

Machine learning for carbon stock prediction in a tropical forest in Southeastern Brazil

Aprendizaje de máquina para la predicción de reservas de carbono
en un bosque tropical en el sureste de Brasil

Daniel Dantas ^{**}, Marcela de Castro Nunes Santos Terra ^a,
Luis Paulo Baldissera Schorr ^a, Natalino Calegario ^a

* Corresponding autor: ^aFederal University of Lavras, Department of Forest Sciences,
Lavras, Minas Gerais, Brazil, tel.: 5538991237493, dantasdaniel12@yahoo.com.br

SUMMARY

The increasing awareness of global climate change has drawn attention to the role of forests as mitigators of this process as they act as carbon sinks to the atmosphere. Understanding the process of carbon storage in forests and its drivers, as well as presenting consistent models for their estimation, is a current demand. In this sense, the aim of this study was to evaluate the performance of machine learning techniques: support vector machines (SVM) and to propose a new nonlinear model extracted from the training of an artificial neural network (ANN) in the modeling of above ground carbon stock in a secondary semideciduous forest. SVM and ANN construction and training process considered independent variables selected by stepwise: minimum DBH (diameter of breast height - 1.3 m), maximum DBH, mean DBH, total height and number of trees, all by plot. SVM and the model extracted from ANN were applied to the data set intended for validation. Both techniques presented satisfactory performance in modeling carbon stock by plot, with homogeneous distribution and low dispersion of residues and predicted values close to those observed. Analysis criteria indicated superior performance of the model extracted from the artificial neural network, which presented a mean relative error of 6.94 %, while the support vector machine presented 13.52 %, combined with lower bias values and higher correlation between predictions and observations.

Key words: artificial intelligence, artificial neural networks, support vector machines, forest biomass.

RESUMEN

La conciencia de la sociedad en relación a los cambios climáticos globales ha llamado la atención sobre el papel de los bosques como mitigadores de este proceso, ya que actúan como sumideros de carbono en la atmósfera. Comprender el proceso de almacenamiento de carbono en los bosques y sus determinantes, así como presentar modelos consistentes para su estimación es una demanda actual. En este sentido, el objetivo de este estudio fue evaluar el desempeño de las técnicas de máquina de vectores de soporte (SVM) y proponer un nuevo modelo no lineal extraído del entrenamiento de una red neuronal artificial (RNA) para modelar la cantidad de carbono sobre el suelo en un Bosque secundario estacional semideciduo. El proceso de construcción y entrenamiento de SVM y RNA consideró variables independientes seleccionadas por *stepwise*: DAP mínimo (diámetro de altura del pecho - 1.3 m), DAP máximo, DAP promedio, altura total promedio y número de árboles, todo por unidad de muestreo. La SVM y el modelo extraído de la RNA se aplicaron al conjunto de datos para su validación. Ambas técnicas mostraron un desempeño satisfactorio en la modelación de la cantidad de carbono por unidad de muestreo, con distribución homogénea y baja dispersión de residuos y valores pronosticados cercanos a los observados. Los criterios de análisis utilizados indicaron un desempeño superior del modelo extraído de la red neuronal artificial, que presentó un error relativo promedio de 6.94 %, mientras que la máquina de vectores de soporte presentó 13.52 % junto con valores de sesgo más bajos y una mayor correlación entre predicciones. y observaciones.

Palabras clave: inteligencia artificial, redes neuronales artificiales, máquinas de vectores de soporte, biomasa forestal.

INTRODUCTION

Forests provide numerous ecosystem services, such as regulation of biogeochemical cycles, pollution control and food supply. Among the most acclaimed ecosystem services provided by forests are the atmospheric carbon (CO₂)

sequestration and its storage (Canadell and Raupach 2008, Chazdon *et al.* 2016). This service is of strategic importance in mitigating ongoing climate change because it acts directly in controlling global warming (Bonan 2008).

In this context, the quantification of the carbon stock present in the most varied types of forests constitutes

an important tool for monitoring this ecosystem service (Scolforo *et al.* 2015). Many initiatives have been taken to quantify carbon stocks in forests, both by direct means (Dantas *et al.* 2021) and through estimates from related data (indirect methods) (Cordeiro *et al.* 2018).

Carbon stock estimation by indirect methods employs modeling and simulation techniques. Historically, modeling of forest attributes has relied on approaches based on statistical models (*e.g.* Melo *et al.* 2017). These approaches share space today with computational approaches of artificial intelligence/machine learning, such as artificial neural networks, support vector machines, decision trees, among others, which have been gaining space as tools for forest data analyses, modeling, estimation and production prognosis. These tools have provided gains in the quality of estimates and predictions (Vendruscolo *et al.* 2015).

Artificial Neural Network (ANN) is a processor consisting of simple processing units (artificial neurons), based on neurons found in the human brain, that calculate certain functions. These units are layered and connected to each other by weights that store experimental knowledge and weight the inputs of each unit. Thus, the acquired knowledge becomes available for use (Braga *et al.* 2007, Dantas *et al.* 2020a).

The most notable features in ANNs are their ability to learn and to generalize information. In other words, ANNs are able, through a learned example, to generalize assimilated knowledge to an unknown data set. Another interesting feature of ANN is the ability to extract non-explicit features from a set of information provided as examples (Haykin 2001).

Support vector machines (SVM) have also proven to be an interesting alternative for mathematical modeling of complex systems (Heddam and Kisi 2018). They are simple techniques in their conceptual basis and capable of solving extremely complex real problems. SVM is a supervised learning technique that is trained to classify different categories of data from various disciplines (Haykin 2001). These have been used for two-class classification problems and are applicable on both linear and non-linear data classification tasks. SVM creates a hyperplane or multiple hyperplanes in a high-dimensional space, and the best hyperplane in them is the one that optimally divides data into different classes with the largest separation among the classes (Steinwart and Christmann 2008).

Initially, SVM techniques were successfully applied as a data classification methodology (Tong and Koller 2001). They were later extended to regression tasks through the following approaches: support vector regression (SVR) and least-square support vector machines (LS-SVM) (Cherkassky and Mulier 1998, Dantas 2020b).

Compared to ANN, SVM has the advantage of leading to an exact solution, that is, a global optimum (Haykin 2001). However, finding a final SVM model may present computational complexity because it requires solving a quadratic programming model and solving a set of nonlinear equations.

The present study aims at evaluating the performance of the support vector machine and artificial neural network techniques, and at proposing a new nonlinear model to the modeling of above ground biomass (carbon stock), using dendrometric variables as inputs, in a secondary semideciduous seasonal forest. It is proposed, as a hypothesis, that (a) machine learning techniques are suitable in modeling above ground biomass, (b) it is possible to extract accurate above ground biomass equations from the artificial neural network training process.

METHODS

Study area and data collection. The study area corresponds to a secondary semideciduous seasonal forest, located in Lavras, Minas Gerais, Brazil, under the coordinates 21° 14' S and 45° 00' W, with average altitude of 900 m (figure 1). The climate is classified as Köppen's Cwb, with dry winters and mild summers (Alvares *et al.* 2013). Mean annual rainfall is around 1,500 mm and mean annual temperature is 19.4 °C (Marques *et al.* 2019). The forest is heterogeneous and presents dominance of tree species of the genus *Anadenanthera*, popularly known as "angico". Data come from 105 sample plots (10x10 m) launched in the area. In each plot, all trees with diameter at breast height (DBH - 1.3 m from the ground) higher than or equal to 5 cm and their respective heights were measured.

From the data collected in the field, the following variables were obtained by plot: minimum DBH (DBHmin), mean DBH (DBHmed), maximum DBH (DBHmax), minimum total Height (Hmin), mean total Height (Hmed), maximum total Height (Hmax), Mean Square Diameter (Dq) and Number of Trees (N).

Above Ground Biomass (AGB) was estimated by tree individual according to the equation proposed by Chave *et al.* (2014), using DBH and total tree height and average basic wood density of 0.620 g cm⁻³. The estimate was performed using the software R (R Core Team 2018), using the BIOMASS package (Réjou-Méchain *et al.* 2017). The estimate of AGB was converted to carbon stock in Mg ha⁻¹, according to Thomas and Martin (2012), a methodology consisting in multiplying AGB by 0.471, which according to the authors corresponds to carbon concentration in tropical forest angiosperms tissues.

Independent variables selection. First, a selection of independent variables was performed by the stepwise method, based on the Akaike Information Criterion (AIC). Thus, the combination of variables that make up the model with the lowest AIC is considered the best. subsequently, the selected variables were used as inputs to model carbon stock by plot through machine learning techniques.

Machine learning algorithms. For carbon stock modeling, support vector machines (SVM) and artificial neural networks (ANNs) were used. The SVM construction was ba-

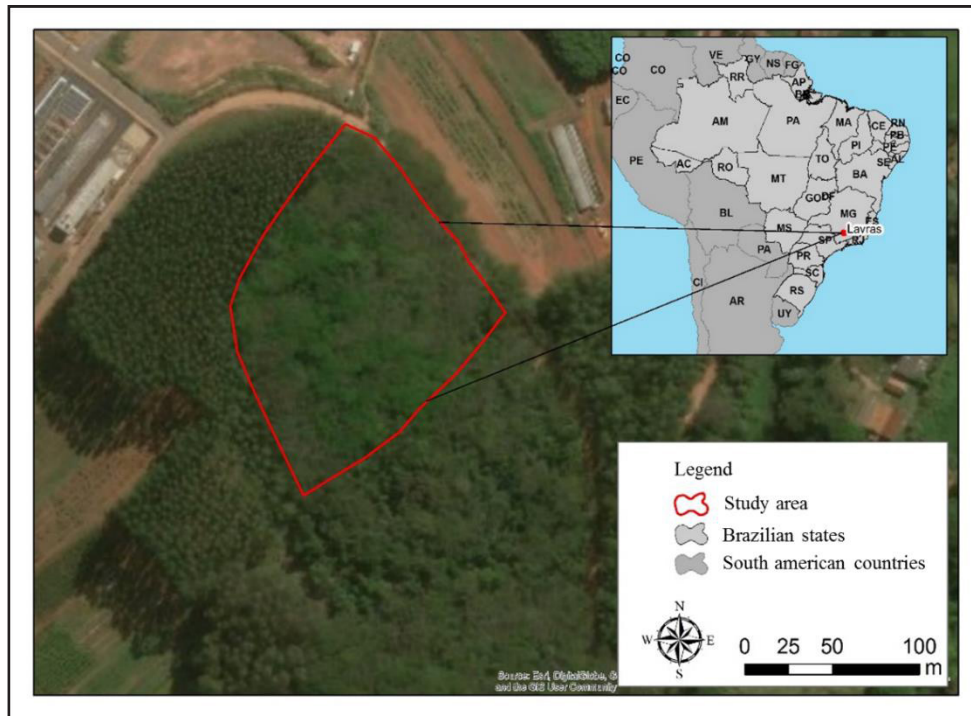


Figure 1. Study area. Secondary semideciduous forest path located in Lavras, Minas Gerais, Brazil. (Adapted from Terra *et al.* 2019).
Área de estudio. Bosque semideciduo estacional en Lavras, Minas Gerais, Brasil (Adaptado de Terra *et al.* 2019).

sed on the supervised machine learning process described by Haykin (2001) and Steinwart and Christmann (2008), where there is a set of paired-order n samples (\mathbf{X}, \mathbf{Y}) , where \mathbf{X} is a matrix of explanatory variables of the sample and \mathbf{Y} is the expected value vector of the sample. Based on this information, taking as input a vector of variables, a chosen function predicts the expected value of the sample. A linear function is given by the form $f(\mathbf{X}) = \langle \mathbf{W}, \mathbf{X} \rangle + b$, where \mathbf{W} is a weight vector.

The type IV error function, also known as *eps-regression*, was used, being the RBF (Radial Basis Function) type Kernel function. Kernel functions offer an alternative solution by designing data in a space with large characteristics to increase the computational power of machine learning, making it possible to represent nonlinear phenomena (Cristianini and Shawe-Taylor 2000). This procedure was performed in software R, version 3.4.1, through the *e1071* package (Meyer *et al.* 2019).

Trained ANNs were Multilayer Perceptron (MLP), composed of an input layer, an intermediate layer and an output layer. The algorithm used was the resilient backpropagation, in which the learning rate was automatically defined by the *neuralnet* package, with values ranging from 0.01 to 1.12.

The choice of the number of neurons in the hidden layer was made using k -fold. This methodology randomly subdivides the database into k subgroups (Ali and Pazzani 1996, Cigizoglu and Kisi 2006). The value of k was

10 subgroups, with a proportion of 90 % for training and 10 % for testing (Diamantopolou 2005), applying cross-validation. Different numbers of neurons, ranging from 1 to 20, were tested.

Logistics (or sigmoidal) was the activation function used, with a range from 0 to 1, which implies limiting the amplitude of the outputs and inputs. Consequently, data were normalized, which consists of transforming the values of each variable to values between 0 and 1. Linear standardization was obtained through equation [1] (Soares *et al.* 2011) and considers the minimum and maximum value of each variable in the transformation of values, maintaining the original distribution of data (Valença 2010).

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

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 standardization range, b : upper limit of the standardization range.

The stopping criterion for the ANN training process was the maximum number of 100.000 cycles, or the mean square error of less than 1 %, and training was terminated when one of these criteria was met. At the end of training, the best ANN was selected based on the lowest mean square error.

A nonlinear equation for tree biomass prediction was extracted from the artificial neural network. Consequently, we generated a system of equations with coefficients resulting from weights generated by the neurons of ANN. This system was used to predict the carbon stock of the plots that comprised the validation database.

Data were divided into two groups: 70 % for ANN training and SVM construction and 30 % for validation of both techniques. Among the data intended for ANN training, 70 % were used in the training phase and 30 % in the test phase.

SVM and ANN performance evaluation. SVM and ANN performances were evaluated in the training and validation phases. Accordingly, the techniques were used to predict the carbon stock in the data set intended for validation, *i.e.* data that had not been used in training. The prediction quality analysis was performed using Mean Relative Error (MRE %) (equation 2), Bias (equation 3), Root Mean Square Error (RMSE %) (equation 4) (Leite and Andrade 2002, Siipilehto 2000), graphs of residuals distribution, graphs of estimated versus observed carbon stocks and the correlation coefficients between estimated and observed values.

$$MRE (\%) = \frac{(\bar{Y}_i - Y_i)}{Y_i} \times 100 \quad [2]$$

$$Bias (\%) = 100 * \frac{1}{n} * \sum_{i=1}^n \left[\frac{(\hat{Y}_i - Y_i)}{Y_i} \right] \quad [3]$$

$$RMSE (\%) = \left[\sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n}} / \bar{Y} \right] * 100 \quad [4]$$

where: Y_i represents the observed value, \hat{Y}_i the estimated value, n the number of observations and \bar{Y} the average of the observed values.

RESULTS

The forest with predominance of *Anadenanthera sp.* contained an average of tree carbon stock (AGC) of 94.25 Mg ha⁻¹. Descriptive statistics of the variables used are presented in table 1.

Stepwise method indicated through the Akaike information criterion that the variables minimum DBH, maximum DBH, average DBH, average H and N are those that have a stronger influence on carbon stock variability, and were, therefore, selected for modeling. It should be noted that DBHmed and Hmed represent the main tree growth trends in each plot; DBHmin and DBHmax, the lower and upper limits of diameter growth, respectively; and N represents the density of individuals in each plot. Figure 2 presents the scatter plots between the variables and their respective confidence intervals and the distribution of each variable.

The configurations obtained with the construction of the support vector machine, which resulted in a machine with 44 support vectors, are presented in table 2.

Regarding the approach by artificial neural networks, figure 3 illustrates the architecture and weights obtained

Table 1. Descriptive statistics of dendrometric variables of a secondary seasonal semideciduous forest located in Lavras, Minas Gerais, Brazil.

Variables dendrométricas de un bosque semideciduo estacional secundario ubicado en Lavras, Minas Gerais, Brasil.

Variable	Mean	Minimum	Maximum	Standard deviation	CV(%)
C	94.25	15.65	245.09	52.57	71.14
DBHmin	5.84	5.00	9.07	0.90	15.34
DBHmax	29.61	12.50	55.25	8.20	29.96
DBHmed	12.72	8.69	19.83	2.50	20.27
Hmin	6.65	5.24	11.43	1.41	21.12
Hmax	22.00	14.86	29.00	2.39	11.05
Hmed	13.18	9.84	18.83	1.76	13.38
N	14.25	3.00	30.00	4.96	34.90
Dq	14.59	9.16	22.79	3.11	22.74

Where: C = carbon stock (Mg.ha⁻¹), DBHmin = minimum breast height diameter (DBH) of the sample plot (cm), DBHmax = maximum DBH of the sample plot (cm), DBHmed = mean DBH of the sample plot (cm), Hmin = minimum total height of the sample plot (m), Hmax = maximum total height of the sample plot (m), Hmed = mean total height of the sample plot (m), N = number of trees in the sample plot, Dq = mean square diameter of the sample plot (cm).

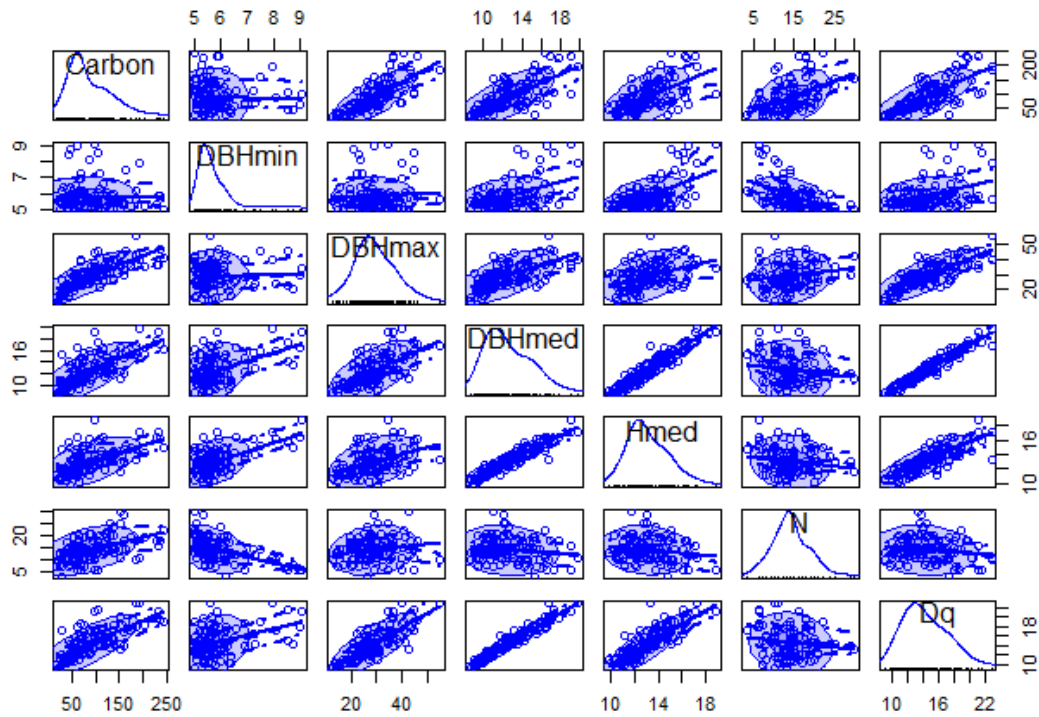


Figure 2. Distribution and dispersion of the independent variables selected by stepwise for carbon stock modelling and their respective confidence intervals.

Distribución y dispersión de las variables seleccionadas y sus respectivos intervalos de confianza.

Table 2. Support Vector Machine Parameters in the estimation of carbon stock in a secondary semideciduous forest.

Parámetros de máquina de vectores de soporte para estimar las existencias de carbono en un bosque semideciduo estacional secundario.

SVM	Parameters
Type	<i>eps-regression</i>
Kernel	Radial basis function
Cost	1
Gamma	0.2
Epsilon	0.1
Numbers of support vectors	40

from the selected ANN that presented the lowest error among the others evaluated, composed by six neurons in the hidden layer.

From the 5-6-1 architecture artificial neural network, an equation system was extracted to predict carbon stock per plot, with coefficients derived from weights generated by neural network neurons. This system was used to predict the carbon stock of plots that make up the database for validation.

Model [5] expresses the relationship between the hidden layer and the response variable, where β_0 is the bias and the other coefficients are weights related to each neuron. Model [6] represents the activation function used in each hidden layer neuron, derived from the logistic model. Finally, model [7] is the result of the relationship between input variables and the respective hidden layer neurons, and a model is generated for each neuron.

$$Carbon' = \beta_0 + \beta_1 * z_1 + \beta_2 * z_2 + \beta_3 * z_3 + \beta_4 * z_4 + \beta_5 * z_5 + \beta_6 * z_6 \quad [5]$$

$$z_n = \left[\frac{1}{1 + e^{-w_i}} \right] \quad [6]$$

$$w_i = \beta_{0,n} + \beta_{1,n} * DBHmin_i' + \beta_{2,n} * DBHmax_i' + \beta_{3,n} * DBHmed_i' + \beta_{4,n} * Hmed_i' + \beta_{4,n} * N_i' \quad [7]$$

where: β_0 : bias, β_n : model coefficient associated with neuron n, β_n : model coefficient between input variable k and neuron n, z_n : n-th hidden layer neuron response, w_i : sum of products between weights and inputs.

The coefficients of the system of equations extracted from the selected artificial neural network are presented in table 3.

The support vector machine and the model extracted from the artificial neural network were applied to the data set intended for validation. The analyzed techniques presented satisfactory performance in the modeling of carbon

stock by plot, due to homogeneous distribution and low dispersion of residues and with predicted values close to those observed, as it can be observed in figures 4 and 5, respectively.

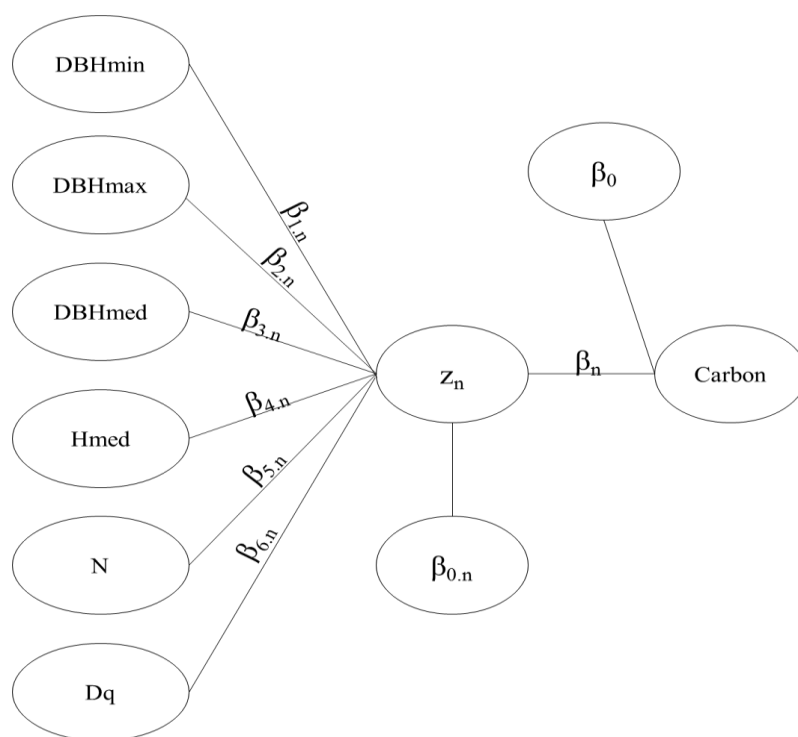


Figure 3. Architecture of the selected artificial neural network, composed of six neurons in the hidden layer. Input variables: DBHmin = minimum plot DBH (cm), DBHmax = maximum DBH of the plot (cm), DBHmed = mean DBH of the plot (cm), Hmed = Average total tree height of the plot (m), N = Number of trees in the plot. Output variable: C = carbon stock (Mg.ha⁻¹), β_0 : bias; β_n : coefficient associated with the neuron n , $\beta_{k,n}$: coefficient of the model between the input variable k and the neuron n , z_n : response of n -th hidden layer neuron.

Arquitectura de la red neuronal artificial seleccionada, compuesta por seis neuronas en la capa oculta. Variables de entrada: DAPmin = DAP mínimo de la parcela (cm); DAPmax = DAP máximo de la parcela (cm); DAPmed = DAP medio de la parcela (cm); HTmed = Altura total promedio de los árboles en la parcela (m); N = Número de árboles en la parcela. Variable de salida: C = stock de carbono (Mg.ha⁻¹).

Table 3. Coefficients (β 's) of each neuron (N) of the hidden layer and of the ANN output layer.

Parámetros (β) de la red neuronal artificial. N representa el número de neuronas.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
ANN	0.8609	-0.7323	0.7619	0.6785	-1.7671	-1.6967	1.0353
N1	-2.3277	2.0301	0.1250	9.7276	-6.8163	6.8954	-
N2	-1.3559	0.6158	2.5005	2.8092	-0.3280	-0.9826	-
N3	-1.7104	2.3549	0.9111	0.6736	-2.8348	6.2268	-
N4	-0.1308	0.1538	0.5098	-1.6662	1.0934	-0.4838	-
N5	0.1174	-1.2125	0.4046	-1.1707	0.2400	-0.3406	-
N6	-0.0674	-2.5130	0.6372	0.7235	-0.2786	1.3868	-

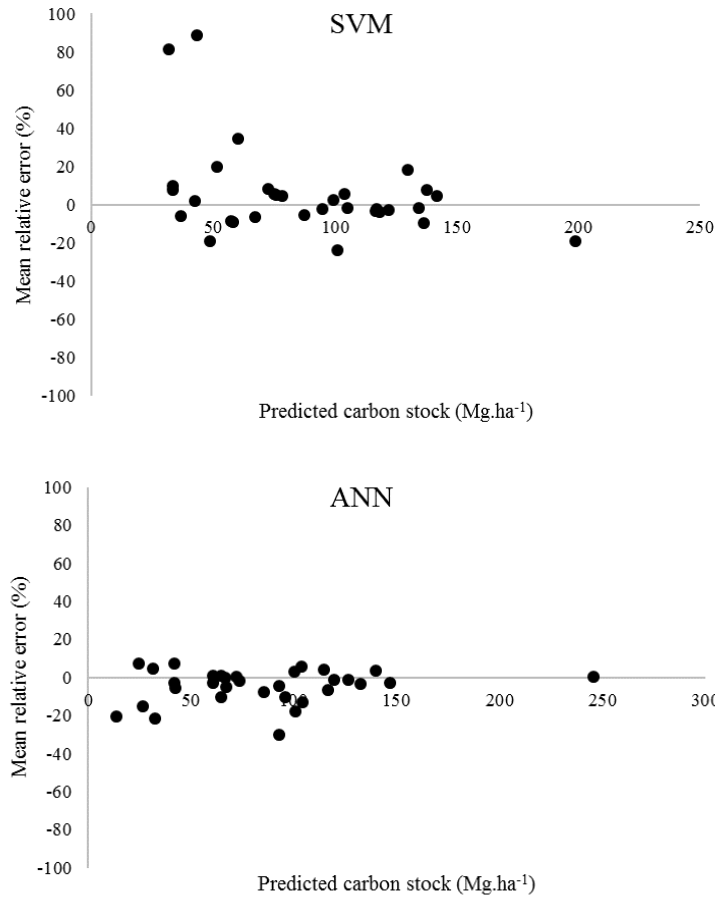


Figure 4. Residuals scatter plots of carbon stock modeling in a secondary semideciduous forest using machine learning algorithms.
Dispersión de residuos del modelado de existencias de carbono en un bosque secundario semideciduo mediante algoritmos de aprendizaje automático.

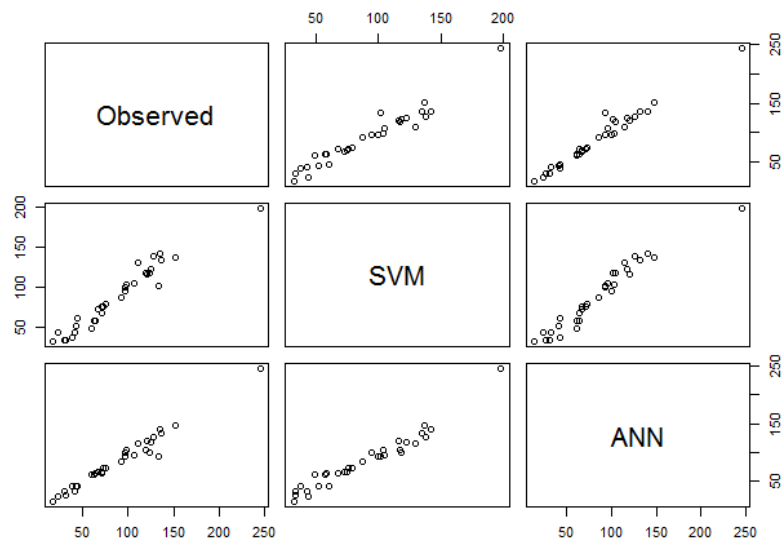


Figure 5. Predicted versus observed graphs of the analyzed techniques (artificial neural network and support vector machine) used to estimate carbon stock in a secondary semideciduous forest.
Criterios para evaluar el desempeño de la máquina de vectores de soporte (SVM) y los algoritmos de la red neuronal artificial (ARN) utilizados para estimar las existencias de carbono en un bosque secundario semideciduo.

According to the graphs (figures 4 and 5), there is a slight superiority of ANN, which presented more concentrated residues around zero and predicted values closer to the real ones. SVM presented higher residual dispersion and two relative error values above 80 %, while ANN presented a maximum relative error of 30 %.

The performance evaluation criteria of the analyzed techniques are presented in table 4. The qualities of the estimates made by SVM and by the model extracted from ANN were evaluated on both data used in the training and validation dataset.

The estimates of the analyzed techniques were strongly correlated with the observed values, showing correlation above 0.99 in the training phase and 0.96 in the validation phase. Error magnitudes, represented by RMSE, were below 10 % in the training phase and 15 % in the validation. The lower the RMSE, the higher the accuracy of the estimates, and the optimal situation when it is zero (Mehtätalo *et al.* 2006). Bias indicated slight underestimation trends in the training phase; whereas in validation, there was a tendency of SVM to overestimate carbon stock values (5.6479 %) and of ANN to underestimate them (-4.6217 %). Mean SVM relative error increased from 6.7689 % in the training phase to 13.5185 % in the validation phase; whereas in ANN, this increase went from 4.8035 % to 6.9375 %.

DISCUSSION

The average of tree carbon stock contained in the forest with predominance of *Anadenanthera* sp. (94.25 Mg ha⁻¹) is above the average carbon stock for this vegetation type in south-central Minas Gerais (55 Mg ha⁻¹, Scolforo *et al.* 2015) and compatible with other local studies in semideciduous seasonal forests in the study region. For instance, Ribeiro *et al.* (2009), quantifying the biomass and tree carbon stock in a mature semideciduous forest in Viçosa, Minas Gerais, Brazil, found 166.67 Mg ha⁻¹ and 83.34 Mg ha⁻¹ for biomass and carbon stocks, respectively. Likewise, Figueiredo *et al.* (2015) that evaluated the dynamics of the tree carbon stock in a semideciduous

forest in Minas Gerais, Brazil, found an average carbon stock of 71.81 Mg ha⁻¹.

Although the average carbon stock was relatively high, its variation among plots (CV%) was also significant. This is mainly due to the fact that the carbon stock variable reflects the variations in other dendrometric variables. Moreover, in natural multiage forests, characteristics such as high ecological complexity, spatial variations in species structure and composition, the presence of clearings and other factors can lead to great variability in biomass/carbon stock values (Soriano-Luna *et al.* 2018) among sample plots.

There is a strong relationship between mean DBH and mean height. Overall, there is a tendency of carbon stock to be positively correlated to the other variables, except for the minimum DBH, in which there was no direct or indirect proportional relationship, which is evidenced by the elliptic shape of the dispersion between these variables. In general, the relationship between carbon and independent variables tended towards linearity. There are, however, some nonlinear behaviors between variables, such as between carbon stock and maximum DBH. In this context, it is worth to emphasize that artificial intelligence/machine learning has the ability to implicitly detect any nonlinear relationship between the response variable and explanatory variables. In addition to the fact that there are no assumptions regarding input data, such as independence and normality, and its high capacity for learning and generalization.

The main purpose of using these techniques, as in classical regression, is their application to data that were not used in their training. ANN presented better performance in the two analyzed phases, training and validation, which indicates its superiority for modeling the carbon stock per plot in the analyzed data set, when compared to SVM.

ANN was able, with the available variables, to explain almost all the variation in carbon stock in the study area. Several studies have demonstrated the superiority of artificial neural networks when compared to other techniques (Özçelik *et al.* 2013, Vendruscolo *et al.* 2015). This superiority can be explained by the ability of neural networks

Table 4. Performance evaluation criteria of the support vector machine (SVM) and artificial neural network (ANN) algorithms used in carbon stock estimation in a secondary semideciduous forest.

Criteria para evaluar el desempeño de la máquina de vectores de soporte (SVM) y la red neuronal artificial (ANN).

	Train		Validation	
	SVM	ANN	SVM	ANN
Mean relative error (%)	6.7689	4.8035	13.5185	6.9375
Bias (%)	-0.2610	-1.2371	5.6479	-4.6217
Root mean square error (%)	7.5858	5.4185	14.5550	10.8191
Coefficient of correlation	0.9900	0.9951	0.9690	0.9828

to detect implicit information and nonlinear relationships between the response variable and explanatory variables provided as examples and to generalize the assimilated knowledge to an unknown data set.

It is worth noting that, although underperforming ANN, SVM was very efficient in estimating carbon stock in the study area. SVM has the advantage over ANN that no evaluation is required after its construction, as it occurs in ANN to select the best network. This is due to the quadratic optimization that occurred during SVM training (Cristianini and Shawe-Taylor 2000) which allows the same result to be obtained for each system configuration whenever applied to the same data set. However, ANNs have more elements to be manipulated; besides, the initialization of neuron parameters occurs at random (Haykin 2001). Thus, each trained network will have slight differences in estimates, even if the same architecture is maintained. These differences highlight the practicality of SVM in relation to ANN as SVM excludes the subjectivity of the operator in choosing the best network to be applied to the database.

Therefore, both approaches were able to explain much of the carbon stock variation in the study area. This is mainly due to the robustness of ANN and SVM. The small part of the carbon stock variation, not explained by the variables in question, is due to the various factors not considered in the present study that are known to affect the variability of carbon stock in forests, such as species diversity, forest size, degree anthropization, among many others and their interactions (McNicol *et al.* 2018)

The results presented in this study provide insights for future assessments of the use of machine learning techniques to obtain carbon stock estimates. Some examples of potential applications are estimates of carbon stock in other forest compartments, such as soil and tree roots, carbon stock estimates through the association between machine learning techniques and remote sensing variables, among others.

CONCLUSIONS

The present study brings important contributions in the modeling of carbon stocks in forests through the use of machine learning. The machine learning techniques performed satisfactorily, and a new model extracted from an artificial neural network for carbon stock prediction in Mg ha⁻¹ is efficient, with potential application in other secondary semideciduous seasonal forests.

Carbon stock modeling using a model extracted from the artificial neural network training presented better performance than that presented by the support vector machine, using the same variables of the analyzed data set.

Determining, modeling and supplying forest carbon stock data are strong scientific and social demands currently, as tree carbon storage is considered a key environmental service in mitigating current climate change by sequestering CO₂ from forest atmosphere.

ACKNOWLEDGMENTS

Authors are grateful for the financial support provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), Federal University of Lavras, MG, Brazil, and Science Forest Department.

REFERENCES

- Ali KM, MJ Pazzani. 1996. Error reduction through learning multiple descriptions. *Machine Learning* 24(3): 173-202. DOI: [10.1007/BF00058611](https://doi.org/10.1007/BF00058611)
- Alvares CA, JL Stape, PC Sentelhas, JLM Gonçalves, G Sparovek. 2013. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift* 22: 711-728. DOI: [10.1127/0941-2948/2013/0507](https://doi.org/10.1127/0941-2948/2013/0507)
- Bonan GB. 2008. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* 320(5882): 1444-1449. DOI: [10.1126/science.1155121](https://doi.org/10.1126/science.1155121)
- Braga AP, APLF Carvalho, TB Ludemir. 2007. *Redes Neurais Artificiais: Teoria e Aplicações*. Rio de Janeiro, Brasil. Editora LTC. 226 p.
- Canadell JG, MR Raupach. 2008. Managing forests for climate change mitigation. *Science* 320(5882): 1456-1457. DOI: [10.1126/science.1155458](https://doi.org/10.1126/science.1155458)
- Chave J, M Réjou-Méchain, A Búrquez, E Chidumayo, MS Colgan, WBC Delitti, A Duque, T Eid, PM Fearnside, RC Goodman, M Henry, A Martínez-Yrizar, WA Mugasha, HC Muller-Landau, M Mencuccini, BW Nelson, A Ngomanda, EM Nogueira, E Ortiz-Malavassi, R Péliissier, P Ploton, CM Ryan, JG Saldarriaga, G Vieilledent. 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology* 20(10): 3177-3190. DOI: [10.1111/gcb.12629](https://doi.org/10.1111/gcb.12629)
- Chazdon RL, EN Broadbent, DMA Rozendaal, F Bongers, AMA Zambrano, TM Aide, P Balvanera, JM Becknell, V Boukili, PHS Brancalion, D Craven, JS Almeida-Cortez, GAL Cabral, BD Jong, JS Denslow, DH Dent, SJ Dewalt, JM Dupuy, SM Durán, MM Espírito-Santo, MC Fandino, RG César, JS Hall, JL Hernández-Stefanoni, CC Jakovac, AB Junqueira, D Kennard, SG Letcher, M Lohbeck, M Martínez-Ramos, P Massoca, JA Meave, R Mesquita, F Mora, R Muñoz, R Muscarella, YRF Nunes, S Ochoa-Gaona, E Orihuela-Belmonte, M Peña-Claros, EA Pérez-García, D Piotto, JS Powers, J Rodríguez-Velazquez, IE Romero-Pérez, J Ruíz, JG Saldarriaga, A Sanchez-Azofeifa, NB Schwartz, MK Steininger, NG Swenson, M Uriarte, MV Breugel, HVD Wal, MDM Veloso, H Vester, ICG Vieira, TV Bentos, GB Williamson, L Poorter. 2016. Carbon sequestration potential of second-growth forest regeneration in the Latin American tropics. *Science Advances* 2(5):e1501639. DOI: [10.1126/sciadv.1501639](https://doi.org/10.1126/sciadv.1501639)
- Cherkassky V, F Mulier. 1998. *Learning from data: Concepts, theory, and methods*. New York, United States. Wiley. 560 p.
- Cigizoglu HK, Ö Kisi. 2006. Methods to improve the neural network performance in suspended sediment estimation. *Journal of Hydrology* 317(3): 221-238. DOI: [10.1016/j.jhydrol.2005.05.019](https://doi.org/10.1016/j.jhydrol.2005.05.019)
- Cristianini N, J Shawe-Taylor. 2000. An introduction to support vector machines: and other kernel-based learning methods.

- New York, United States. Cambridge University Press. 204 p.
- Cordeiro NG, KMG Pereira, MCNS Terra, JM Mello. 2018. Variação temporal do estoque de carbono e volume de madeira em um fragmento de cerrado *sensu stricto*. *Enciclopédia Biosfera* 15(28): 931-942. DOI: [10.18677/EnciBio_2018B76](https://doi.org/10.18677/EnciBio_2018B76)
- Dantas D, LOR Pinto, MCNS Terra, N Calegario, MLR Oliveira. 2020a. Reduction of sampling intensity in forest inventories to estimate the total height of eucalyptus trees. *Bosque* 41(3): 353-364. DOI: [10.4067/S0717-92002020000300353](https://doi.org/10.4067/S0717-92002020000300353)
- Dantas D, N Calegario, FW Acerbi Júnior, SPC Carvalho, MAI Júnior, EA Melo. 2020b. 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](https://doi.org/10.1590/01047760202026012668)
- Dantas D, MCNS Terra, LOR Pinto, N Calegario, SM Maciel. 2021. Above and belowground carbon stock in a tropical forest in Brazil. *Acta Scientiarum. Agronomy* 43:e48276. DOI: [10.4025/actasciagron.v43i1.48276](https://doi.org/10.4025/actasciagron.v43i1.48276)
- Diamantopoulou MJ. 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. *Computers and Electronics in Agriculture* 48(3): 235-244. DOI: [10.1016/j.compag.2005.04.002](https://doi.org/10.1016/j.compag.2005.04.002)
- Figueiredo LTM, CPB Soares, AL Sousa, HG Leite, GF Silva. 2015. Dinâmica do Estoque de Carbono em Fuste de Árvores de uma Floresta Estacional Semidecidual. *Cerne* 21(1): 161-167. DOI: [10.1590/01047760201521011529](https://doi.org/10.1590/01047760201521011529)
- Haykin, S. 2001. Redes neurais: princípios e prática. 2.ed. Porto Alegre, Brasil. Bookman. 900 p.
- Heddam S, O Kisi. 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology* 559: 499-509. DOI: [10.1016/j.jhydrol.2018.02.061](https://doi.org/10.1016/j.jhydrol.2018.02.061)
- Leite HG, VCL Andrade. 2002. Um método para condução de inventários florestais sem o uso de equações volumétricas. *Revista Árvore* 26(3): 321-328. DOI: [10.1590/S0100-67622002000300007](https://doi.org/10.1590/S0100-67622002000300007)
- Marques RFPV, MCNS Terra, VA Mantovani, AF Rodrigues, GA Pereira, RA Silva, CR Mello. 2019. Rainfall Water Quality Under Different Forest Stands. *Cerne* 25(1): 8-17. DOI: [10.1590/01047760201925012581](https://doi.org/10.1590/01047760201925012581)
- Mcnicol IM, CM Ryan, KG Dexter, SMJ Ball, M Williams. 2018. Aboveground carbon storage and its links to stand structure, tree diversity and floristic composition in south-eastern Tanzania. *Ecosystems* 21(4): 740-754. DOI: [10.1007/s10021-017-0180-6](https://doi.org/10.1007/s10021-017-0180-6)
- Mehtätalo L, M Maltamo, A Kangas. 2006. The use of quantile trees in the prediction of the diameter distribution of a stand. *Silva Fennica* 40(3):id333. DOI: [10.14214/sf.333](https://doi.org/10.14214/sf.333)
- Melo EA, N Calegario, AR Medonça, EL Possato, JA Alves, MAI Júnior. 2017. Modelagem não Linear da Relação Hipsométrica e do Crescimento das Árvores Dominantes e Codominantes de *Eucalyptus* sp. *Ciência Florestal* 27(4): 1325-1338. DOI: [10.5902/1980509829895](https://doi.org/10.5902/1980509829895)
- Meyer D. 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-1. Consultado 24 aug. 2019. Disponível em <https://CRAN.R-project.org/package=e1071>
- Özçelik R, MJ Diamantopolou, F Crecente-Campo, U Eler. 2013. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. *Forest Ecology and Management* 306: 52-60. DOI: [10.1016/j.foreco.2013.06.009](https://doi.org/10.1016/j.foreco.2013.06.009)
- R Development Core Team. 2018. R: a language and environment for statistical computing. ViennaR Foundation for Statistical Computing, 2018. Consultado 24 aug. 2019. Disponível em <https://www.r-project.org/>
- Réjou-Méchain M, A Tanguy, C Pioniot, J Chave, B Hérault. 2017. biomass: an r package for estimating above-ground biomass and its uncertainty in tropical forests. *Methods in Ecology and Evolution* 8(9): 1163-1167. DOI: doi.org/10.1111/2041-210X.12753
- Ribeiro SC, LAG Jacovine, CPB Soares, SV Martins, AL Souza, AMB Nardelli. 2009. Quantificação de biomassa e estimativa de estoque de carbono em uma floresta madura no município de viçosa, Minas Gerais. *Revista Árvore* 33(5): 917-926, 2009. DOI: [10.1590/S0100-67622009000500014](https://doi.org/10.1590/S0100-67622009000500014)
- Scolforo HF, JRS Scolforo, CR Mello, JM Mello, ACF Filho. 2015. Spatial distribution of aboveground carbon stock of the arboreal vegetation in Brazilian Biomes of Savanna, Atlantic Forest and Semi-arid woodland. *PLoS ONE* 10(6): 1-20. DOI: [10.1371/journal.pone.0128781](https://doi.org/10.1371/journal.pone.0128781)
- 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>
- Soares FAAMN, EL Flores, CD Cabacinha, GA Carrijo, ACP Veiga. 2011. Recursive diameter prediction and volume calculation of eucalyptus trees using Multilayer Perceptron Networks. *Computers and Electronics in Agriculture* 78(1): 19-27. DOI: [10.1016/j.compag.2011.05.008](https://doi.org/10.1016/j.compag.2011.05.008)
- Soriano-Luna MLA, et al. 2018. Determinants of above-ground biomass and its spatial variability in a temperate forest managed for timber production. *Forests* 9(8): 1-20. DOI: [10.3390/f9080490](https://doi.org/10.3390/f9080490)
- Steinwart I, A Christmann. 2008. Support vector machines. Berlin, Germany. Springer. 601 p.
- Terra MCNS, D Dantas, LOR Pinto, N Calegario, SM Maciel. 2019. Estimativa do estoque de carbono em floresta semidecidual: uma comparação entre regressão e redes neurais artificiais. In Oliveira AC ed. Fontes de Biomassa e Potenciais de Uso. Ponta Grossa, Brasil. Editora Atena. p. 24-35.
- Thomas SC, AR Martin. 2012. Carbon content of tree tissues: A synthesis. *Forests* 3(2):332-352. DOI: [10.3390/f3020332](https://doi.org/10.3390/f3020332)
- Tong S, DKoller. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2: 45-66. DOI: [10.1162/153244302760185243](https://doi.org/10.1162/153244302760185243)
- Valença M. 2010. Fundamentos das redes neurais: exemplos em Java, ver. ampli. 2 ed. Olinda, Brasil. Livro Rápido. 386 p.
- Vendruscolo DGS, R Drescher, HS Souza, JPM Moura, FMD Mamoré, TAS Siqueira. 2015. Estimativa da altura de eucalipto por meio de regressão não linear e redes neurais artificiais. *Revista Brasileira de Biometria* 33(4): 556-569, 2015. DOI: [10.13140/RG.2.1.1742.5684](https://doi.org/10.13140/RG.2.1.1742.5684)