BOSQUE 41(3): 353-364, 2020 DOI: 10.4067/S0717-92002020000300353

Reduction of sampling intensity in forest inventories to estimate the total height of eucalyptus trees

Reducción de la intensidad de muestreo en inventarios forestales para estimar la altura total de eucaliptos

Daniel Dantas [a]*, Luiz Otávio Rodrigues Pinto [a], Marcela de Castro Nunes Santos Terra [a], Natalino Calegario [a], Marcio Leles Romarco de Oliveira [b]

*Corresponding autor: [a] Federal University of Lavras, Departament of Forest Sciences,
Lavras, Minas Gerais, Brazil, tel.: 5538991237493, dantasdaniel12@yahoo.com.br

[b] Federal University of the Jequitinhonha and Mucuri Valleys, Departament of Forest Engineering,
Diamantina, Minas Gerais, Brazil.

SUMMARY

This study aimed at evaluating the performance of different models based on Artificial neural networks (ANN) to estimate the total height of eucalyptus trees (*Eucalyptus* spp.), reducing the number of measurements in the field. Forty-eight ANN were tested, different from each other by the number of trees used as training sample, number of trees used to calculate the dominant height and use of variables (a) categorical, (b) categorical and continuous and (c) continuous, except for the diameter at 1.30 meters above the ground (DBH), used in all combinations. Estimates of height obtained by ANN were compared with values observed and estimates obtained by a hypsometric model. The ANN that showed the best results were used for the height estimation in forest inventory data for further application in the Schumacher and Hall volumetric model. The proposed models were efficient to estimate the total height of eucalyptus trees and allowed the expressive reduction of the number of trees to be measured in forest inventory. The best model found is composed of five trees as training sample, one as test sample and one as validation sample; dominant height coming from the height of the tallest tree in the plot; categorical variable Clone and continuous variables DBH, DBH dominant and basal area of the plot.

Key words: artificial neural network, machine learning, stem volume, Schumacher and Hall.

RESUMEN

El objetivo fue evaluar el desempeño de diferentes modelos basados en Redes Neuronales Artificiales (RNA) en la estimación de la altura total de los eucaliptos, reduciendo el número de mediciones en el campo. Se analizaron 48 RNA, diferentes entre sí por el número de árboles utilizados como muestra de entrenamiento; número de árboles utilizados para calcular la altura dominante; y el uso de (a) variables categóricas, (b) categóricas y continuas y (c) continuas, con la excepción del diámetro a 1,30 m del suelo (DAP), utilizadas en todas las combinaciones. Las estimaciones de altura obtenidas por RNA han sido comparadas con los valores observados y con las estimaciones obtenidas por un modelo hipsométrico. Las RNA que presentaron los mejores rendimientos se utilizaron para estimar la altura en los datos del inventario forestal, para el cálculo posterior del volumen de cada árbol. Los modelos propuestos demostraron ser eficientes para estimar la altura total de los eucaliptos y permitieron la reducción expresiva de la cantidad de árboles que se medirán en el inventario forestal. El mejor modelo encontrado se compone de cinco árboles como muestra de entrenamiento, uno como muestra de prueba y uno como muestra de validación; altura dominante desde la altura del árbol más alto en la parcela; variable categórica clon; y variables continuas DAP, DAP dominante y área basal de la parcela.

Palabras clave: redes neuronales artificiales, altura dominante, Schumacher y Hall.

INTRODUCTION

In forest surveys, some dendrometric variables are measured in the field, highlighting the diameter measured at 1.30 m above the ground (DBH) and total height. DBH is considered as the main variable, since it is a direct measure and easy to obtain. Total height is another variable of great importance, where its measurement is taken indirectly and presents itself as a difficulty in the surveys due to factors such as the difficulty in visualizing the top of trees, time required to complete measurements, among others. These factors, in addition to interfering with the accuracy of measurements, significantly affect the cost of forest inventories.

In 1957, Ker and Smith proposed the use of hypsometric relationships, in which, by measuring the diameters

(DBH) and the heights of some trees in the plot, a height-diameter curve (hypsometric relationship) is obtained and the height of the others can be estimated. Since then, several models for height prediction have been proposed and can be found in literature (Curtis 1967, Inoue and Yoshida 2004, Campos and Leite 2009).

It is known that the quality of hypsometric relationships is influenced by several factors besides DBH, such as forest sites, age, genetic material, silvicultural tracts, among others. The inclusion of these factors in hypsometric models can lead to a gain in the quality of estimates and in biological realism. However, the modeling and quantification of the influences of these characteristics on the variable to be estimated makes this inclusion difficult, since the relations present non-linear characteristics or qualitative (categorical) values (Binoti 2012).

With the advancement of evolutionary computing and the spread of artificial intelligence, artificial neural networks (ANN) have been widely used as an alternative to hypsometric models, for the modeling and prognosis of forest yield. Dantas *et al.* (2020) assessed the quality of the volumetric estimation of *Eucalyptus* spp. trees using machine learning and observed a marked decrease in residual standard error, from 0.0142 m³ (7.9830 %) in the nonlinear fixed-effects regression model to 0.0024 m³ (0.6060 %) in ANN. Freitas *et al.* (2020) evaluated ANN to estimate eucalyptus productivity as a function of environmental variables and what was obtained was ANN with correlation between the estimated and observed mean annual increment of eucalyptus stands at six years of age higher than 85 % and root mean square error less than 15 %.

ANN is an algorithm based on simple processing units (artificial neurons), mimicking the neurons found in the human brain, which calculates specific functions Braga *et al*. (2007). These units are distributed in layers and connected to each other by weights that store the experimental knowledge and weight the inputs of each unit. With that, the acquired knowledge becomes available for use.

The most striking features in ANN are the ability to learn and generalize information. In other words, ANN are able, through a learned example, to generalize the knowledge assimilated to a set of unknown data. Another interesting feature is the ability to extract non-explicit features from a set of information that is provided as examples (Gorgens *et al*. 2015).

One aspect that must be considered, with the adoption of ANN as a modeling tool in forest management, is the possibility of reducing the number of measurements necessary for training the networks, without losing the quality of the estimates. This would result in a decrease in data collection time and cost of forest inventories.

One of the most important pieces of information to determine the potential of a forest in a given region is the variable "volume," the accurate quantification of which is essential in forest management planning. The individual volume serves as a starting point for assessing the wood content in a forest stand and provides support for decisions related to silvicultural practices and timber harvesting and transport. Thus, it is essential that the volume of trees be correctly determined to provide an accurate representation of the sampled population.

The search for methodologies that provide exact estimates and, at the same time, make it possible to reduce the cost and time of measurements is constant, requiring studies that provide subsidies for the manager in the processing of forest inventory data. In this sense, the objective of this work is to propose and evaluate the performance of different models based on ANN in estimating the total height of eucalyptus trees and estimating the total volume in eucalyptus stands. It is proposed as hypothesis that the use of artificial neural networks allows the reduction of the number of heights measurements in forest inventories, without losing the accuracy of the estimates.

METHODS

Data base. The study area consists of 28 management units with four different *Eucalyptus* spp. clones. (MG01, MG02, MG03 and MG04), in the municipality of Minas Novas, Minas Gerais, Brazil, totaling 900 hectares. The climate of the region is characterized as tropical dry climate, Aw type, according to the Köppen climate classification, with average annual temperature of 22.2 °C, with dry winters, and rainy summers with high temperatures. Average annual total precipitation is 961 mm (Alvares *et al.* 2013).

The data for this study came from forest inventories in plantations aged 4 years, planted at 3 x 3 m spacing. In the forest inventory, 100 rectangular sample units with an area of 870 m² were measured. A total of 9,378 individuals were measured. In each plot, the diameter, in centimeters, at 1.30 m above the ground (DBH) of all trees was measured; the total height (Ht), in meters, of 20 trees; and the total height, in meters, of the five dominant and codominant trees (Hd).

Data processing. For processing, three different forms of dominant heights were considered for each plot: using the highest tree in the plot (Hd1), the average of the two tallest trees (Hd2) and the average of the three tallest trees (Hd3); the basal area of the plot (Gparc), in m²; and the dominant DBH (DBHd) of each plot, resulting from the average DBH of the five dominant trees.

Four groups were created, different from each other by the number of trees in each plot used in the training of ANN: (a) G1, consisting of one tree as a training sample, one as a test sample and one as a validation sample; (b) G3, consisting of three trees as a training sample, one as a test sample and one as a validation sample; (c) G5, composed of five trees as a training sample, one as a test sample and one as a validation sample; (d) GT, composed of all trees (except the trees of the test and validation samples) as a training sample, one as a test sample and one as a validation sample. None of the groups contained the trees that were used to calculate the dominant height of each plot. In addition, the trees in the test, validation and training samples were different so that, during training, the network used the specified number of trees in each sample.

Training of artificial neural networks. To obtain ANN to estimate the total height of the trees, the ANN were trained. This procedure consisted of adjusting their weights, using a learning algorithm that extracts characteristics from the data and aims at generating a network that performs the task of interest (Binoti *et al.* 2014). The training was performed in R, version 3.4.1, using the neuralnet package (Günther and Fritsch 2010).

Trained ANN were Multilayer Perceptron (MLP) networks, consisting of an input layer, an intermediate layer, and an output layer. The algorithm used was resilient backpropagation, where the learning rate was set automatically by the package neuralnet, with values ranging from 0.01 to 1.12. The number of neurons in the intermediate layer was chosen using the k-fold. This methodology randomly subdivides the database into k subgroups (Wong *et al.* 2017). The k value was 10 subgroups, with 90 % for training and 10 % for testing (Diamantopolou 2010), applying cross validation. Different numbers of neurons, ranging from 1 to 20, were tested.

The activation function used was logistic (or sigmoid), with an interval from 0 to 1, which limits the amplitude of outputs and inputs. Therefore, data were normalized, which consisted of transforming the values of each variable into values ranging from 0 to 1, using equation [1] (Soares et al., 2011). This equation considers the minimum and maximum value of each variable in the value transformation, maintaining the original data distribution (Valença 2010).

$$x' = \frac{(x - x_{min}) * (b - a)}{(x_{max} - x_{min})} + a \qquad [1]$$

Where:
x': normalized value.
x: original value.
x_{min}: minimum value of the variable.
x_{max}: maximum value of the variable.
a: lower limit of the normalization range.
b: upper limit of the normalization range.

The stopping criterion of the ANN training process was a maximum number of 100,000 cycles, or a mean squared error less than 1 %, stopping the training when meeting one of the criteria. At the end of the training, the best ANN were selected, based on the smallest mean squared error.

In each group (G1, G3, G5 and GT) 12 combinations among variables were obtained. Thus, ANN training sessions were: (a) three sessions without the dominant height as one of the input variables, (b) three sessions with the dominant height Hd1, (c) three sessions with the dominant height Hd2 and (d) three sessions with the dominant height Hd3. Each set of training sessions was subdivided as follows: (i) one session with the categorical variable Clone and the continuous variable DBH, (ii) one session with the categorical variable Clone and the continuous variables DBH, DBHd and Gparc, (iii) one session without categorical variables and with continuous variables DBH, DBHd and Gparc. In each session, 50 ANN were trained, and the network with the best performance was retained based on the training, testing and validation values of determination coefficient and sum of squares of errors. At the end of this process, 48 networks were obtained, one for each combination between the four groups and the 12 training sessions.

To assess the quality of ANN training, the heights of all individuals used in the network adjustments were estimated and the Bias % and the Root Mean Square Error (RMSE %) were calculated (Siipilehto 2000, Leite and Andrade 2002).

Bias and RMSE are used as a parameter in choosing the networks that showed the best performance in the training phase, however, this does not guarantee that they will be able to make a good generalization in an unknown database. To assess the performance of the ANN in generalization, the five best ANN were selected, based on the values of Bias and RMSE, and these were used to estimate the heights of the trees that had no height measured in the field. For the remainder, the heights measured in the inventory were maintained.

Hypsometric and volumetric models. The hypsometric model cited by Campos and Leite (2009) [2] was adopted as a reference in estimating heights, due to its good performance, which can be attributed to the use of dominant height as one of the variables independent of the model (Leite and Andrade 2003). For this, the model was adjusted by Clone using all trees with measured height. The dominant height used was obtained by the average of the five trees with the highest height in each plot, since this is the standard procedure already adopted by the forest company. After adjusting the model, it was applied to the inventory data to estimate the height of trees that had no height measured in the field.

After estimating heights, using the hypsometric model and the five best ANN, linear equations of Schumacher and Hall (1933) [3] were adjusted by Clone, using taper data obtained by accurately estimating the cubic volume of 159 trees, in which Ht, DBH and diameters were measured at the base of the trees (at 0.1 m high) and at heights of 0.5 m, 1 m, 1.5 m and 2 m and, from this section, every 2 m. Individual volumes were obtained using the Smalian formula. The adjusted equations were applied to the inventory data and the volumes of each stem were estimated, considering the six heights obtained for each one (one height estimated by the hypsometric model and five heights estimated by the five best ANN). Finally, the volume per hectare for each plot was estimated and the average volume per hectare for each plot was estimated.

$$\ln(Ht) = \beta_0 + \beta_1(\frac{1}{DBH}) + \beta_2\ln(Hd) + e \qquad [2]$$

$$\ln(vol) = \beta_0 + \beta_1\ln(DBH) + \beta_2\ln(Ht) + e \qquad [3]$$

Where:
Ht = total height, in meters.
DBH = diameter, in centimeters, at 1.30 m in height of the tree.
Hd = dominant height, in meters, from the average height of the 5 highest trees in the plot.
β_0, β_1 and β_2 = parameters of the model.
e = random error.
vol = volume in m³.

Evaluation of estimates. The quality of the height estimates was evaluated in the calculation of the total volume, per sample plot and per management unit, the Average Relative Error (ERM) between estimated volumes (Vest), from the heights estimated by the five ANN, and the observed volume (Vobs), derived from the heights estimated by the hypsometric model; distribution graphs of estimated and observed volumes; and the correlation coefficients between estimated and observed volumes.

RESULTS

The networks trained with the dominant height as an input variable showed better performance and less square sum of errors (figure 1) in the phases of training, testing and validation by the software used, in all four groups.

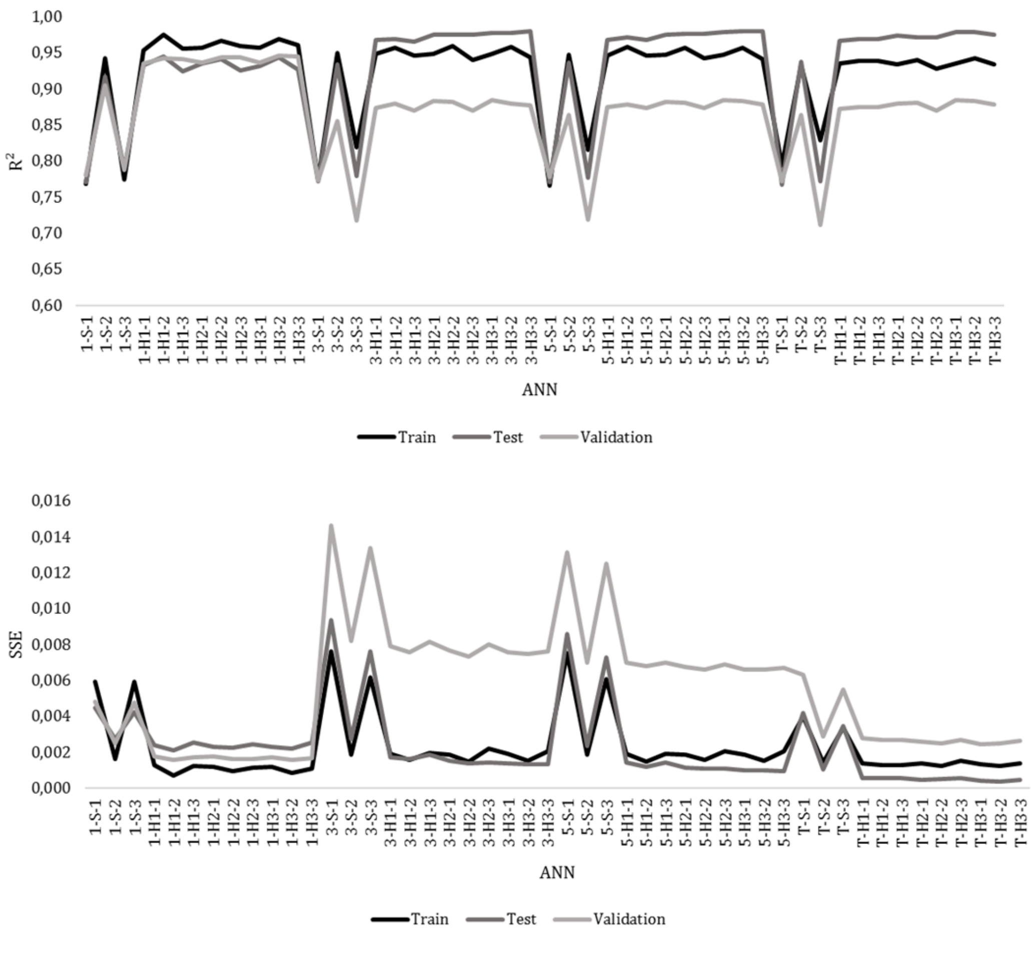

Figure 1. Performance graphs: determination coefficient (R²) and sum of squares of errors (SSE) of the Artificial neural networks (ANN) obtained. In "*ANN X-Y-Z*", X represents the number of individuals in the training sample (T being for all), Y represents the number of dominant trees (S for none) and Z represents which variables are used as input, in addition to the dominant height (1 for Clone and diameter at 1.30 m from the ground (DBH); 2 for all; 3 for DBH, basal area (Gparc) and dominant diameter (DBHd).

Gráficos de rendimiento: coeficiente de determinación (R²) y suma de cuadrados de errores (SSE) de las Redes Neuronales Artificiales (RNA) obtenidas. En "RNA X-Y-Z", X representa el número de individuos en la muestra de entrenamiento (T es para todos), Y representa el número de árboles dominantes (S para ninguno) y Z representa qué variables se usan como entrada, además del dominante altura (1 para clon y diámetro a 1,30 m del suelo (DAP); 2 para todos; 3 para DAP, área basal (Gparc) y diámetro dominante (DAPd).

With ANN, the heights of trees with known height were estimated. As a result, each tree has an observed height and 48 estimated heights. The values of Bias and RMSE calculated to evaluate the performance of the networks are presented in tables 1 and 2, respectively. Bias values close to zero indicate less error tendencies in the estimates. Negative values indicate overestimates and positive values indicate underestimates. RMSE values indicate the average magnitude of the error.

Bias and RMSE indicated different network performances according to the input variables used. When analyzing the different dominant heights considered in the training of the networks, it appears that the highest Bias values were found in networks that did not have a dominant height as an-input variable, with a tendency to overestimate the height values. The networks trained without the Hd variable showed Bias between -11.82 and 3.51 %, while in the networks where Hd was used, there was a smaller variation, from -5.61 to 2.43 %. The different ways of calculating the dominant height studied showed Bias values close to each other. With values between -4.79 and 2.30 % for Hd1; -5.61 and 2.43 % for Hd2; and -5.53 and 1.87 % for Hd3.

Due to larger bias, networks without Hd also presented higher magnitudes of error, which can be verified by the higher values of RMSE. Networks without Hd had an average RMSE of 5.62 %, while in networks with Hd1 the average was 3.73 %, in networks with Hd2, 3.79 % and 3.70 % in networks with Hd3. Regarding the maximum values of RMSE, in networks without Hd, the value of 12.55 % was verified, and in networks with Hd, the maximum value was 7.54 %. Among the different types of Hd, as well as in Bias, maximum values of RMSE were

Table 1. Bias values for all artificial neural networks. Groups G1, G3, G5 and GT represent the number of trees used in the training sample, with T for all. The networks differ as follows: H indicates the number of trees used as dominant (S for none); considering the variables used as input for ANN training, 1 represents Clone and diameter at 1.30 from the ground (DBH), 2 represents all variables, and 3 represents DBH, basal area (Gparc) and dominant diameter (DBHd).

Valores de sesgo para todas las redes neuronales artificiales. Los grupos G1, G3, G5 y GT representan el número de árboles utilizados en la muestra de entrenamiento, con T para todos. Las redes difieren de la siguiente manera: H indica el número de árboles utilizados como dominantes (S para ninguno); considerando las variables utilizadas como entrada para el entrenamiento RNA, 1 representa el clon y el diámetro a 1,30 m desde el suelo (DAP); 2 representa todas las variables; y 3 representa DAP, área basal (Gparc) y diámetro dominante (DAPd).

	Bias (%)											
Group	S-1	S-2	S-3	H1-1	H1-2	H1-3	H2-1	H2-2	H2-3	H3-1	H3-2	H3-3
G1												
MG01	2.55	1.90	-11.82	2.30	1.38	-4.59	2.43	0.67	-3.74	-3.74	1.57	-3.26
MG02	-0.88	0.82	3.40	0.83	1.05	1.35	0.88	1.16	1.24	1.24	1.16	1.38
MG03	-0.02	0.53	1.20	0.82	0.58	1.83	0.66	0.74	1.86	1.86	0.71	1.87
MG04	-0.33	1.79	-4.33	1.68	1.70	1.28	1.64	1.84	1.13	1.13	1.73	1.19
G3												
MG01	-4.24	-1.01	-11.55	-1.51	-0.74	-4.79	-1.59	-1.07	-5.61	-1.27	-1.27	-5.53
MG02	-2.07	-1.24	2.44	-1.38	-1.27	-0.85	-1.39	-1.29	-0.61	-1.42	-1.42	-0.82
MG03	-1.14	-1.05	-0.31	-0.99	-1.01	-0.46	-0.93	-0.85	-0.59	-0.82	-0.82	-0.77
MG04	-2.40	-1.18	-4.90	-1.16	-1.20	-1.62	-1.25	-1.14	-1.26	-1.10	-1.10	-1.16
G5												
MG01	-1.49	0.04	-11.00	-1.09	-0.87	-3.97	-1.00	-0.86	-5.10	-0.60	-0.80	-4.50
MG02	-1.02	-0.76	2.93	-0.88	-0.75	-0.40	-0.83	-0.74	-0.24	-0.92	-0.83	-0.31
MG03	-1.19	-0.56	0.10	-0.58	-0.55	0.04	-0.57	-0.55	-0.34	-0.41	-0.56	-0.38
MG04	-1.61	-0.74	-4.46	-0.76	-0.68	-1.01	-0.79	-0.67	-0.73	-0.71	-0.76	-0.65
GT												
MG01	-0.86	-0.06	-10.08	-0.61	-0.18	-4.42	-0.38	-0.11	-5.17	-0.39	-0.88	-4.32
MG02	-0.89	-0.37	3.51	-0.25	-0.40	0.24	-0.30	-0.33	0.20	-0.31	-0.26	0.19
MG03	-0.56	-0.15	0.76	-0.08	-0.03	0.07	-0.07	0.09	0.26	0.04	-0.18	0.08
MG04	-0.82	-0.30	-3.85	-0.22	-0.17	-0.41	-0.18	-0.22	-0.18	-0.25	-0.21	-0.22

Table 2. Root mean square error (RMSE) for the ANN obtained. G1, G3, G5 and GT represent the number of trees used in the training sample, with T for all. "Y-Z" differentiates networks as follows: Y for the number of trees used as dominant (S for none), Z for the variables used as input (1 for Clone and diameter at 1.30 from the ground (DBH), 2 for all, 3 for DBH, basal area (Gparc) and dominant diameter (DBHd)).

Error cuadrático medio (RMSE) para el RNA obtenido. G1, G3, G5 y GT representan el número de árboles utilizados en la muestra de entrenamiento, con T para todos. "Y-Z" diferencia las redes de la siguiente manera: Y para el número de árboles utilizados como dominantes (S para ninguno); Z para las variables utilizadas como entrada (1 para Clon y diámetro a 1,30 m desde el suelo (DAP); 2 para todos; 3 para DAP, área basal (Gparc) y diámetro dominante (DAPd)).

	RMSE (%)											
Clones	S-1	S-2	S-3	H1-1	H1-2	H1-3	H2-1	H2-2	H2-3	H3-1	H3-2	H3-3
G1												
MG01	9.19	5.26	12.55	6.31	3.76	6.14	6.42	3.69	6.31	6.31	3.64	6.01
MG02	6.19	3.66	5.51	3.48	3.09	3.50	3.36	3.33	3.32	3.32	3.24	3.42
MG03	4.45	3.20	3.72	2.68	2.64	3.23	2.57	2.62	3.21	3.21	2.51	3.19
MG04	6.40	4.77	6.74	3.95	3.71	3.66	3.92	3.84	3.57	3.57	3.87	3.61
G3												
MG01	7.81	3.54	12.50	5.16	3.47	6.40	5.34	3.27	7.54	3.83	3.83	7.27
MG02	6.72	4.34	5.33	4.33	4.24	4.24	4.23	4.15	4.12	4.16	4.16	4.13
MG03	4.67	3.15	3.61	2.42	2.41	2.14	2.29	2.27	2.15	2.29	2.29	2.18
MG04	6.94	3.58	6.69	3.40	3.19	3.66	3.37	3.14	3.60	3.14	3.14	3.30
G5												
MG01	6.67	3.00	11.62	4.90	3.67	5.62	4.98	4.01	7.20	4.72	3.83	6.90
MG02	6.48	3.85	5.32	3.94	3.71	3.85	3.84	3.66	3.74	3.83	3.63	3.69
MG03	4.54	2.81	3.51	2.19	2.16	2.03	2.07	2.12	2.04	2.01	2.09	2.05
MG04	6.73	3.46	6.32	3.21	2.97	3.33	3.20	3.02	3.20	3.22	3.05	3.25
GT												
MG01	6.90	3.22	11.22	5.42	4.24	6.82	5.90	4.49	7.31	5.53	4.01	6.82
MG02	6.37	3.91	5.65	3.99	3.95	3.92	3.90	3.77	3.97	3.83	3.71	3.77
MG03	4.55	2.98	3.68	2.49	2.47	2.53	2.45	2.44	2.59	2.41	2.43	2.46
MG04	6.75	3.85	5.97	3.46	3.45	3.64	3.50	3.45	3.55	3.55	3.47	3.53

found close to each other: 6.82, 7.54 and 7.27 % for Hd1, Hd2 and Hd3 respectively. Indicating that the use of the height of only one dominant tree can generate networks with good estimation capacity.

It shows that, in addition to Hd, the simultaneous absence of the categorical variable Clone in the training process of networks negatively influenced the quality of the estimates. Among the ANN trained without Clone and without Hd as input variables, Bias values varied between -11.82 and 3.51 % and the maximum RMSE value was 12.55 %. There was a tendency to overestimate heights, especially in Clone MG01, where the average height of trees is lower. As networks did not have the information of maximum heights or information that differentiated Clones (Clone), the same pattern verified in the other Clones was applied, in which trees are bigger. In networks where the Clone variable was used and Hd was not used in training, Bias values were between -4.24 and 2.55 % and the maximum RMSE verified was 9.19 %. The networks trained with Clone and Hd presented Bias between -3.74 and 2.43 % and maximum RMSE of 5.26 %.

With the use of continuous variables (DBH, DBHd, Gparc and Hd), it was possible to obtain even lower values of Bias and RMSE. Among the networks that used categorical variables and did not use continuous variables, except DBH, there were Bias values between -3.74 and 2.43 % and maximum RMSE of 6.42 %; while in the networks trained with continuous and categorical variables, Bias values were between -1.42 and 1.84 % and maximum RMSE of 4.49 %.

The ANN with all the trees used as a training sample resulted in estimates with lower values of Bias and RMSE, however, there is no significant difference between the values observed for these networks and those whose training samples are made up of smaller numbers of trees. For networks with a tree as a training sample, Bias values varied between -11.82 and 3.40 % and the maximum RMSE was 12.55 %. For networks with three trees in the training sample, Bias values varied between -11.55 and 2.44 % and the maximum RMSE was 12.50 %. Nets with five trees in the training sample showed Bias values between -11.00 and 2.93 % and maximum RMSE of 11.62 %. For networks with all trees in the training sample, Bias between -10.08 and 3.51 % and maximum RMSE of 11.22 %.

Considering the performance of networks (lowest values of Bias and RMSE), the best five were selected, whose minimum and maximum RMSE and Bias values are shown in table 3.

With the adjustment of the hypsometric model used as a reference in this study, four equations were obtained, one for each Clone. The coefficients of the equations and their determination coefficients (R^2) are shown in table 4. In all adjusted equations, the parameters associated with the coefficients (DBH and Hd) were significant by the t test ($P < 0.05$).

All equations showed lower values than those shown by ANN, when compared by the determination coefficient (R^2). The lowest value found in training the networks was 0.7661 (figure 1), while the highest value found in adjusted hypsometric models was 0.7655 (table 4). This result corroborates with Haykin (2001), who showed that ANN may have a higher estimation capacity than that of the regression models. The parameters and coefficients of determination (R^2) were obtained by adjusting the volumetric model of Schumacher and Hall (1933), by Clone. In all adjusted equations, the parameters associated with the coefficients (DBH and Ht) were significant by the t test ($P < 0.05$).

The volumes per hectare, for each sample plot, estimated by the adjusted models of Schumacher and Hall using, in addition to the DBH, the heights estimated by the five best ANN, indicated differences in the estimates. Considering the hypsometric model cited by Campos and Leite (2009), adjusted by Clone, as a reference for the height estimation and comparing the estimates of this with the height estimates by ANN, it can be observed, in general, a trend of Mean Relative Error (MRE %) less than 10 % (figure 2). In the ANN where Hd was used as an input variable in training, there is less dispersion of percentage errors around zero, indicating higher precision of the estimates. In the ANN where Hd was not used, despite having a low

Table 3. Artificial neural network (ANN) with the best performances and their minimum and maximum Bias values and average of Root mean square error (RMSE). G3 and G5 represent the number of trees used in the training sample. "Hd" differentiates networks according to the number of trees used as dominant height (S for none); 2 represents the variables used as input (Clone, diameter at 1.30 m from the ground (DBH), basal area (Gparc) and dominant diameter (DBHd)).

Red neuronal artificial (RNA) con los mejores rendimientos y sus valores de sesgo mínimo y máximo y el promedio del error cuadrático medio (RMSE). G3 y G5 representan el número de árboles utilizados en la muestra de entrenamiento. "Hd" diferencia las redes de acuerdo con el número de árboles utilizados como altura dominante (S para ninguno); 2 representa las variables utilizadas como entrada (Clon, diámetro a 1,30 m desde el suelo (DAP), área basal (Gparc) y diámetro dominante (DAPd)).

ANN	RMSE	Minimum bias (%)	Maximum bias (%)
G5-Hd1-2	3.13	-1.09	-0.58
G5-Hd3-2	3.15	-0.83	-0.56
G5-Hd2-2	3.20	-0.86	-0.55
G5-Hd2-2	3.21	-1.29	-0.85
G5-S-2	3.28	-0.37	-0.06

Table 4. Estimates of adjusted parameters (βi) for the hypsometric model, by Clone, and their respective determination coefficients (R^2). All parameters were significant ($P < 0.05$).

Estimaciones de los parámetros ajustados (βi) para el modelo hipsométrico, por clon, y sus respectivos coeficientes de determinación (R^2). Todos los parámetros fueron significativos ($P < 0,05$).

Clone	β0	β1	β2	R^2
MG01	-13.76672	-59.31327	12.22367	0.6519
MG02	-16.70950	-92.29903	14.63143	0.7038
MG03	-17.45423	-98.43312	14.90976	0.7655
MG04	-0.85420	-113.46220	9.81136	0.6539

variation in the values of Bias and RMSE, there is a high dispersion of percentage errors (figure 2).

It is noted that the MRE trend was closer to zero in the networks trained with five trees in the training sample. However, this superiority is not so significant as to compromise the use of the network trained with three trees in the training sample, since the MRE values in this network were, in general, below 5 %.

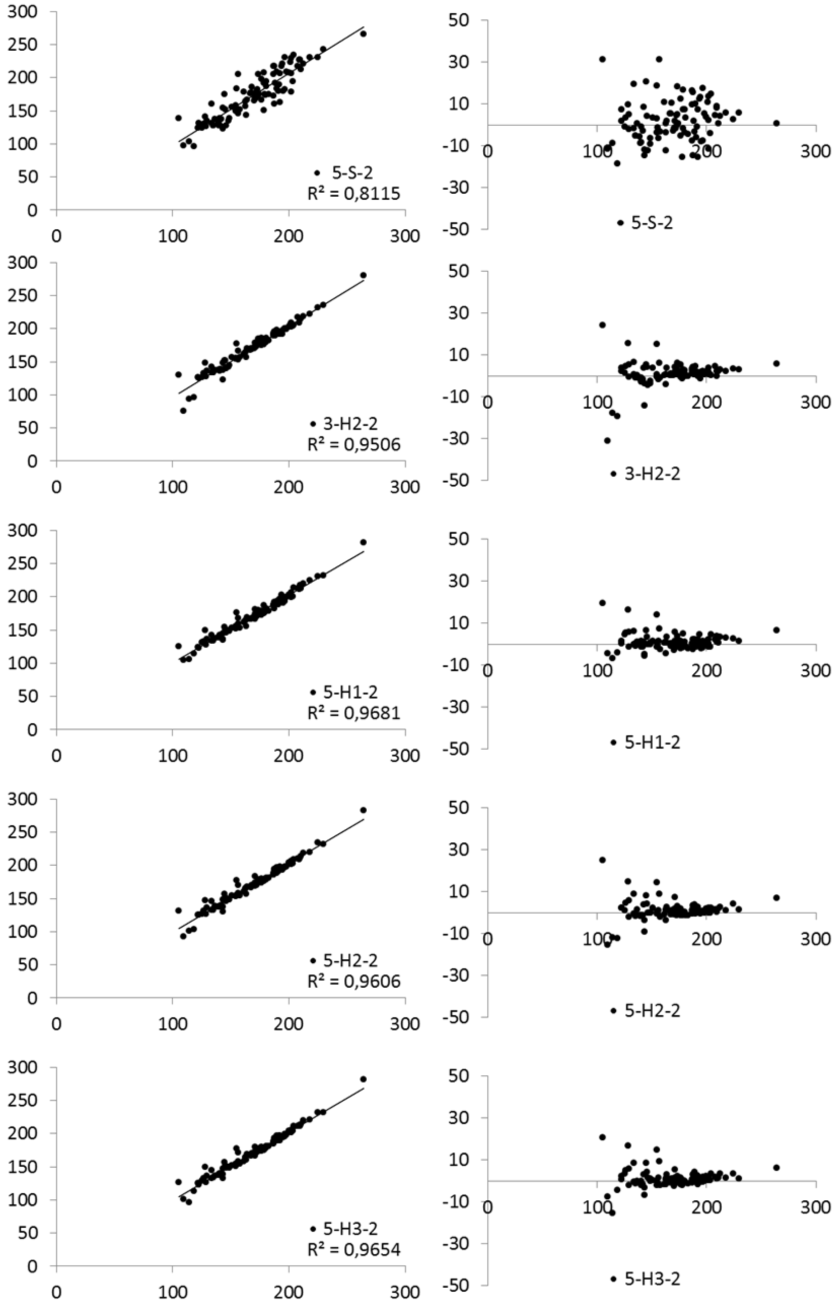

Figure 2. Observed (x) and estimated (y) volumes per hectare and their correlation coefficients (R²). Dispersion of percentage errors (y) as a function of total observed volumes (x) per hectare. Groups 3 and 5 represent the number of trees used in the training sample. "H" differentiates networks according to the number of trees used as dominant (S for none); 2 represents the variables used as input (Clone, diameter at 1.30 m from the ground (DBH), basal area (Gparc) and dominant diameter (DBHd)).

Volúmenes observados (x) y estimados (y) por hectárea y sus coeficientes de correlación (R²). Dispersión de errores porcentuales (y) en función de los volúmenes totales observados (x) por hectárea. Los grupos 3 y 5 representan el número de árboles utilizados en la muestra de entrenamiento. "H" diferencia las redes de acuerdo con el número de árboles utilizados como dominantes (S para ninguno); 2 representa las variables utilizadas como entrada (clon, diámetro a 1,30 m desde el suelo (DAP), área basal (Gparc) y diámetro dominante (DAPd)).

Considering the estimated volumes per hectare, for each plot, it is confirmed that Hd contributed significantly to obtain more accurate estimates (table 5).

In the 5-S-2 network, nine MRE values above 10 % were verified, in a total of 28 estimates, and the maximum MRE verified was 26 %. In 3-H2-2, 5-H2-2 and 5-H3-3 networks, only a value above 10 % was verified, for the same number of estimates. The highest MRE values are 12 % for the 3-H2-2 network, 13 % for the 5-H2-2 network, and 11 % for the 5-H3-3 network. The 5-H1-2 network did not present a MRE value above 10 %, with 7 % being the maximum value. The lowest average of the MRE modules was also verified for this network (table 6).

DISCUSSION

Results show that the use of variables, both categorical and continuous, that manage to represent the characteris-

Table 5. Estimated volumes (Vol, m³ ha^{-1}) considering the heights estimated by the hypsometric model used as a reference, and by the five best Artificial neural network, as well as their mean relative error (MRE).

Volúmenes estimados (m³ ha^{-1}) considerando las alturas estimadas por el modelo hipsométrico utilizado como referencia, y por las cinco mejores redes neuronales artificiales, así como su error relativo medio (ERM).

Management unit	Model	3-H2-2		5-S-2		5-H1-2		5-H2-2		5-H3-2	
	Vol	Vol	MRE	Vol	MRE	Vol	MRE	Vol	MRE	Vol	MRE
1	126.9	132.0	4.0	130.6	2.9	130.2	2.6	130.6	2.9	130.0	2.4
2	204.1	208.9	2.3	217.4	6.5	206.9	1.4	205.9	0.9	206.7	1.3
3	166.8	176.5	5.8	192.4	15.3	171.0	2.5	170.8	2.4	170.0	1.9
4	233.1	244.1	4.7	236.4	1.4	243.9	4.6	246.2	5.6	244.5	4.9
5	187.7	192.0	2.3	211.4	12.6	191.2	1.9	189.9	1.2	191.6	2.1
6	176.7	176.5	-0.1	149.4	-15.5	175.5	-0.7	178.1	0.8	173.7	-1.8
7	106.6	93.6	-12.2	95.9	-10.1	100.6	-5.6	92.1	-13.4	94.3	-11.5
8	165.9	167.9	1.2	167.9	1.2	167.9	1.2	168.1	1.3	168.4	1.5
9	140.9	139.0	-1.4	132.9	-5.7	141.6	0.5	140.3	-0.5	140.9	0.0
10	159.0	164.7	3.6	171.4	7.8	163.9	3.1	164.2	3.3	164.5	3.5
11	177.5	178.5	0.6	168.2	-5.3	174.8	-1.6	177.5	0.0	176.1	-0.8
12	159.8	170.7	6.8	163.3	2.2	171.8	7.5	173.2	8.4	168.9	5.7
13	129.4	124.1	-4.1	119.5	-7.7	131.0	1.2	126.9	-2.0	131.8	1.8
14	133.7	133.3	-0.3	130.6	-2.3	134.9	0.9	134.6	0.7	134.2	0.4
15	169.5	170.5	0.6	192.0	13.3	175.3	3.4	168.4	-0.7	168.9	-0.4
16	192.5	191.1	-0.8	213.7	11.0	194.2	0.9	192.0	-0.3	194.1	0.8
17	167.6	172.6	2.9	178.0	6.2	167.9	0.2	168.3	0.4	168.0	0.2
18	175.6	180.9	3.0	187.7	6.9	176.2	0.3	175.2	-0.2	175.8	0.1
19	187.0	188.7	0.9	192.3	2.8	188.7	0.9	187.0	0.0	186.9	-0.1
20	153.4	153.8	0.3	154.0	0.4	152.3	-0.7	152.8	-0.4	151.8	-1.0
21	143.7	151.9	5.7	181.3	26.2	153.7	7.0	155.9	8.5	156.4	8.8
22	167.3	168.6	0.8	162.6	-2.8	164.0	-1.9	166.1	-0.7	164.7	-1.5
23	188.9	189.3	0.2	209.9	11.1	195.3	3.4	191.8	1.5	194.3	2.9
24	185.2	188.0	1.5	180.6	-2.5	185.2	0.0	187.0	1.0	184.5	-0.4
25	138.8	143.2	3.2	147.4	6.2	141.8	2.2	143.5	3.4	143.3	3.2
26	132.2	130.0	-1.7	122.8	-7.1	132.0	-0.2	130.9	-1.0	131.9	-0.3
27	183.5	187.1	2.0	167.1	-9.0	180.0	-2.0	185.4	1.0	182.6	-0.5
28	121.7	129.5	6.3	134.4	10.4	126.7	4.1	129.3	6.2	127.3	4.6

Table 6. Average, minimum and maximum mean relative error (MRE) values generated by the 5 Artificial neural networks, per management unit. G3 and G5 represent the number of trees used in the training sample. "Hd" differentiates networks by the number of trees used as dominant (S for none); 2 represents the variables used as input (Clone, diameter at 1.30 from the ground (DBH), basal area (Gparc) and dominant diameter (DBHd)).

Valores promedio, mínimo y máximo de error relativo medio (ERM) generados por las cinco Redes Neuronales Artificiales, por unidad de manejo. G3 y G5 representan el número de árboles utilizados en la muestra de entrenamiento. "Hd" diferencia las redes por el número de árboles utilizados como dominantes (S para ninguno); 2 representa las variables utilizadas como entrada (clon, diámetro a 1,30 m desde el suelo (DAP), área basal (Gparc) y diámetro dominante (DAPd)).

ANN	Minimum MRE (%)	Average MRE (%)	Maximum MRE (%)
G5 – Hd1 - 2	0.00	2.23	7.00
G5 – Hd3 – 2	0.00	2.29	11.00
G5 – Hd2 – 2	0.01	2.46	13.00
G3 – Hd2 – 2	0.13	3.26	12.00
G5 – S – 2	0.38	7.58	26.00

tics of the plots, especially the Clone variable, is important in the training of ANN to obtain estimates with better accuracy, since these variables provide information about the specificities of each Clone, field or project, reducing, for example, the generalization of characteristics observed in a given Clone to others with different behaviors. It is worth mentioning that in the data used, considering the categorical variables, only Clone information was available, the introduction of additional information, such as soil type, terrain preparation, precipitation, spatial arrangement, radiation, among others, can contribute to increase the quality of the estimates.

The use of the Hd variable contributed to the improvement of the estimates and the use of the height of the largest tree in the plot resulted in ANN with performances similar to those presented by the networks trained with Hd coming from the average height of more than one dominant tree in the plot.

The reduction in the number of trees used as a training sample did not significantly affect the performance of the networks. The use, for example, of five trees as a training sample can already provide a considerable gain of time and cost reduction in the forest inventory and the difference between the maximum RMSE of networks trained with all trees and networks trained with five trees was only 0.40 %.

In the case of the forest company, from which the data used were obtained, the number of Ht measured per plot, which is 25 (20 normal trees and five dominant trees), could be reduced to eight (seven normal trees, where five would be used in the training sample, one in the test and one in the validation; and one dominant tree). Enabling a reduction in the measurement time and, consequently, in the cost of the forest inventory, increasing the efficiency of the measurement team.

Binoti *et al*. (2013), in a study on the effect of reducing Ht measurements on the precision obtained by ANN, evaluated the estimates obtained by reducing the number of plots with measured Ht and also concluded that it is possible to reduce the number of measurements without loss of accuracy. Still according to the authors, it is possible to reduce the cost of the forest inventory through the application of ANN in the estimation of the Ht of the trees.

According to Leite and Andrade (2003), the dominant height allows representing different productive capacities of the places where the plots are located. This is important since the relationship between total height and DBH of trees can differ among plots located in areas with lower, medium or higher productivity.

The networks with the highest precision were those with training samples composed of five trees per plot (however, the number of trees can be reduced to three without major losses in accuracy); use of the dominant height variable, regardless of how many trees are used in its calculation (1, 2 or 3); and categorical and continuous variables that differentiate the different extracts, such as the Clone, Gparc and DBHd variables.

More specifically, the best performance was presented by the 5-H1-2 network, in whose training five trees were considered as a training sample, one as a test sample and one as a validation sample; dominant height from the height of the highest tree in the plot, categorical variable Clone and continuous variables DBH, DBHd and Gparc (figure 3).

From the artificial neural network an equation system was extracted to predict the individual tree height of *Eucalyptus* spp., with coefficients resulting from the weights generated by the ANN. Model (4) expresses the relationship between the hidden layer and the response variable, where β_0 is the bias, and the other coefficients are the weights related to each neuron. Model (5) represents the activation function used in each neuron of the hidden layer, derived from the logistic model. Finally, the model (5) is the result of the relationship between the input variables and the respective hidden layer neurons, generating a model for each neuron.

$$Ht' = \beta_0 + \beta_1 * z_1 + \beta_2 * z_2 + \beta_3 * z_3 + \beta_4 * z_4 + \beta_5 * z_5 \quad [4]$$

$$z_n = \left[\frac{1}{1 + e^{-w_i}}\right] \quad [5]$$

$$w_i = \beta_{0.n} + \beta_{1.n} * DBH_i' + \beta_{2.n} * \ DBHd_i' + \beta_{3.n} * \ Gparc_i' + \beta_{4.n} * \ Clone_i + \beta_{5.n} * \ Hd_i' \quad [6]$$

Where β_0: bias, β_n: coefficient of the model associated with neuron n, $\beta_{k.n}$: coefficient of the model between input variable k and neuron n, z_n: response of the n-th neuron of the hidden layer, w_i: sum of the products between the weights and the inputs.

The coefficients of the system of equations extracted from the artificial neural network are presented in table 7.

It can be inferred, therefore, that the ANN performed satisfactorily in estimating the total height of the trees studied, for later obtaining the individual volumes and per unit area. Therefore, this tool is applicable to the processes of estimating the total height of eucalyptus trees, allowing the reduction of the number of measurements required per plot without significant interference in the accuracy of the estimates obtained.

Another important aspect to be considered, due to the ease provided to the modeler, is that, unlike regression models, adjustments by extract are not necessary, since a single ANN is representative for all extracts (Haykin 2001).

Diamatopoulou (2005) reports that the quality of the estimates obtained through the ANN is due to their ability to model several variables and overcome certain problems found in forest data, such as non-linear relationships, non-Gaussian distributions, outliers and data failures.

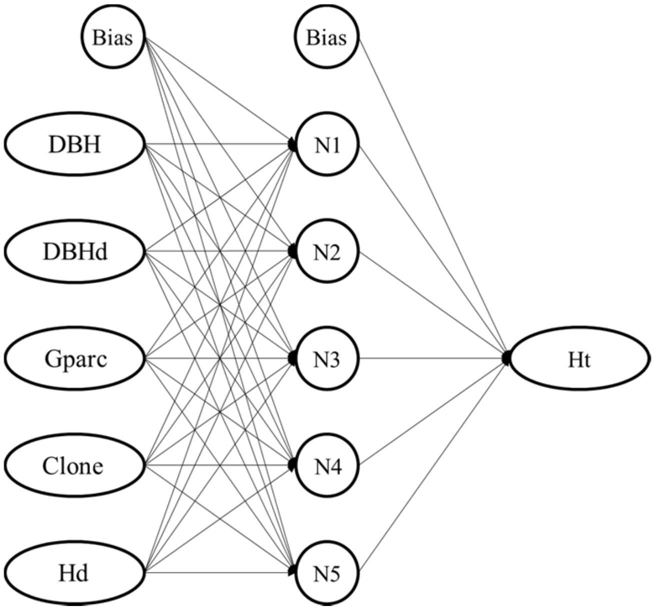

Figure 3. Architecture of the best ANN, with five neurons in the hidden layer

Arquitectura de la mejor RNA, con cinco neuronas en la capa oculta.

Table 7. Parameters (β's) of the artificial neural network. N represents the neuron.

Parámetros (βi') de la red neuronal artificial. N representa la neurona.

	β0	β1	β2	β3	β4	β5
RNA	0.8609	-0.7323	0.7619	0.6785	-1.7671	-1.6967
N1	-2.3277	2.0301	0.6158	2.5005	2.8092	-0.3280
N2	-1.3559	0.6158	2.3549	0.9111	0.6736	-2.8348
N3	-1.7104	2.3549	0.1538	0.5098	-1.6662	1.0934
N4	-0.0674	-2.5130	-1.2125	0.4046	-1.1707	0.2400
N5	2.5005	2.8092	-2.5130	0.6372	0.7235	-0.2786

CONCLUSIONS

The present study considerably improves the modeling of the height and log volume of *Eucalyptus* spp. trees, using machine learning. The technique performed satisfactorily, and the models based on Artificial neural networks proposed in this study to estimate the total height of eucalyptus trees are efficient and their application is recommended due to the expressive reduction of the number of tree heights to be measured in the field.

The model that presents the best performance, according to the data used, consists of five trees as a training sample, one as a test sample and one as a validation sample; dominant height from the height of the highest tree in the plot; categorical variable Clone and continuous variables: diameter at 1.30 m in height from the base of the tree, dominant diameter and basal area of the plot.

ACKNOWLEDGMENTS

The authors thank the Postgraduate Program in Forestry Engineering at the Federal University of Lavras (PPGEF-UFLA) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for the financial support in carrying out this study.

REFERENCES

Alvares CA, JL Stape, PC Sentelhas, G Moraes, J Leonardo, G Sparovek. 2013. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift* 22:711-728. DOI: https://dx.doi.org/10.1127/0941-2948/2013/0507

Binoti DHB, MLMS Binoti, HG Leite. 2013. Redução dos custos em inventário de povoamentos equiâneos. *Revista Brasileira de Ciências Agrárias* 8:125-129. DOI: https://dx.doi.org/10.5039/agraria.v8i1a2209

Binoti MLMS, DHB Binoti, HG Leite, SLR Garcia, MZ Ferreira, R Rode, AAL Silva. 2014. Redes neurais artificiais para estimação do volume de árvores. *Revista Árvore* 38:283-288. DOI: http://dx.doi.org/10.1590/S0100-67622014000200008

Binoti MLMS. 2012. Emprego de Redes Neurais Artificiais em Mensuração e Manejo Florestal. Tese (Doutorado em Engenharia Florestal). Viçosa –Minas Gerais, Brasil. Universidade Federal de Viçosa. 130 p.

Braga AP, APLF Carvalho, TB Ludemir. 2007. Redes Neurais Artificiais: Teoria e Aplicações. Rio de Janeiro, Brasil. Editora LTC. 248 p.

Campos JCC, HG Leite. 2009. Mensuração florestal: perguntas e respostas. 3. ed. Viçosa, Brasil. UFV. 636 p.

Curtis R. 1967. Height-diameter and height-diameter-age equations for second-growth Douglas-fir. *Forest Science* 13:365-375. DOI: https://doi.org/10.1093/forestscience/13.4.365

Dantas D, N Calegario, FWA Júnior, SPC Carvalho, MAI Júnior, EA Melo. 2020. Multilevel nonlinear mixed-effects model and machine learning for predicting the volume of *Eucalyptus* spp. trees. *CERNE* 26(1): 48-57. DOI: 10.1590/01047760202026012668

Diamantopoulou MJ. 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. *Computers and Electronics in Agriculture* 10:235-244. DOI: https://doi.org/10.1016/j.compag.2005.04.002

Freitas CS, HN Paiva, JCL Neves, GE Marcatti, HG Leite. 2020. Modeling of eucalyptus productivity with artificial neural networks. *Industrial Crops and Products* 164:112149. DOI: https://doi.org/10.1016/j.indcrop.2020.112149

Gorgens EB, A Montaghi, LCE Rodriguez. 2015. 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* 116:221-227. DOI: https://doi.org/10.1016/j.compag.2015.07.004

Günther F, S Fritsch. 2010. Neuralnet: Training of neural networks. *The R journal*. 2(1):30-38. Accessed in sep. 2018. Available in https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf

Haykin S. 2001. Redes neurais: princípios e prática. 2. ed. Porto Alegre, Brasil. Bookman. 898 p.

Inoue A, S Yoshida. 2004. Allometric model of the height–diameter curve for even-aged pure stands of Japanese cedar (*Cryptomeria japonica*). *Journal of Forest Research* 9:325–331. DOI: https://doi.org/10.1007/s10310-004-0085-z

Ker J, J Smith. 1957. Sampling for height-diameter relationships. *Journal of Forestry* 55:205-207. DOI: https://doi.org/10.1093/jof/55.3.205

Leite HG, VCL Andrade. 2003. Importância das variáveis altura dominante e altura total em equações hipsométricas e volumétricas. *Revista Árvore* 27:301-310. DOI: http://dx.doi.org/10.1590/S0100-67622003000300005

Leite HG, VCL Andrade. 2002 Um método para condução de inventários florestais sem o uso de equações volumétricas. *Revista Árvore* 26:321-328. DOI: http://dx.doi.org/10.1590/S0100-67622002000300007

Schumacher FX, FS Hall. 1933. Logarithmic expression of timber-tree volume. *Journal of Agricultural Research* 47:719-734.

Siipilehto J. 2000. A comparison of two parameter prediction methods for stand structure in Finland. *Silva Fennica* 34:331-349. DOI: http://dx.doi.org/10.14214/sf.617

Statsoft. 2014. Statistica (data analysis software system). Version 10. Accessed in May. 2020. Available in www.statsoft.com.br

Recibido: 17/05/20
Aceptado: 17/09/20