ISSN 0104-7760
v.26 n.3 2020

CERNE

Jadson Coelho de Abreu[1a+], Carlos Pedro Boechat Soares[2b], Helio Garcia Leite[2c], Daniel Henrique Breda Binoti[2], Gilson Fernandes da Silva[3]

ALTERNATIVES TO ESTIMATE THE VOLUME OF INDIVIDUAL TREES IN FOREST FORMATIONS IN THE STATE OF MINAS GERAIS, BRAZIL

ABREU, J. C. de; SOARES, C. P. B.; LEITE, H. G.; BINOTI, D. H. B.; SILVA, G. F. da. Alternatives to estimate the volume of individual trees in forest formations in the state of Minas Gerais-Brazil. **CERNE**, v. 26, n. 3, p.393-402, 2020

Keywords:
Artificial neural networks
Mixed linear model
Regression
Support-vector machine

HIGHLIGHTS

Evaluation of machine learning algorithms to estimate volume in native forests.

Comparison of machine learning algorithms with fixed and mixed regression models.

Mixed models showed better results than machine learning algorithms.

Machine learning methods were not superior in native forests.

Historic:
Received 06/02/2020
Accepted 07/04/2020

ABSTRACT

The objective of this study was to compare different alternatives to estimate the stem volume of individual trees in four different forest formations in the Minas Gerais state, Brazil. The data were obtained in a forest inventory procedure performed by the Minas Gerais Technological Center Foundation. The stem volumes were computed by the Smalian expression up to the outside bark diameter equal to 4 cm. The volume data of outside bark, diameters (DBH) and total heights were used to fit a Schumacher and Hall equation for each forest formation, considering the structures of the linear fixed and mixed models. Next, 100 Multilayer Perceptron artificial neural networks (ANN) were trained in a supervised manner. In addition, we evaluated eight support-vector machine regression (SVMR). The criteria to evaluate the performance of all the alternatives studied were: the correlation between the observed and estimated volumes, the square root of the mean square error and the frequency distribution by percentage relative error class. After the analyzes, all the alternatives were verified to estimate the volume of the individual trees in the different forest formations. Although the alternatives presented close statistics in the validation process, the graphical analysis of the error distribution showed greater precision of the estimates of the mixed linear models for the four formations. Given the results, it is concluded that there is no absolute superiority of one alternative over the others, and that all of them should be evaluated to find the one which best describes or explains the dataset.

+Correspondence:
jadson.abreu@ueap.edu.br

DOI:
10.1590/01047760202026032728

[1] Amapa State University, Macapá, Amapá, Brazil. ORCID: 0000-0001-9273-7533[a],
[2] Federal University of Viçosa, Viçosa, Minas Gerais, Brazil.
[3] Federal University of Espirito Santo, Department of Forestry Engineering, Jerônimo Monteiro, Espírito Santo, Brazil

INTRODUCTION

One of the main purposes of forest inventories is to estimate tree volumes (Machado and Figueiredo Filho, 2009), in which it is necessary to define the measuring unit at which the volume is to be expressed, the minimum tree inclusion diameter and the way of obtaining the estimates (Soares et al., 2011). In this sense, there are several methods to estimate tree volume, among them: form factors, shape quotient, volume equations, multi-volume or tapering equations (Burkhart and Tome, 2012).

Volume equations are commonly used to estimate the volume of trees in forest inventory procedures due to the accuracy of their estimates. They are an expression in which the wood volume is presented as a function of other quantities or variables of the tree (usually the diameter at breast height (DBH) and the height), which can be directly measured or estimated by non-destructive means (Campos and Leite, 2017). However, the volume equations in tropical native forests are fitted considering all species, thus decreasing their accuracy due to the data heterogeneity. This procedure is common due to a lack of trees of the species in all diameter classes, in most cases preventing the fit of specific volume equations for each of them.

In addition to the above mentioned methods, tree volumes can be estimated through the use of artificial neural networks - ANN (Silva et al., 2009; Görgens et al., 2014; Souza et al., 2018), mixed models (Hall and Clutter, 2004; Gouveia et al., 2015) and support-vector machine regression (SVMR) (Cordeiro et al., 2015; Binoti et al., 2016), which have mainly been explored in commercial species plantations. However, the methodological alternatives are still little used in estimating the tree volume in native forests in Brazil, so that there is a gap which can be filled by developing studies in different natural forest typologies or biomes where it is difficult to estimate the volume of individual trees with precision due to the great heterogeneity of species, as well as the sizes and shapes of the stems/trunks found in these environments.

Machine learning methods can reduce most of the data heterogeneity, as this previously mentioned information can be entered into the model as categorical variables and improve the model accuracy. In addition, this information can be considered as a random effect in a mixed model structure, generating several equations depending on the random parameters.

In view of the above, the hypothesis of this study is that machine learning (ANN and SVMR) methods and the inclusion of random effects in the linear mixed models improve the volume estimate precision in comparison to the Schumacher and Hall model in different forest formations in Minas Gerais state, Brazil.

MATERIAL AND METHODS

Database

The data used in this study came from the forestry inventory carried out by the Minas Gerais Technological Center Foundation in four forestry formations in the State. The following is a brief description of the study formations based on Cetec (1995).

Cerrado *sensu stricto*

This formation comprises the "sensu strictu" cerrado, meaning that it is characterized by typical cerrado vegetation, with a predominance of tree-shrub individuals, and as a rule present tortuous stems, thick bark and a predominant height of 4 to 5 m. This formation has wide geographical distribution, significantly occurring in the northwest, north, jequitinhonha and central-north regions of the state.

Primary forest

Primary forest are forest formations originating in evergreen or semi-deciduous, however, in the present case not considering those formations located in the alluvial plains, marginal to the water courses or in their sources.

They predominantly consist of arboreal elements with high shafts and great diameters, and with a significant occurrence of noble species.

Secondary forest

Secondary forest comprises the strata of *capoeirão*, *capoeira* and *capoeirinha*, being constituted by evergreen or semi-deciduous vegetation formations in different regeneration stages developed from cutting or burning of preexisting virgin forest. This formation comes from the sprouting of primary forest stumps, roots and the germination of previously fallen seeds on the soil. These formations are predominantly located in the south, southeast and northeast regions of the State.

Jaíba transitional forest

Jaiba transitional forest is a forest complex comprising deciduous, semi-deciduous and transition forms between these and hyperxerophilous caatinga, which occurs in the Jaíba area and surroundings. It was distinguished from similar types due to occupying a relatively large area and having very different characteristics from the different types of forests existing in other areas of the State. Its wood yield is approximately equal to that of the mesophilic forest, standing at around 240 st/hectare. The municipalities with the highest occurrence of these formations are: Manga, Itacarambi, Januária and Varzelândia.

The number of sample trees in each formation was established according to the distribution proportional of the trees in the respective diameter classes, totaling 1479 trees, namely: 414 in the Cerrado sensu stricto; 266 in Primary forest; 448 in Secondary forest; 351 in Jaíba Transitional forest.

Information in the rigorous cubing process was collected from all trees to identify the species by measuring the diameters at 1.30 m of height (DBH) and the total heights, counting the number of branches, and measuring the outside bark diameters along the stems. The stem volumes of individual trees were obtained by successive application of the Smalian formula (Soares et al., 2011), considering sections of 1 m length and the minimum commercial outside bark diameter equal to 4 cm. Descriptive statistics were made for all variables by forest formation.

Linear fixed and mixed models

Mixed models are used to model the random parts of forests by including a matrix of variances and covariance (Resende et al., 2014). In addition, this modeling approach analyzes hierarchically structured data more efficiently than other approaches, and can increase the accuracy of the estimates (Hao et al., 2015). These models have three fundamental aspects: the estimation and hypothesis testing of fixed effects, prediction of random effects and estimation of variance components.

The linear mixed model was in the following form (Wu, 2009), where β are fixed effects, b_i are random effects, x_i is a design matrix containing covariates of individual i, z_i, is a design matrix, e$_i$ are random errors, R_i is a n x n variance-covariance matrix within individual measurements (R_i= I x s^2, where I is an identity matrix), and G is the variance-covariance matrix of random effects.

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i, i = 1,2,\ldots,n. \qquad [1]$$

$$b_i \sim N\left(0,G\right), \qquad \varepsilon_i \sim N(0,R_i) \qquad [2]$$

The Schumacher and Hall model (1933) in its linearized form with the outside bark volume, diameters and total tree height (fixed effects) data was initially adjusted for each forest formation, and its functional relationship was determined as per. In which: *Ln* = Napierian logarithm; *V* = commercial volume of the stem including the outside bark, in m^3; *dbh* = outside bark diameter at 1.30m aboveground, in cm; *H* = total height, in m; β_0 to β_2 = model parameters; ε = random error.

$$LnV = \beta_0 + \beta_1 Lndbh + \beta_2 LnH + \varepsilon \qquad [3]$$

The adjustments of the equations referring to model 1 were performed using the Restricted Maximum Likelihood method with the *glm*^{2} package in the R software program (R Development Core Team, 2014).

In order to verify the influence of the inclusion of random effects on the accuracy of the equations, the Schumacher and Hall model (1933) was modified considering the structure of a mixed linear model, including random slopes and inclination coefficients, considering the species in each forest formation as random effects, defining the following model. In which: β_0, β_1 and β_2 = fixed parameters of the model; a$_i$ = random intercept for the ith species; b_{1i}, b_{2i} = random slope coefficients for the ith species.

$$LnV = (\beta_0 + a_i) + (\beta_1 + b_{1i})Lndbh + (\beta_2 + b_{2i})LnH + \varepsilon \qquad [4]$$

The adjustments of the equations referring to the mixed models (models 2) were performed by the Restricted Maximum Likelihood method using the *nlme* package in the R software program (R Development Core Team, 2014).

The variances of the errors in this study were considered to be homogeneous, since the logarithmic transformation of the data usually provides attendance to this assumption of the classical regression model, as well as the covariance of the errors equal to zero using longitudinal data in the analyzes (Gujarati and Porter, 2011).

The results of the inclusion of the random effects on the intercept and slopes of the models were verified using the maximum likelihood ratio (MLR) test (Resende et al., 2014), where the significance of the difference (D) between the *deviances* (-2log(L)) for the models with and without the random effect was verified by comparing the calculated value with the tabulated value by the χ^2 test, with 1 degree of freedom and 5% significance.

Thus, the model selected as the best model for each forest formation at the end of this modeling process could be the complete mixed linear model or a partial model, which means with the random effect only being associated with some parameters of the model, or the model still considering only the fixed effects due to the non-significance of the random effects.

The following evaluation criteria were used in order to avoid personal judgments in evaluating the adjustments of the equations for fixed and mixed effects models, being calculated in the original dependent variable volume unit (m^3): correlation coefficient ($r_{y\hat{y}}$) between the observed and estimated volumes and the root-mean-square error (RMSE) (Silva et al., 2009; and Binoti et al., 2015), and analysis of the relative error percentage.

Artificial neural networks and support-vector machine

ANNs are computational models inspired by the nervous system of living beings. They have the ability to acquire and maintain knowledge (based on information) and can be defined as a set of processing units characterized by artificial neurons, which are interconnected by a large number of interconnections (artificial synapses), and represented by vectors/synaptic weight matrices (Macukow, 2016).

First, 100 Multilayer Perceptron artificial neural networks (ANN) using the Backpropagation and Simulated Annealing training algorithms with the sigmoid activation function were separately trained in a supervised way for the four studied forest formations (70% of the data).

The input variables (inputs) in training the networks (ANN) were: DBH, total height (H), number of branches and species (categorical variable); while the output variable was the volume of the stem outside bark. The stop training criteria adopted for the ANNs were: root-mean-square error (<0.001) or number of cycles (equal to 3000).

Next, eight configurations formed from two error functions and four kernel functions were tested for training the support-vector machine regression (SVMR). The optimized error functions were: type I and type II functions, given by:

Type I function:

Subject to the following restrictions. In which: w = coefficient vector; c = error penalty parameter; ξ, ξ^* = variables that characterize, respectively, the error above and below the e - tube; i = training cases; total number of training cases; $\varphi.(x_i)$ = Kernel used; b = bias; y_i = output data and e = maximum allowed error.

$$\boldsymbol{Minimize}\, \frac{1}{2}.\boldsymbol{w}^T \boldsymbol{w} + \boldsymbol{C}.\sum_{i=1}^{N} \xi_i + \boldsymbol{C}.\sum_{i=1}^{N} \xi_i^* \quad [5]$$

$$w^T.\varphi.(x_i) + b - y_i \leq \varepsilon + \xi_i^* \quad [6]$$

$$y_i - w^T.\varphi.(x_i) - b \leq \varepsilon + \xi_i \quad [7]$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, N \quad [8]$$

Type II function:

Subject to the following restrictions:

In which: v = parameter that regulates the number of support vectors.

$$\boldsymbol{Minimize}\, \frac{1}{2}.\boldsymbol{w}^T \boldsymbol{w} - \boldsymbol{C}\left(\boldsymbol{v}.\boldsymbol{\varepsilon} + \frac{1}{\boldsymbol{N}}.\sum_{i=1}^{N} \left(\xi_i + \xi_i^* \right) \right) \quad [9$$

$$\left(w^T.\varphi.(x_i) + b \right) - y_i \leq \varepsilon + \xi_i \quad [10]$$

$$y_i - \left(w^T \cdot \varphi \cdot (x_i) + b \right) \leq \varepsilon + \xi_i^* \quad [11]$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, N, \varepsilon \geq 0 \quad [12]$$

The Kernel functions evaluated were: linear, polynomial, radial basis function (RBF) and sigmoid (Table 1).

In which: $K(X_i.X_j) = \varphi(X_i), \varphi(X_j)$ and represents the kernel function applied to the input data; g=shape parameter; d = polynomial degree; C = error penalty parameter.

TABLE 1 Kernel functions tested on support-vector machine regression.

Kernel type	Function	Parameters
Linear	$K(X_i.X_j)$	-
Polynomial	$K(X_i.X_j) = (\gamma.X_i.X_j + C)^a$	g, d, C
RBF	$K(X_i.X_j) = e^{\overline{\left(-\gamma \lvert X_i - X_j \rvert^2\right)}}$	g
Sigmoidal	$K(X_i.X_j) = \tanh(\gamma.X_i.X_j + C)$	g, C

The same input variables (inputs) for training the networks (ANN) were considered for training the eight support-vector machine regression (SVMR) configurations: DBH, total height (H), the number of branches, in addition to the variable categorical species; and the stem plus the outside bark volume as the output variable.

All the training of the artificial neural networks (ANN) and support-vector machine regression (SVMR) were performed in NeuroForest 4.06 (Neuroforest, 2017) and R software program (R Development Core Team, 2014), respectively. The evaluation criteria were the same as those used for the regression analysis.

Validation of alternatives

In order to compare the performance between the artificial neural networks (ANNs) and the support-vector machine regression (SVMR) configurations, 30% of the database was used as test samples, i.e. samples not used in of the ANN training and the SVMR adjustment, and then the following statistics were calculated: correlation coefficient between the observed and estimated volumes ($r_{y\hat{y}}$), and the root-mean-square error (RMSE) in percentage (%); as well as a graphical analysis of the distribution of the frequencies per class of errors in percentage (Silva et al., 2009; Binoti et al., 2015).

As there was no data separation for validating the equations for the fixed and mixed models, the comparison between these methodological alternatives and the ANN and SVMR was performed by separating the estimates in the equation adjustment database referring to 30% of the validation database of ANN and SVMR validation and calculating the statistics described above to enable comparison.

RESULTS AND DISCUSSION

Data description

Considering the total number of sample trees used in the analyzes (1479), tree diameters (DBH) ranged from 3.80 to 66.20 cm and heights from 2.30 to 33.40 m (Table 2). The forest formations where the smallest and largest number of trees were covered were the primary forest and the secondary forest, respectively. In terms of the number of species, the Jaíba Transitional Forest and Secondary Forest formations were those with the lowest (36) and the highest (112) amounts, respectively.

Fixed and mixed effects models

The Schumacher and Hall model equation (1933) (only fixed effect) adjusted well in the four formations (Table 3), and the parameter estimates were all statistically significant (p-value <0.05). The inclusion of the species as a random effect in the mixed model structure was significant in the four formations. It should be noted that only the random coefficient associated with the height variable was not significant for the Cerrado (p-value> 0.05). All three coefficients were significant for the other formations (p-value <0.05).

Artificial neural networks (ANNs) and support-vector machine regression (SVMR)

The ANN and SVMR results by forest formation are presented in Table 4. The ANN and SVMR configurations which presented the best training statistics are provided. The Backpropagation algorithm was the best for the ANN related to the Cerrado formation. The best training algorithm for the other formations was Simulated Annealing, with 1 neuron in the hidden layer.

The best SVMR for the Cerrado training was with the RBF kernel function and type II optimized error function (RBF-II) (Table 4). The kernel function selected as the best for the other formations was the Polynomial, also with the optimized type II error function (Polynomial-II).

The ANN configurations which had species as categorical variables did not present the best training statistics, and the ANN configurations which presented the best statistics did not have a categorical variable. Inclusion of the species only improved the results in the secondary forest and transient forest of Jaíba for the SVMR.

The SVMR presented lower errors (RMSE) and higher correlations between the observed and estimated volumes in the Cerrado, Secondary forest and Jaíba Transitional forest formations, while only in the Primary Forest formation for ANN.

Validation of the alternatives

In the validation process of the evaluated methodologies, a high correlation (Table 5) was observed between the observed and estimated volumes, with values between 96.74% and 98.68%. In terms of root-mean-square error (RMSE), the estimates were from 19.83% to 51.20% for the evaluated methodological alternatives.

The histograms of the percentage error frequencies (Figure 1) show that the mixed linear model was the alternative with the best performance in the validation process for the four formations, with a concentration of errors close to zero and lower amplitude of distribution.

TABLE 2 Descriptive statistics of tree-sample dendrometric variables for four forest formations in the state of Minas Gerais, Brazil.

Formation	Variables	Minimum	Maximum	Mean ($\bar{X}$)	*Deviation (S)*	N°. of Observations	Species
Cerrado *sensu stricto*	*DBH* (cm)	3.8000	29.3000	8.9804	4.3195	414	70
	H (m)	2.3000	18.3000	5.9326	2.1215		
	N°. of branches	0.0000	38.0000	2.5338	3.6780		
	Volume (m^3)	0.0027	0.7413	0.0428	0.0748		
Primary forest	*DBH* (cm)	4.5000	66.2000	17.0094	10.5126	266	98
	H (m)	4.8000	33.4000	14.1150	5.3599		
	N°. of branches	0.0000	52.0000	5.7143	7.8827		
	Volume (m^3)	0.0046	5.4508	0.3612	0.6042		
Secondary forest	*DBH* (cm)	3.5000	49.7000	12.0746	6.0389	448	112
	H (m)	4.9000	24.1000	11.0638	3.0575		
	N°. of branches	0.0000	70.0000	4.1942	6.6028		
	Volume (m^3)	0.0031	2.0284	0.1188	0.2097		
Jaíba Transition forest	*DBH* (cm)	3.8000	42.7000	13.4071	6.2108	351	36
	H (m)	3.9000	17.0000	10.3416	2.7126		
	N°. of branches	0.0000	29.0000	4.9202	4.6763		
	Volume (m^3)	0.0040	1.2883	0.1422	0.1647		

TABLE 3 Estimates of fixed effects parameters; correlation coefficient between the observed and estimated volumes in the original measurement scale, in m^3, ($r_{y\hat{y}}$); root-mean-square error (RMSE); maximum likelihood ratio (MLR) estimates for random effects; and level of significance (p-value).

Formation	Model/effect	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$r_{y\hat{y}}$(%)	RMSE (%)	MLR	p-value
Cerrado	Fixed	-9.33578	2.18800	0.53568	97.85	39.72		
	Mixed	-9.31700	2.16848	0.55075	98.19	33.41	4.80	0.0284
Primary forest	Fixed	-9.60426	2.15763	0.69592	98.18	42.69		
	Mixed	-9.60039	2.16958	0.68031	98.64	24.31	12.19	0.0068
Secondary forest	Fixed	-9.52586	1.98448	0.85141	97.04	48.07		
	Mixed	-9.52776	1.97831	0.85705	97.95	35.69	10.70	0.0135
Jaíba Transitional forest	Fixed	-9.65836	2.09329	0.82044	98.08	28.00		
	Mixed	-9.67200	2.08317	0.82635	98.64	19.06	20.13	<0.001

TABLE 4 Structure of artificial neural networks (ANNs) and support-vector machine regression (SVMR) selected in the training process for four forest formations in the state of Minas Gerais and their respective statistics in the training process.

Forest Formation	Method	Structure/ Type	Categorical variables	$r_{y\hat{y}}$ (%)	RMSE (%)
			Species		
Cerrado	ANN	4-22-1	N	97.28	42.85
	SVMR	RBF-II	N	98.48	32.33
Primary forest	ANN	4-1-1	N	99.22	22.22
	SVMR	Polynomial-II	N	95.22	70.01
Secondary forest	ANN	4-1-1	N	98.51	28.70
	SVMR	Polynomial-II	Y	99.53	16.06
Jaíba Transition forest	ANN	4-1-1	N	98.90	18.45
	SVMR	Polynomial-II	Y	99.29	14.63

In which: Y or N represents the presence or not of the species and/or forest formation as a categorical variable.

DISCUSSION

In analyzing the significance of the coefficients of the fixed effect model in Table 3, we can confirm the importance of the explanatory variables of diameter and height in the volumetric model (Calegario et al., 2005). Similar results were found by Chicorro et al. (2003); Scolforo et al. (2008); Rufini et al. (2010) and Stolariková et al. (2014). All parameters of the equations were statistically significant (p-value < 0.05).

The model predicts random coefficients for each species (BLUP) with the inclusion of random factors, and instead of the regression curve tending towards the sample mean, it predicts a curve for each species, i.e. the random factors. The values which would be incorporated into the model error end up being incorporated and explained by random factors.

Equations adjusted for a group of species or forest formation are more common due to the lack of tree-samples for all forest species (Huff et al., 2018). If response variable information is available for a new species in a mixed model, then random coefficients are obtained and estimated by considering the species-specific response rather than just an average (or expected) response to the population. In the mean response of the population, it is assumed that the vector of the random coefficients for a new individual has an expected value of zero (Burkhart and Tomé, 2012).

The performance of a mixed model may be better than that of the model with only fixed effects when a sample is available to predict the random parameters (BLUP) (Temesgen et al., 2008; Huff et al., 2018). This behavior was observed for the four studied forest formations by including the species as a random effect in the volumetric model. Other variables can be inserted as random effects in addition to this variable, such as region and local quality classes (Ou et al., 2016), precipitation, soil, elevation and other geographical characteristics observed to increase the accuracy of the estimates (Meng et al., 2007).

The histogram of the distribution of residuals confirms the good performance of the mixed model (Figure 1), and it should be noted that there was a greater dispersion of errors for the Cerrado formation (up to -60%). However, this can be considered insignificant since it is a small number of estimates compared to the sample, and does not strongly interfere with the model accuracy (Costa et al., 2012). This result confirms our second hypothesis, namely that the inclusion of the species as a random effect would improve the estimate of the Schumacher and Hall model.

Although the artificial neural networks (ANNs) were accurate in the training phase, their performance was not good over the test data in the validation process. In the distribution of errors (Figure 1), it was observed that the networks overestimated the smaller volumes and their performance in some formations was lower than the Schumacher and Hall model (fixed effect model). This result is already in line with one of our hypotheses,

TABLE 5 Artificial neural network (ANN) structures, support-vector machine regression (SVMR), fixed and mixed effects models selected in the training process for four of the forest formations in the State of Minas Gerais and their respective statistics in the validation process.

Forest Formation	Method	Structure/	Variables	$r_{y\hat{y}}$(%)	RMSE
		Type	Species		
Cerrado	ANN	4-22-1	N	98.00	31.30
	SVMR	RBF-II	N	97.60	37.99
	Regression	Fixed	N	96.74	41.35
	Regression	Mixed	Y	97.43	35.47
Primary forest	ANN	4-1-1	N	98.19	29.07
	SVMR	Polynomial-II	N	98.57	24.67
	Regression	Fixed	N	98.65	35.57
	Regression	Mixed	Y	98.68	28.15
Secondary forest	ANN	4-1-1	N	97.60	43.85
	SVMR	Polynomial-II	Y	96.93	51.20
	Regression	Fixed	N	97.93	42.84
	Regression	Mixed	Y	98.27	37.92
Jaíba Transitional forest	ANN	4-1-1	N	97.70	22.80
	SVMR	Polynomial-II	Y	98.20	19.83
	Regression	Fixed	N	97.27	27.43
	Regression	Mixed	Y	97.87	21.86

In which: Y or N represents the presence or not of the species and/or forest formation as categorical variable or random effect.

in which machine learning methods would present the best results. Similar results were found by Görgens et al. (2015), whose scatter plots of the eucalyptus plantations obtained by ANN showed overestimation for lower volumes and underestimation for larger ones when compared to Random Forest, a support-vector machine regression. These results are probably due to some algorithms having difficulty learning lower values, and that they end up overestimating the lowest values when the residual is calculated in relative form. Araújo (2015) found good results for ANN for this same data set, but using different combinations of DBH and height variables, as well as different algorithms such as: NEAT and *Skyp Layer Connections*.

The difference in ANN and SVMR performance in the training and validation process evidences the careful separation of data in these two modeling phases. In this study, the data selection was performed in a random manner and because there were no sample trees in all diameter classes and in all forest formations, and so the data were unbalanced in such a way that good training and not such good validation were performed. In the

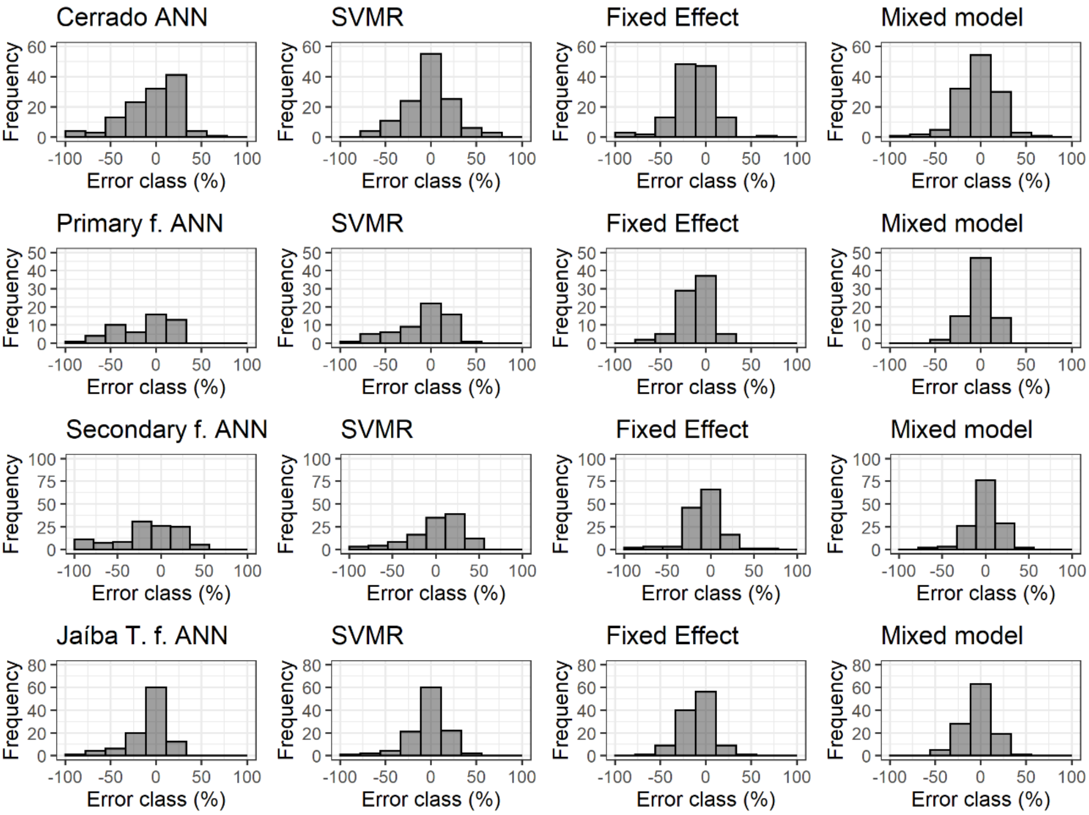

FIGURE 1 Histogram of frequency by error class for the methodological alternatives: artificial neural networks (ANN), support vector machine regression (SVMR), regression - fixed effect, and regression - mixed model for four forest formations in Minas Gerais state.: $\text{error}(\%) = \frac{(y-\hat{y})}{y}.100$.

case of regression models (fixed and mixed effects), which can be adjusted using the method of least squares and maximum likelihood, the estimates refer to the mean values (expected) (Gujarati and Porter, 2011), and therefore presented better results than ANN and SVMR.

Superior results with regression models to ANNs have already been found, such as by Özçelik et al. (2013) who modeled the hypsometric relationship of *Juniperus excelsa* in southern and southwestern Turkey using mixed models; and Mendonça et al. (2018), who adjusted fixed height models for *Zeyheria tuberculosa*. This shows that although ANN is an efficient and accurate method for predicting dendrometric variables, other methodologies may also sometimes present better results.

The support-vector machine regression (SVMR) was statistically the best in the training phase. However, the SVMR presented the same problems as ANN in the validation phase. Although they presented good accuracy, they could not "learn" to estimate small volumes in some formations.

Cordeiro et al. (2015) found more accurate volume estimates for SVMR compared to the Schumacher and Hall model in an *Acacia mangium* plantation in the state of Amapá; however, there was no separation of data for training and validation in this study. Costa Filho (2019) found higher precision of the SVMR to estimate height for *Pinus taeda* plantations in São Paulo. Binoti et al. (2016) found close values between the estimates obtained by the Schumacher and Hall model (regression) and SVMR, also for eucalyptus plantations in the South of Bahia. It is observed that all these works were for artificial plantations with little variation in the data. For unequal forests, we can mention Abreu et al. (2017), who found a good fit to estimate volume with a support vector machine in semi-deciduous seasonal forest in the state of Minas Gerais; and Montaño (2016), who in addition to modeling dendrometric variables for commercial plantations, also used SVMR for dry biomass in tropical forests. However, the SVMR in the latter study were lower than the allometric models suggested by Chave et al. (2014).

As a result, we can infer that machine learning methods in artificial plantations are much better compared to traditional methods. However, this does not always happen when the data comes from native forests.

If there were more individuals per species, they would probably have a positive influence on the results, because despite a large number of species, most of them are represented by a single individual, and have situations that they are in training and are not in validation, or vice and vice versa. In addition, we can insert more categorical variables in future studies such as relief, climate, and soil, among others.

CONCLUSION

With the development of this study, it was verified that all the alternatives were precise in estimating the volume of the individual trees in the different forest formations.

Despite presenting satisfactory results, the machine learning methods were not superior to the mixed model and the conventional regression model.

The species as a random effect considerably improved the fit of the Schumacher and Hall model.

Given the results, it is concluded that there is no absolute superiority of a methodological alternative to the others, and that they should all be evaluated to find the one which best describes or explains the dataset.

ACKNOWLEDGEMENTS

The authors acknowledge the Federal University of Viçosa for funding this research and the Minas Gerais Technological Center Foundation for making the data available.

REFERENCES

ABREU, J. C; SOARES, C. P. B; LEITE, H. G. Assessing alternatives to estimate the stem volume of a seasonal semi-deciduous forest. **Revista Floresta**, Curitiba, v. 47, n. 4, p. 375-382, dez. 2017.

ARAÚJO, A. C. S. C. Estimativa volumétrica de formações florestais no estado de Minas Gerais com redes neurais artificiais. 2015. 70 p. (Dissertação – Mestrado em Ciência Florestal) – Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2015.

BINOTI, M. L. M. S.; LEITE, H. G.; BINOTI, D. H. B.; GLERIANI, J. M. Prognose em nível de povoamento de clones de eucalipto empregando redes neurais artificiais. **CERNE**, v. 21, p. 97-105, 2015.

BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G.; ANDRADE, A. V.; NOGUEIRA, G. S.; ROMARCO, M. L.; PITANGUI, C. G. Support vector machine to estimate volume of eucalypt trees. **Revista Árvore**, Viçosa, v.40, n.4, p.689-693, 2016.

BURKHART, H. E.; TOMÉ, M. Modeling forest trees and stands. Dordrecht, Springer, 2012. 458p.

CALEGARIO, N.; MAESTRI, R.; LEAL, C. L.; DANIELS, R. F. Estimativa do crescimento de povoamentos de *Eucalyptus* baseada na teoria dos modelos não lineares em multinível de efeito misto. **Ciência Florestal**, Santa Maria, v. 15, n. 3, p. 285-292. 2005.

CAMPOS, J. C. C.; LEITE, H. G. **Mensuração florestal**: perguntas e respostas. 5 ed. Viçosa, Editora UFV, 2017. 636p.

CENTRO TECNOLÓGICO DE MINAS GERAIS. Determinação de equações volumétricas aplicáveis ao manejo sustentado de florestas nativas no estado de Minas Gerais e outras regiões do país**: relatório final**. Belo Horizonte, 1995.

CHAVE, J., REJOU-MECHAIN, M., BURQUEZ, A., CHIDUMAYO, E., COLGAN, M., DELITTI, W., DUQUE, A., EID, T., FEARNSIDE, P., GOODMAN, R., HENRY, M., MARTINEZ-YRIZAR, A., MUGASHA, W., MULLER-LANDAU, H., MENCUCCINI, M., NELSON, B., NGOMANDA, A., NOGUEIRA, E., ORTIZ-MALAVASSI, E., PELISSIER, R., PLOTON, P., RYAN, C., SALDARRIAGA, J. E VIEILLEDENT, G. . Improved allometric models to estimate the aboveground biomass of tropical trees**. Global Change Biology**, v.20, n.10, p:3177–3190, 2014.

CHICORRO, J. F.; RESENDE, J. L. P.; LEITE, H. G. Equações de volume e de taper para quantificar multiprodutos da madeira em floresta atlântica. **Revista Árvore**, Viçosa, v.27, n.6, p.799-809, 2003.

CORDEIRO, M. A.; PEREIRA, N. N. J.; BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE. H. G. Estimativa do volume de *Acacia mangium* utilizando técnicas de redes neurais artificiais e máquinas vetor de suporte. **Pesquisa florestal brasileira.** Colombo, v. 35, n. 83, p. 255-261, 2015.

COSTA, T. R.; CAMPOS, L.; CYSNEIROS, F. J. A.; CUNHA FILHO, M. Modelos lineares mistos: uma aplicação na curva de lactação de vacas da raça Sindi**. Revista Brasileira de Biometria**. São Paulo, v. 30, n. 1, p. 23 - 49, 2012.

COSTA FILHO, S. V. S.; ARCE, J. E.; MONTANÃ, R. N. R.; PELISSARI, A. L. Configuração de algoritmos de aprendizado de máquina na modelagem florestal: um estudo de caso na modelagem da relação hipsométrica. **Ciência Florestal**, Santa Maria, v. 29, n. 4, p. 1501-1515, 2019.

GOUVEIA, J. F.; SILVA, J. A. A.; CARACIOLO, R. L. C.; GADELHA, F. H. L.; LIMA FILHO, L. M. A. Modelos volumétricos mistos em clones de *Eucalyptus* no polo gesseiro do Araripe, Pernambuco. **Revista Floresta**, Curitiba, v. 45, n. 3, p. 587 - 598, 2015.

GÖRGENS, E. B.; LEITE, H. G.; GLERIANI, J. M.; SOARES, C. P. B.; CEOLIN, A. Influência da arquitetura na estimativa de volume de árvores individuais por meio de redes neurais artificiais. **Revista Árvore,** Viçosa, v. 38, p. 289-295, 2014.

GÖRGENS, E. B.; MONTAGHI, A.; RODRIGUEZ, L. C. E. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. **Computers and Electronics in Agriculture**, v. 116, p: 221-227, 2015.

GUJARATI, D. M.; PORTER, D. C. **Econometria Básica**. 5.ed. Buckman. 2011. 920p.

HALL, D. B.; CLUTTER, M; Multivariate Multilevel Nonlinear Mixed Effects Models for Timber Yield Predictions. **Biometrics**, v.60, p.16–24, 2004.

HAO, X.; YUJUN, S.; XINJIE, W.; JIN, W.; YAO, F., Linear mixed-effects models to describe individual tree crown width for China-Fir in Fujian Province, **Southeast China.** PLoS ONE v.10, n.4, 2015.

HUFF, S.; POUDEL, K. P.; RITCHIE, M.; TEMESGEN, H. Quantifying aboveground biomass for common shrubs in northeastern California using nonlinear mixed effect models. **Forest Ecology and Management**, v.424, p:154–163, 2018.

MACHADO, S. D.; FIGUEIREDO FILHO, A. **Dendrometria.** 2ed. Editora unicentro, 2009. 316p.

MACUKOW, B. Neural Networks – State of Art, Brief History, Basic Models and Architecture**. Lecture Notes in Computer Science**, v. 9842, pp. 3–14, 2016.

MENDONÇA, A. R.; SILVA, J. C.; AOZANI, T. S.; SILVA, E. R.; SANTOS, J. S.; BINOTI, D. H. B.; SILVA, G. F. Estimação da altura total de árvores de ipê felpudo utilizando modelos de regressão e redes neurais artificiais. **Revista Brasileira de Biometria**, Lavras, v.36, n.1, p.128-139, 2018.

MENG, Q.; CIESZEWSKI, C. J.; MADDEN, M.; BORDERS, B. A linear mixed-effects model of biomass and volume of trees using Landsat ETM+ images. **Forest Ecology and Management,** v. 244, p:93-101, 2007.

MONTAÑO, R. A. N. R. Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal. 2016**.** 102f. Tese (Doutorado), Curso de informática, Universidade Federal do Paraná, 2016.

NEUROFOREST. Disponível em: <http://neuroforest.ucoz.com/>. Acesso em: 07 julho. 2017.

ÖZÇELIK, R.; DIAMANTOPOULOU, M. J.; CRECENTE-CAMPO, F.; ELER, U. Estimating *Crimean juniper* tree height using nonlinear regression and artificial neural network models. **Forest Ecology and Management,** v.306, p.52–60, 2013.

OU, G.; WANG, J.; XU, H.; CHEN, K.; ZHENG, H.; ZHANG, B.; XUELIANO, S.; XU, T.; XIAO, Y. Incorporating topographic factors in nonlinear mixed-effects models for aboveground biomass of natural Simao pine in Yunnan, China. **Journal of Forestry Research**, v. 27, n.1, p:119–131, 2016.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. URL http://www.R-project.org/

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência. **Suprema gráfica e Editora Ltda**, 2014. 881p.

RUFINI, A. L.; SCOLFORO, J. R. S.; OLIVEIRA, A. D.; MELLO, J. M. Equações volumétricas para o cerrado sensu stricto, em Minas Gerais. **CERNE,** Lavras, v. 16, n. 1, p.1-11. 2010.

SOARES, C. P. B.; PAULA NETO, F.; SOUZA, A. L. **Dendrometria e inventário florestal**. 2ed. Viçosa, Editora UFV, 2011. 272p.

SOUZA, S. R. R.; SILVA, J. A. A.; FERREIRA, T. A. E.; GUERA, O. G. M. Redes neurais para estimativa volumétrica de clones de eucalyptus spp. No pólo gesseiro do Araripe. **Revista Brasileira de Biometria**, Lavras, v.36, n.3, p.715-729, 2018.

SCHUMACHER, F. X.; HALL, F. D. S. Logarithmic expression of timber-tree volume. **Journal of Agriculture Research**, v. 47, n. 9, p. 719-734, 1933.

SILVA, M. L. M.; BINOTI, D. H. B.; GLERIANI, J. M.; LEITE, H. G. Ajuste do modelo de Schumacher e Hall e aplicação de redes neurais artificiais para estimar volume de árvores de eucalipto. **Revista Árvore**, Curitiba, v. 33, n. 6, p. 1133-1139, 2009.

STOLARIKOVÁ, R.; ŠÁLEK, L.; ZAHRADNÍK, D.; DRAGOUN, L.; JEŘÁBKOVÁ, L.; MARUŠÁK, R.; MERGANIČ, J. Comparison of tree volume equations for small-leaved lime (*Tilia cordata* Mill.) in the Czech Republic. **Scandinavian Journal of Forest Research**, v. 29, n. 8, p:757–763, 2014.

TEMESGEN, H.; MONLEON, V.J.; HANN, D.W. Analysis and comparison of nonlinear tree height prediction strategies for Douglas-fir forests. **Canadian Journal of Forest Research**, v.38, p: 553–565, 2008.

WU, L. **Mixed effects models for complex data**. New York: Chapman and Hall/CRC, 2009.