syx%), root mean square error (RMSE) and Akaike Information Criteria (AIC). The student's t-test (α = 0.05) was used to evaluate the significance of regression coefficients. The VIF statistics were obtained for each of the coefficients of the multiple regression models through the expression: $\mathrm{VIF}(\hat{\beta}_i) = 1/(1 - R_i^2)$, where: R_i^2 = coefficient of determination obtained by the regression of the explanatory variable $X_i (i = 2, \dots, p)$ with the other independent variables.

Linear regression residues were evaluated for adherence to the assumptions of normality (*Shapiro-Wilk*, $\alpha = 0.05$), homoscedasticity (*Breusch-Pagan*, $\alpha = 0.05$), and autocorrelation (*Durbin-Watson*, $\alpha = 0.05$). The graphs of residuals versus adjusted values (dependent variable scale) and the Normal Probability Paper graph (Q-Q plot) were also analyzed. All analyses were performed using several packages of the statistical environment R, version 3.4.3 (Table S1). The adjustment of all linear regression models was done using the "lm" function of the "stats" package available in the R-base.

Regression k-NN - (non-parametric modeling)

The instance-based learning algorithm (*k*-NN) is a type of supervised and non-parametric method (HECHENBICHLER; SCHLIEP, 2004). In its most basic version, the use of *k*-NN requires three essential elements: a set of labeled objects, a distance or similarity metric to calculate the distance between objects and the value of k (number of nearest neighbors) (LI *et al*., 2017).

To implement the algorithm of the nearest neighbor, we used the interface of the "caret" package that has the *k*-NN (*k*-nearest neighbors) method (KUHN, 2016). This simpler version only needs to find the *k* value (number of nearest neighbors) of optimal fit. By default, the Euclidean distance was used. The training set (75%) was used to find the optimal adjustment of hyperparameter *k* for the *k*-NN estimator using repeated cross-validation, and the test set was used to evaluate the generalization capacity of the best model built on new data. Therefore, the test set was excluded from the construction process of the *k*-NN models. The performance comparison between the approaches (linear regression and *k*-NN) was performed on the test set.

The selection of relevant variables is an important task in the process of constructing the *k*-NN models. Therefore, the marginal importance of covariates for the *k*-NN estimator was measured using metrics independent of the adjusted model. This approach can be applied using the 'filterVarImp' function of the 'caret' library (see KUHN, 2016). By this method, the default is to adjust smoothed curves between each predictor and the response variable using the LOESS regression method, allowing estimates to be obtained within a local neighborhood window. The LOESS regression has the 'span' argument that controls the size of the neighborhood and therefore the degree of smoothing of the regression and can range from 0 to 1. Here, we use the default (span = 0.75) of the 'loess' function of the R-base stats package. After adjusting the individual models, the coefficient of determination (R^2) values are taken and computed as a relative measure of the importance of the variable. To evaluate the importance, all the predictors used in the linear regression models and other relations of diameter and height were considered.

To train the *k*-NN regression model, the *k*-fold cross-validation method was used (KUHN; JOHNSON, 2013), and a manual grid with *k* candidate values optimum setting of the hyperparameter was established. In the learning process, to reduce the computational overhead intrinsic to the algorithms based on instances, we established a maximum *k* of 25 nearest neighbors, starting with 2-NN.

The performance of the *k*-NN model, in each disjointed subset k of validation, was measured using RMSE *(Root Mean Square Error)* metrics, rRMSE *(Relative Root Mean Square Error)* and R^2 *(R-squared)* (KVALSETH, 1985; KUHN, JOHNSON, 2013). In addition, the Pearson correlation coefficient (r) was calculated to quantify the correlation between the observed and predicted values for each variant *k*-NN and bias. The final performance of each *k*-NN model was obtained by generating the arithmetic mean of the estimates in the 50 disjointed subsets *k* of the 5x10-*folds* CV scheme. All analyses were performed using several packages of the statistical environment R, version 3.4.3.

RESULTS

The Pearson correlation coefficients (r) between the dependent and independent variables contained in the artificial form factor models ($f_{1.3}$), were significant at the level of $\alpha = 0.05$, except for the correlation between $f_{1.3}$ and $h/d_{1.3}$ (r =-0.062^{ns}; FF7). The f1,3 in its natural form showed a degree of association of less than 70% with its regressors, in most cases. However, there were strong correlations between $f_{1.3}$ and $(d_{0.3}.d_{0.7})/d_{1.3}^2$ (r =0.943*; FF4, FF9 and FF10), $f_{1.3}$ and $(d_{0.3}.d_{0.5})/d_{1.3}^2$ (r =0.873*; FF5), and also $f_{1.3}$ and $(d_{0.5}/d_{1.3})^{2}$(r =0.864*; FF5).

Severe multicollinearity effects on model coefficient estimates were detected for most multiple linear regression models, except for the models FF4, FF9 and FF10 (VIF < 10); for the models with VIF > 10 there were correlations between predictor variables greater than 80%.

The FF5 model was not considered satisfactory since it presented nonsignificant coefficients β_2 and β_3 ($\alpha = 0.05$). The *F-Snedecor* statistic was significant ($\alpha = 0.01$) for all adjusted form factor models, thus rejecting the null hypothesis of non-regression. The *Durbin-Watson* test ($\alpha = 0.05$) did not show evidence of autocorrelation of the residuals for the fitted models, in other words, the residuals are independent. The diagnosis of homoscedasticity and normality of the residues was done only for the FF4, FF9 and FF10 models, since they were the only ones that had significant differences for all their regression coefficients ($\alpha = 0.05$), without multicollinearity problems (VIF < 10) , acceptable R^2 (> 86%) and Syx% less than 5% (Table S2).

The null hypothesis of homoscedasticity of variance was rejected (*Breusch-Pagan*; $\alpha < 0.05$) for the FF4 and FF10 models. The FF9 model was the only one that had a normal distribution of residuals (*Shapiro-Wilk*; $\alpha = 0.05$); the normal distribution of the residuals was verified through the Quantile-Quantile plot by analyzing the proximity of the points (theoretical quantile and standardized residuals) to the red line. For this model the residuals were homoscedastic and non-tendentious, with an average close to zero. The residuals also showed a linear pattern, as suggested by the smoothing curve (red line in the residual vs. predicted plot) (Figure 1).

In this study, the FF9 model had the best performance with excellent fit quality statistics. For this model the degree of explanation of the dependent variable ($f_{1.3}$) by the regressors was 99% and the Syx% was less than 10%. The equation DH2 used to adjust $d_{0.3}$ was the only one that met all the linear regression assumptions. With respect to the models for prediction of $d_{0.5}$, the DH4 model was indicated by its lower AIC value.

Machine learning algorithms have hyperparameters that need to be tuned in order to avoid overfitting and to find the best compensation bias-variance model. Here, we used the basic version of *k*-NN which required finding the optimal value of *k* (number of nearest neighbors) for which the loss function RMSE was minimal. To obtain an indication of the covariant space most appropriate to the *k*-NN estimator, before submitting the data to the learning process we identified the marginal importance of potential predictors (those used in linear regression models and others) to the estimator *k*-NN (Figure S1), and then the performance of the models was estimated using repeated cross-validations.

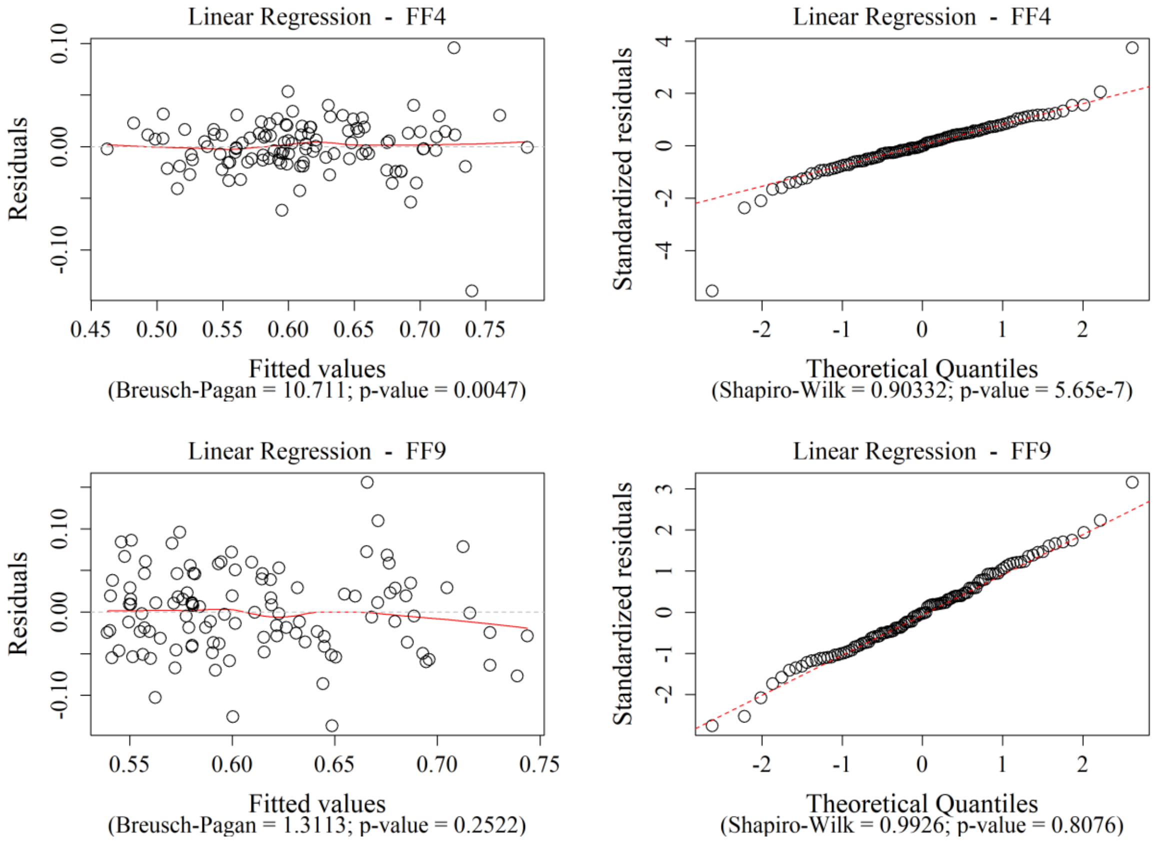

Figura 1. Resíduos versus valores ajustados (escala da variável dependente) e gráfico quantil-quantil para os modelos sem autocorrelação, multicolinearidade e coeficientes da regressão nulos.

Figure 1. Residual versus fitted values (dependent variable scale) and quantile-quantile plot for models without autocorrelation, multicollinearity and null coefficients of regression.

From the analysis of the marginal importance of the covariates to the *k*-NN estimator, we obtained the result that using the covariate $(d_{0.3}.d_{0.7})/d_{1.3}^2$ (logarithmic or not) would be the best approach to obtain a more accurate model. Indeed, the marginal application of both covariates ensured better performance to the estimator, but the performance did not differ between the two; therefore, the use of the covariate without the logarithm was preferred. Finally, using the covariate $(d_{0.3}.d_{0.7})/d_{1.3}^2$ (non-logarithmic) and 15 nearest neighbors (*k* = 15) of the instance was predicted, and this was the configuration with better generalization capacity (lower RMSE) and was therefore used to predict the artificial form factor.

The hyperparameter *k* has a significant influence on the performance of the *k*-NN estimator. The rRMSE curve in cross-validation shows the expected pattern, i.e., delineating the "U" shape. In general, the curve had a decreasing pattern reaching the best performance when *k* = 15 and from that point the increase of *k* caused a decrease in accuracy. In contrast, the learning set error decreases as the flexibility (< *k*) of the model increased (Figure 2a). Thus, comparison of the curves shows that *k* = 15 represents the best balance between bias and variance such that it minimizes the prediction error of the form factor. The RMSE, rRMSE, r and bias values in the cross-validation, when using the 15-NN model, were 0.0379 m³, 6.19%, 0.8492 and 0.198%, respectively. Other quality metrics in the validation and training sets, as a function of *k*, are also presented in Table S3.

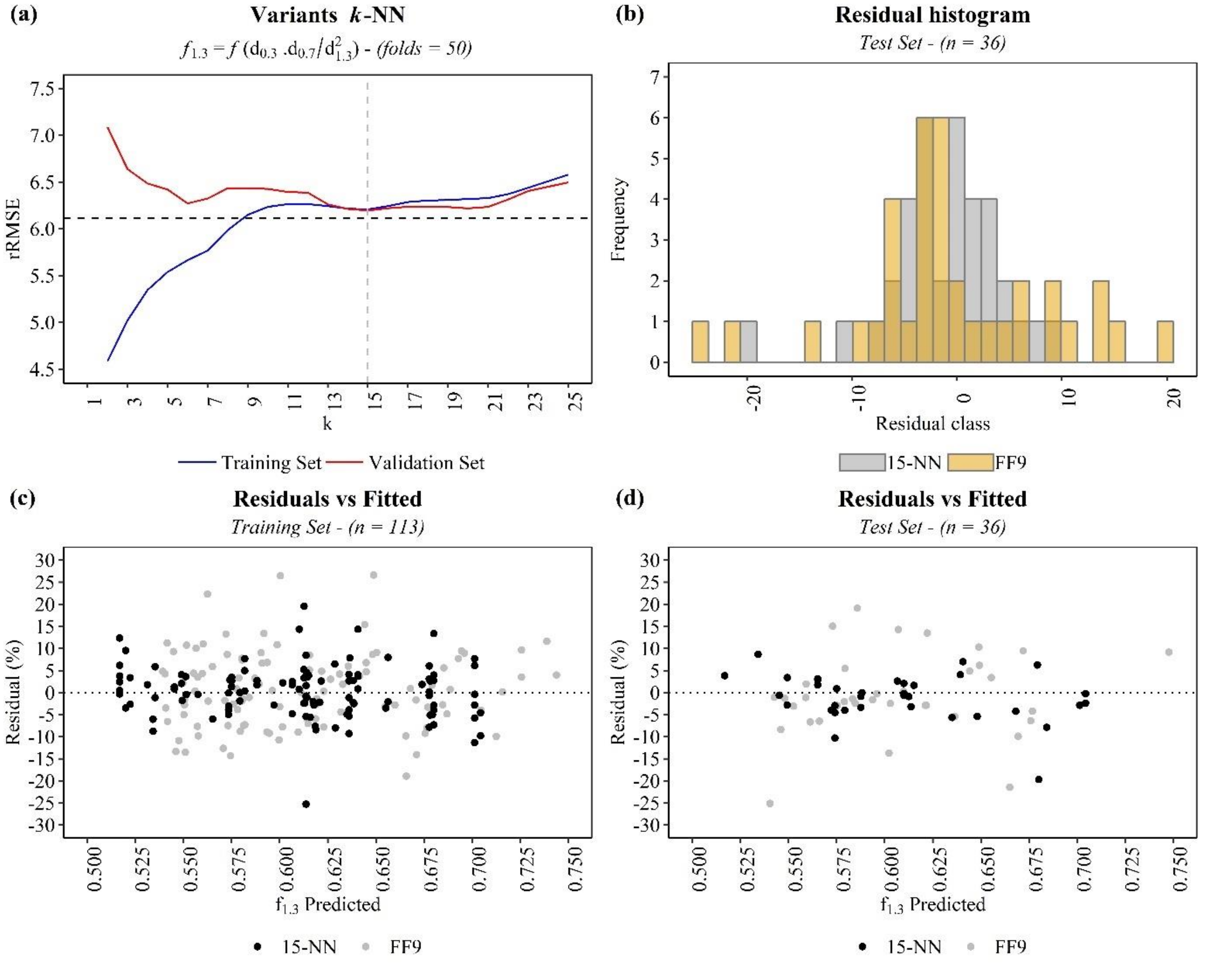

Figura 2. a) Perfil da medida *rRMSE* nos conjuntos de treino e validação, como função de *k*. Em que: r*RMSE* = Raiz do Erro Quadrático Médio Relativo; *k* = número de vizinhos mais próximos; $f_{1,3}$ = fator de forma artificial; $d_{1,3}$ = diâmetro medido a 1,30 m do solo; $d_{0,3}$ e $d_{0,7}$ = diâmetros relativos Hohenadl a 30% e 70%, respectivamente, da altura do fuste comercial da árvore; linha vertical pontilhada = representa o melhor *k* na validação; linha horizontal pontilhada = representa desempenho no conjunto de teste usando a configuração *k*-NN ótima; b) Histogramas de frequência residuais para o dados de teste; c) Resíduos versus valores estimados para os dados de treinamento; e d) Resíduos versus valores estimados no conjunto de teste. Em que: 15-NN = modelos *k*-NN escolhido que usa 15 vizinhos mais próximos e a

covariável de maior importância marginal para predizer o fator de forma artificial; FF9 = modelo de regressão linear selecionado.

Figure 2. a) Profile of the rRMSE measure in the training and validation sets as a function of *k*. Where: rRMSE = Relative Mean Square Error Root; *k* = number of nearest neighbors; $f_{1.3}$ = artificial form factor; $d_{1.3}$ = diameter measured at 1.30 m from the ground; $d_{0.3}$ and $d_{0.7}$ = Hohenadl relative diameters at 30% and 70% respectively of commercial height of the tree trunk; dotted vertical line = represents the best *k* in the validation; dotted horizontal line = represents performance in the test set using the optimal *k*-NN configuration; b) Residual frequency histograms for test data; c) Residual versus estimated values for the training data; and d) Residual versus fitted values in the test set. Where: 15-NN = chosen *k*-NN models that uses 15 nearest neighbors and the covariant of greater marginal importance to predict the artificial form factor; FF9 = selected linear regression model.

The analysis of frequency histograms indicates that the 15-NN model has the highest concentration of residual values close to zero (Figure 2b), and this fact is confirmed by the residual plot versus fitted values (Figure 2c). The training set had a discrepant point (observed $f_{1.3}$ = 0.8214), and both models (FF9 and 15-NN) had greater difficulty predicting it. In general, the quality measures done using the training set were better for the 15-NN model.

The resubstitution error is not a good parameter to evaluate a model's prediction ability using new data. Therefore, the predictive quality of the best RL model (FF9) and the best *k*-NN configuration (15-NN) (adjusted to the training set, n = 113) was validated in the test set (n = 36). At tree level, the 15-NN model was more accurate to predict the individual form factor. The average values of r, R², RMSE, rRMSE and bias were 0.8749, 0.7655, 0.0376 m³, 6.12% and 1.51% for the 15-NN model. For the FF9 model, we found r = 0.5494, R² = 0.2743, RMSE = 0.0608 m³, rRMSE = 9.88% and bias = 1.32% (Figure 2d and S2).

DISCUSSION

Linear regression (LR) is a simple parametric method with consolidated use in the prediction of forest biometric variables. Most commonly, the estimation of single or multiple RL parameters is done using the ordinary least squares (OLS) method. By this method, the choice of the appropriate model should ensure, in addition to good quality of fit statistics, the non-violation of assumptions of independence, homoscedasticity and normality of errors. If there is a violation of any of these hypotheses, the inferences (confidence intervals and hypothesis tests) based on the LR model may be erroneous or unreliable. For multiple models, it is important to verify the existence of collinearity or multicollinearity, which refers to the strong linear correlation between predictor variables.

In this study, the selection of the best linear model to predict the artificial form factor was critical for the diagnosis of the residuals, the significance of the regression coefficients and the lack of a strong correlation between predictors. However, finding a linear model for $f_{1.3}$ prediction that met all the assumptions of LR was not trivial. In an initial analysis, most of the adjusted multiple models had their applicability rejected by the indication of existence of multicollinearity and / or non-significant regression coefficients. Subsequently, the difficulty was in finding models with homoscedastic and normal residuals.

The models FF2 and FF6 were indicated to predict the artificial form factor of *Tectona grandis* (DRESCHER *et al.*, 2010; CUNHA NETO *et al.*, 2016) and the FF5 model was recommended for *Pinus elliottii* (DRESCHER *et al.*, 2001) were not successful in predicting $f_{1.3}$ of trees in the current study. Here, both models presented strong evidence of multicollinearity, and specifically the FF5 model revealed non-significant regression coefficients.

It is common that high bivariate correlations between predictor variables (r > 0.75) lead to multicollinearity problems (MAROCO, 2010). This situation was verified for the FF2, FF5 and FF6 models, whose covariates showed linear correlations higher than 0.85 in most cases and had severe multicollinearity effects indicated by the VIF statistic. Here, the models with VIF > 10 had their general applicability rejected, because it becomes difficult to evaluate the relative importance of the predictor variables to explain the variation in the dependent variable (LIAO; VALLIANT, 2012). Moreover, as the variance inflation factor increases, the variance of the regression coefficient also increases (DIAMANTOPOULOU, 2005).

Multicollinearity does not affect the mathematics of model determination but influences the interpretation of each parameter estimation and on the predictive power of the constructed model. In multiple regression, the presence of severe multicollinearity causes instability in the estimates of the coefficients and consequently the quality of the interpretation of the coefficients associated with the covariates, which is one of the greatest advantages of the method. In practical terms, the loss of interpretability of the coefficients is observed, for example, by the change of signal for the regression coefficient, or even by a sensitive change in its magnitude with the inclusion or removal of model covariates. Moreover, the addition of correlated variables also causes an increase in the standard errors of the model parameter estimates resulting in less informative prediction intervals.

The Bartlett test (α = 0.05) showed homoscedasticity of variance within the groups of methods, and Analysis of Variance (ANOVA) was used to compare the volume averages between the different methods. The results indicated that the estimated average volumes using the different methods did not differ between them and also were equal to the observed mean volume ($p = 0.7693$) (Figure S3 and Tables S4). Acceptance of the null hypothesis of ANOVA reveals the efficacy of the methods to estimate the average individual volume.

Despite this, it is also interesting to compare the individual volume estimates using the artificial form of each tree predicted by the LR and *k*-NN models, the observed average form factor ($f_{1.3}$ observed mean factor values = 0.61), and the form factor 0.7 ($f_{1.3} = 0.7$) with the actual volume obtained by rigorous cubing (Huber). At tree level, individual volume estimates were more biased when using the average form factors approach, with a tendency to overestimate (Figure 3b and 3c). In general, the mean bias was -4.48% and -18.89% when using the mean factors $f_{1.3} = 0.61$ or $f_{1.3} = 0.7$, respectively. On the other hand, the use of individual artificial form factors estimated by the 15-NN model implied a smaller bias in the estimation of the individual volume, as well as more accurately estimated the total volume (sum of the volume of individual trees) (Figure 3a). The bias values were 1.56% and 0.39% for the FF9 and 15-NN models, respectively.

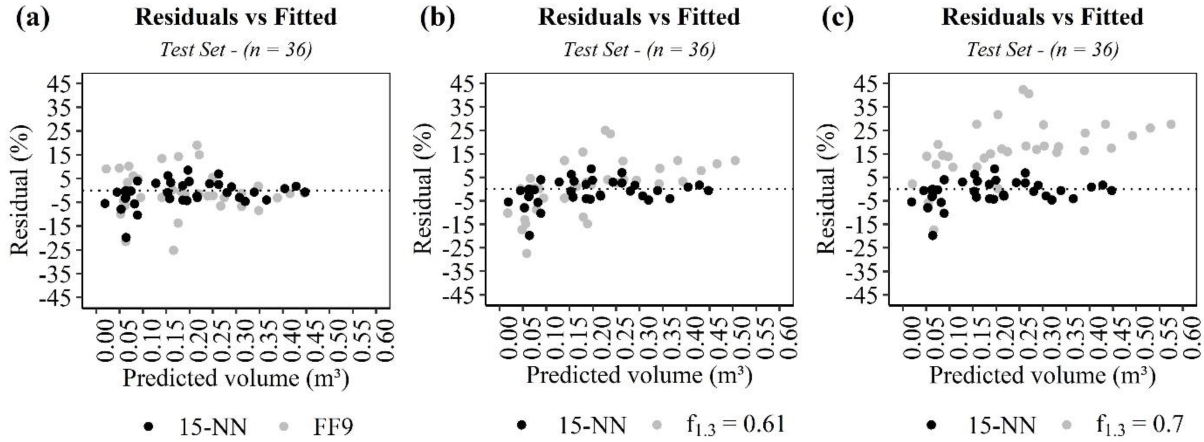

Figura 3. Resíduos versus volumes estimados de para observações do conjunto de teste (n = 36). São mostradas as diferenças residuais entre os modelos 15-NN e FF9 (a), 15-NN e $f_{1.3}$ médio = 0.61 (b) e, 15-NN e $f_{1.3}$ = 0.7.

Figure 3. Residual versus fitted volumes for test set observations (n = 36). The residual differences between the 15-NN and FF9 (a), 15-NN and $f_{1.3}$ average = 0.61 (b) and, 15-NN and $f_{1.3}$ = 0.7 models are shown.

Using non-parametric methods such as *k*-NN has advantages over classical linear regression methods. One benefit is that it does not require prior specification to a statistical model that describes the relation between the target response and the predictors; another is to not require compliance with statistical assumptions (FEHRMANN *et al*., 2008). Some older and even more recent studies have demonstrated the potential of the method based on instances to predict forest variables. Fehrmann *et al.* (2008) reported that the *k*-NN approach was superior to linear regression and mixed-effect models for estimating individual tree biomass of the species *Picea abies* (L.) Karst and *Pinus sylvestris* L. from data from the National Forest Inventory of Finland. Sanquetta *et al.* (2015) mentioned an increase up to 16.5% in the reduction of the standard error of estimation for models based on the famous Schumacher-Hall model in trees from environmental restoration plantations in the Brazilian Atlantic Forest biome. Sanquetta *et al.* (2018) reported that the k-NN approach improved the volume estimates of *Cryptomeria japonica* in southern Brazil, providing a more accurate estimate of total and commercial volumes.

The process of selecting variables in machine learning models, such as *k*-NN, is fundamental to the success of predictive modeling. In general, two strategies are described to evaluate the importance of a predictor variable in machine learning tasks: a) those that use the information of the adjusted model; and b) those that do not use the information from the adjusted model (KUHN; JOHNSON, 2013). Unfortunately, for the *k*-NN estimator there is still no approach based on the fitted model. In this study, the LOESS regression coefficient values were used as a relative measure of the importance of the covariate. Using this approach to identify potential predictors for *k*-NN was quite attractive as it allowed us to more efficiently find the optimal model fit; otherwise, an extremely tedious process of experimenting with predictors and their combinations would be necessary.

CONCLUSIONS

The results of this study permitted the following conclusions:

- The use of the *k*-NN estimator to predict the artificial form factor and thus to obtain less biased estimates of the individual tree volumes was evaluated and considered superior to linear regression and average form factors.
- In the case of multiple linear regression, the greatest difficulty was to obtain models without confounding effects between independent variables. The models adjusted here were successful in modeling the form factor of other species, but in this study suffered serious effects of collinearity or multicollinearity.
- The use of the form factor 0.7 as a generic approach to correct the volume of trees is not indicated, as it was evident that the individual volume estimates suffered from high bias.

REFERENCES

ADEKUNLE, V. A. J.; NAIR, K. N.; SRIVASTAVA, A. K.; SINGH, N. K. Models and form factors for stand volume estimation in natural forest ecosystems: a case study of Katarniaghat Wildlife Sanctuary (KGWS), Bahraich District, India. **Journal of Forestry Research**, v. 24, n. 2, p. 217-226, 2013.

BOTH, O. W.; CARVALHO, T. G.; DRESHER, R. Determinação de fator de forma artificial para *Tectona grandis* linn f., em povoamento equiâneo situado em Monte Dourado, Estado do Pará, Brasil. **Multitemas**, n. 40, p. 61-74, 2011.

CUNHA NETO, F. V.; VENDRUSCOLO, D. G. S.; DRESCHER, R. Artificial form factor equations for *Tectona grandis* in different spacings. **African Journal of Agricultural Research**, v. 11, n. 37, p. 3554-3561, 2016.

DIAMANTOPOULOU, M. J. Artificial neural networks as an alternative tool in pine bark volume estimation. **Computers and Electronics in Agriculture**, v. 48, n. 3, p. 235-244, 2005.

DIAMANTOPOULOU, M. J.; MILIOS, E. Modelling total volume of dominant pine trees in reforestations via multivariate analysis and artificial neural network models. **Biosystems Engineering**, v. 105, n. 3, p. 306-315, 2010.

DRESCHER, R.; SCHNEIDER, P. R.; FINGER, C. A. G.; QUEIROZ, F. L. C. Fator de forma artificial de *Pinus elliottii* Engelm para a região da serra do sudeste do estado do Rio Grande do Sul. **Ciência Rural**, v. 31, n.1, p. 37 - 42, 2001.

DRESCHER, R.; PELISSARI, A. L.; GAVA, F. H. Fator de forma artificial para povoamentos jovens de *Tectona grandis* em Mato Grosso. **Pesquisa Florestal Brasileira**, v. 30, n. 63, p. 191 – 197, 2010.

EASTAUGH, C. S. Relationships between the mean trees by basal area and by volume: reconciling form factors in the classis Bavarian yield and volume tables for Norway spruce. **European Journal of Forest Research**, v. 133, n. 5, p. 871-877, 2014.

FEHRMANN, L.; LEHTONEN, A.; KLEINN, C.; TOMPPO, E. Comparison of linear and mixed-effect regression models and k-nearest neighbour approach for estimation of single-tree biomass. **Canadian Journal of Forest Research**, v. 38, n. 1, p. 1-9, 2008.

HECHENBICHLER K.; SCHLIEP K. Weighted k-nearest-neighbor techniques and ordinal classification. **Institute Für Statistik Sonderforschungsbereich**, v. 399, p. 1-16, 2004.

IBÁ. **Indústria Brasileira de Árvores**. O Relatório Anual IBÁ. 2017. 80p.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Springer, 2013, 600p.

KUHN, M. Contributions from Jed Wing, Steve Weston, Andre Williams, ..., Can Candan and Tyler Hunt. **caret: Classification and Regression Training**. R package version 6.0-73, 2016.

KVALSETH, T. O. Cautionary Note about R2. **The American Statistician**. v. 39, n.4, p. 279-285, 1985.

LI, CHANGGENG; QIU, ZHENGYANG; LIU, CHANGTONG. An Improved Weighted *k*-Nearest Neighbor Algorithm for Indoor Positioning. **Wireless Personal Communications**, v. 96, n. 2, p. 2239-2251, 2017.

LIAO, D.; VALLIANT; R. Variance inflation factors in the analysis of complex survey data. **Statistics Canada**. v. 38, n. 1, p. 53-62, 2012.

MAROCO, J. **Análise Estatística** - Com utilização do SPSS. Lisboa: 3 ed. 2010, 824p.

MARTINS, E. R.; BINOTI, M. L. M. S.; LEITE, H. G.; BINOTI, D. H. B.; DUTRA, G. C. Configuração de redes neurais artificiais para estimação do afilamento do fuste de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias.** v.11, n.1, p.33-38, 2016.

PACHECO, J. M.; FIGUEIREDO FILHO, A.; DIAS, A. N. MACHADO, S. A.; LIMA, R.; ROVEDA, M. Effect of Spacing on the Form Factor of Pinus taeda L. **Australian Journal of Basic and Applied Sciences**, v. 11, n. 14, p. 139-143, 2017.

SANQUETTA, C. R.; WOJCIECHOWSKI, J.; CORTE, A. P. D.; RODRIGUES, A. L.; MAAS, G. C. B. On the use of data mining for estimating carbon storage in the trees. **Carbon Balance and Management**, v. 8, n. 6, p. 1-9, 2013.

SANQUETTA, C. R.; WOJCIECHOWSKI, J.; CORTE, A. P. D.; BEHLING, A.; NETTO, S. P., RODRIGUES, A. L.; SANQUETTA, M. N. Comparison of data mining and allometric model in estimation of tree biomass. **BMC bioinformatics**, v. 16, n. 247, p 1-9, 2015.

SANQUETTA, C. R.; PIVA, L. R.; WOJCIECHOWSKI, J.; CORTE, A. P.; SCHIKOWSKI, A. B. Volume estimation of *Cryptomeria japonica* logs in southern Brazil using artificial intelligence models. **Southern Forests: a Journal of Forest Science**, v. 80, n. 1, p. 29-36, 2018.

SOCHA, J.; KULEJ, M. Variation of the tree form factor and taper in European larch of Polish provenances tested under conditions of the Beskid Sadecki mountain range (Southern Poland). **Journal of Forest Science,** v. 53, n. 12, p. 538-547, 2007.

SOUZA, D. S., NIEVOLA, J. C., SANTOS, J. X., WOJCIECHOWSKI, J., GONÇALVES, A. L., CORTE, A. P. D. & SANQUETTA, C. R. (2019). k-Nearest Neighbor Regression in the Estimation of *Tectona grandis* Trunk Volume in the State of Pará, Brazil, **Journal of Sustainable Forestry**, v. 38, n. 8, p. 755-768, 2019.

TENZIN, J.; WANGCHUK, T.; HASENAUER, H. Form factor functions for nine commercial tree species in Bhutan. **Forestry: An International Journal of Forest Research**, v. 90, n. 3, p. 359-366, 2017.