academicJournals

Vol. 11(39), pp. 3771-3776, 29 September, 2016
DOI: 10.5897/AJAR2016.11533
Article Number: B98AE5360826
ISSN 1991-637X

http://www.academicjournals.org/AJAR

African Journal of Agricultural Research

Full Length Research Paper

Selection of input vectors for estimation of aboveground biomass of *Mimosa scabrella* Benth. using an artificial neural network

Aline Bernarda Debastiani*, Ana Paula Marques Martins, Carlos Roberto Sanquetta, Sebastião do Amaral Machado, Ana Paula Dalla Corte and Edilson Urbano

Post Graduation in Forest Engineering, Federal University of Parana, City of Curitiba, State of Paraná, Brazil.

Received 8 August, 2016; Accepted 19 September, 2016

The objective of this study was to evaluate the effect of input vectors in an artificial neural network (ANN) and determine their best combination to estimate the individual dry biomass of native bracatinga. The dataset consisted of 178 trees of *Mimosa scabrella* Benth. (bracatinga) from the Metropolitan Region of Curitiba. The ANN used was a Multi-Layer Perceptron; the learning algorithm was the Levenberg-Marquardt, consisting of an occult layer where 50% of the data were used for training, 25% for cross-validation and the other 25% for the test. The input vectors were all the variables collected in the field, such as: diameter at breast height (dbh), total height (ht), crown height (hc), stem height (hf), crown diameter (dc) and age (i). The treatment 1 consisted of all the vectors; after the MLP trained, the Garson algorithm was executed for obtaining relative contribution of each vector; the less important vector was deleted and the MPL was retrained (treatment 2) and so on until only one vector was left. Based on the coefficient of determination and root mean square error, treatment 3 provided the best performance (i, hc, ht and dbh), followed by treatment 6 (dbh). The method of selecting attributes by the Garson algorithm was remarkable and provided the definition of essential vectors, allowing minimal costs and optimizing the performance of the MLP.

Key words: Bracatinga, multi-layer perceptron, Garson algorithm, relative contribution.

INTRODUCTION

Studies on forest biomass are done with various objectives, among which it is possibility to highlight the quantification of nutrient cycling, the quantification for energetic ends and as information base for studies on carbon sink (Trautenmüller, 2015). Sanquetta (2002) affirms that in researches focused on the carbon fixation, the biomass is one of the most relevant factors, and for this reason, it must be determined and estimated precisely, or else there will not be consistency in the quantification of the carbon fixed in forest ecosystems.

According to Sanquetta et al. (2015), though biomass is important for the quantification of carbon in the plants, its

*Corresponding author. E-mail: aline.debastiani@gmail.com.

direct determination is complex, expensive and destructive. For this reason, the quantification may be performed in an indirect way, which consists of estimates generally made by mathematical relations, as reasons or regressions, with data coming from forest inventory. The direct quantification may also be performed with data from a remote sensing in a geographic information system (Silveira et al., 2008).

However, when used in natural forests, the use of the referred techniques becomes more complex due to the great floristic, physiognomic and phenological diversity of this forest type, and so may present limitations (Sanquetta et al., 2015). Besides, a single equation may not be able to reproduce a large variation regarding natural forests, in addition to the need of this technique to attend to some prerequisites as additivity and linearity, residue independence, homoscedasticity and residue normality (Osborne and Waters, 2002).

In contrast, some machine learning techniques have presented themselves as promising alternatives to conventional statistical methods. Among these techniques, the artificial neural networks (ANN) have been used successfully in the forest sector in estimating dendrometric variables, such as height (Ferreira et al., 2014; Binoti et al., 2013), volume (Binoti et al., 2014; Cordeiro et al., 2015), thinning (Martins et al., 2016; Mendonça et al., 2015) and growth and production (Castro et al., 2013).

However, this technique presents some difficulties in understanding how they reach such estimates, and for this reason are named as black box system, for it is not possible to know what are the intermediate relations between the input and output variables, knowing only that this relation consists of adjusting synaptic weights.

There are techniques commonly utilized that seek to clarify how the ANN reaches such estimates, for example: the neural interpretation diagram technique (Gozlan et al., 1999), sensitivity analysis (Olden and Jakson, 2002) and the Garson algorithm (Garson, 1991). From these, the Garson algorithm has proven to be quicker and more decisive to be strictly quantitative (Kalteh, 2008).

It is necessary for the acquisition of some answers in the forest sector, the measurement of dendrometric variables with some level of difficulty in obtainment, it is of utmost importance to determine which are the input vectors and their effects, as well as identify the best combination of variables that produce the most accurate response. In this sense, the Garson algorithm is highlighted; it consists of considering the variation of absolute values of synaptic weights between the vectors of the input layer and the output layer, with the objective of determining the relative relevance of each vector of the input layer (Garson, 1991).

Thus, the objective of the present research was to evaluate the contribution of input vectors of an ANN to estimate the individual aboveground biomass of *Mimosa scabrella* Benth. using the Garson (1991) algorithm and determine the best set of input vectors.

MATERIALS AND METHODS

Data collection

In this study, data from native bracatinga trees of the Metropolitan Region of Curitiba were used. The total data set was 178 trees, which were sampled seeking representation in the age classes and diameters. The methodology was obtaining biomass according to Sanquetta (2002). More details may be obtained in Urbano (2007).

The measured variables were: age (years), crown diameter (m), crown height (m), stem height (m), total height (m) and diameter of breast height (cm) collected at 1.3 m from the ground. With the objective to verify the correlation between the input variables, the analysis of linear correlation of Pearson was performed between these and output variables, in this case, the total individual aboveground dry biomass (kg).

Artificial neural network used

The ANN used was the multi-layer perceptron (MLP) applied to the Matlab 2014a software in the Neural Network Toolbox. The learning algorithm used was the Levenberg-Marquardt backpropagation due to its quickness and stable convergence, being basically an integration of classic methods of Error Back Propagation and Gauss-Newton (Hagan and Menhaj, 1994). A learning rate of 0.01 was adopted; this represents how much of the error is backpropagated.

A tangent hyperbolic sigmoidal activation function, which compresses the answer to a known interval, from -1 to 1 was used. The activation functions to prevent the saturation and attenuation of the input signal (Haykin, 2001).

The MLP constituted of a hidden layer, in which, according to Atkinson and Tatnal (1997), generally is sufficient. In the hidden layer, the number of neurons (units of signal processing between the input layer and the output layer) varied ±5 with the satisfactory number of neurons, which, according to Heath (2010), corresponds to a relation of 10 times more training samples (Equation 1) than weights or the also named synaptic weights (Equation 2), which concentrates the knowledge of the network through the weighting of the connection between the neurons of the input layer and the hidden layer. Ten initializations of synaptic weights were also evaluated.

$$Neurons = (-1 + (Train\_samples - O)/(I + O + 1))/\mathrm{Re}ason \quad (1)$$

$$Weights = (I + 1) * Neurons + (Neurons + 1) * O \quad (2)$$

Where, Neurons: number of satisfactory neurons; Train_samples: number of training samples; O: number of output vectors; I: number of input vectors; Reason: reason of about 10 times more training samples than the number of hidden layer weights.

For each combination of input vectors (treatments), many configurations of the MLP were evaluated (initialization of weights and number of neurons of the hidden layer), and the best performance configuration was selected from the determination coefficient (R^2) in the testing phase. The total dataset was subdivided in three parts, constituting 50, 25 and 25% of training, cross validation and the testing statistics calculation, respectively. The training consists of the process in which the input value is presented as the ANN and the corresponding answer (output

vector), adjusting the weights and connections to obtain the expected output. After the MLP was trained, the individual aboveground dry biomass (ba) for the whole dataset was obtained. The cross validation technique was used as the criterion to stop training and to avoid overfitting of the MLP, being done in a distinct dataset. The goal is to build a MLP in a manner that the same may simulate, for a distinct dataset (testing), the answer variable and obtain good performance in this phase to be considered well trained. The MLP input vectors were the dendrometric variables: Age (i), crown diameter (dc), crown height (hc), stem height (hf) and total height (ht) and diameter at breast height (dbh). These were standardized to the same scale to improve convergence (FU, 1994). The MLP output vector was configured to correspond to the individual dry aboveground biomass (ba).

Relative contribution of vectors

For the synaptic weight that presented the largest R^2 for treatment 1 (all vectors), the relative contribution of each input vector was calculated through the Garson algorithm (1991) (Equation 3). The vector with the lowest relative contribution was removed from the configuration of the next treatment, and this process was repeated until only one value remained.

$$CR_{ik} = \frac{\sum_{j=1}^{L}\left(\frac{w_{ij}}{\sum_{r=1}^{N} w_{rj}} v_{jk}\right)}{\sum_{i=1}^{N}\left(\sum_{j=1}^{L}\left(\frac{w_{ij}}{\sum_{r=1}^{N} w_{rj}} v_{jk}\right)\right)} \quad (3)$$

Where, CR_{ik} represents the influence percentage of each input vector *i* on the output vector *k*; $\sum_{r=1}^{N} w_{rj}$ is the sum of the weights connecting the input layer *i* and the *j* neuron; *N* corresponds to the total of input vectors; *L* corresponds to the total of neurons of the hidden layer, V_{jk} corresponds to the weights of the connection between the *j* neuron and the *k* input vector.

The relative contribution of Garson seeks to determine the best combination of input vectors, based on the selection of attributes, optimizing the adjustment process.

Statistical analysis

The performance at the different treatments of the MLP was calculated by the adjustment statistics and model selection, which is the determination coefficient (R^2) (Equation 4) and the root of the mean quadratic error in percentage (RMSE) (Equation 5), in addition to the graphical analysis of residue dispersion.

$$R^2 = \frac{[\sum(ba_{obsi} - \bar{ba}_{obs})*(ba_{simi} - \bar{ba}_{simi})]^2}{\sum(ba_{obs} - \bar{ba}_{obsi})^2 * \sum(ba_{simi} - \bar{ba}_{simi})^2} \quad (4)$$

$$RMSE\% = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(ba_{simi} - ba_{obsi})^2}}{ba_{máx} - ba_{mín}} * 100 \quad (5)$$

Where, ba_{obsi}: observed aboveground biomass (kg); ba_{simi}: simulated aboveground biomass (kg); $\bar{ba}$: average aboveground biomass observed (kg); $ba_{mín}$: minimum observed value of aboveground biomass (kg); $ba_{máx}$: maximum observed value of aboveground biomass (kg); *n*: number of observations.

RESULTS AND DISCUSSION

As shown in Table 1, the bracatinga have an age range varying from 4 to 17 years. The lowest variability was found in total height (CV = 20.01%) and the highest for the individual dry aboveground biomass (CV = 108.13%). Since treatment is done with native bracatinga trees, the high data variability was already expected.

The Pearson linear correlation (r) between the input and output variables was calculated with the objective to determine the existing relations between them (Table 2). The vector corresponding to the dbh was the variable correlating to ba (r = 0.954), followed by dc (r = 0.794) and i (r = 0.751), with a strong positive correlation, indicating that the higher the dbh, dc, and the older the tree is, the more individual dry aboveground biomass is produced by it.

Table 3 shows the input vectors, the architecture and the adjustment statistics for training and testing for each treatment. The number of neurons in the hidden layer ranged from 3 (treatment 5) to 6 (treatment 3), representing a low number of neurons needed to estimate the individual dry aboveground biomass of bracatinga.

The adjustment statistics for the training phase resulted in R^2 varying from 0.948 (treatment 2) to 0.962 (treatment 4), and the RMSE varying from 5.281 (treatment 2) to 4.343% (treatment 5). It should be noted that the RMSE for the testing tends to decrease from treatment 1, which contains all possible vectors, until treatment 3, due to the removal of less relevant vectors according to their relative contribution. Since this treatment (3) was the one with the best performance of adjustment statistics (R^2 = 0.934 e RMSE = 8.304%), demonstrating the importance of the Garson algorithm in the selection of attributes and the ideal composition of the input vectors to estimate dendrometric variables.

Having determined the treatment with the best input vectors, the removal of less relevant vectors from following treatments resulted in the decrease of model adjustment statistics, with the exception of treatment 6 that was classified as the second best performance (R^2 = 0.919, RMSE = 9.388%).

The relative contribution of each vector at the treatments may be observed in Figure 1, which served as base for the composition of input vectors of each treatment. Treatment 1 was composed of all input vectors, in which, by the relative contribution, the least important vector, in this case hf, was removed from the composition of treatment 2; followed by the removal of the dc vector from the composition of treatment 3, and

Table 1. Descriptive statistics of the input and output vectors to estimate the individual aboveground biomass of *Mimosa scabrella* Benth. in the Metropolitan Region of Curitiba.

Variables	i (years)	dc (m)	hc (m)	hf (m)	ht (m)	dbh (cm)	ba (kg)
Minimum	4	0.55	1.70	2.56	9.15	4.80	6.90
Average	9	3.11	5.16	8.95	14.11	13.64	101.32
Maximum	17	8.85	12.00	15.45	21.80	35.00	586.90
CV%	31,31	55.48	39.43	27.74	20.01	46.61	108.13

CV%: Coefficient of variation.

Table 2. Pearson linear correlation (r) between the input and output vectors to estimate the individual dry aboveground biomass for the *Mimosa scabrella* Benth. in the Metropolitan Region of Curitiba.

Variables	i (years)	dc (m)	hc (m)	hf (m)	ht (m)	dbh (cm)	ba (kg)
i (years)	1.000	-	-	-	-	-	-
dc (m)	0.700	1.000	-	-	-	-	-
hc (m)	0.386	0.586	1.000	-	-	-	-
hf (m)	0.332	0.286	-0.231	1.000	-	-	-
ht (m)	0.570	0.674	0.517	0.713	1.000	-	-
dbh (cm)	0.759	0.836	0.588	0.339	0.722	1.000	-
ba (kg)	0.751	0.794	0.573	0.281	0.660	0.954	1.000

Table 3. Performance of the MLP with regards to the different combinations of input vectors, chosen based on the attribute selection described by the Garson algorithm (1991).

S/N	Input vectors	Architecture	Training		Test	
			R^2	RMSE (%)	R^2	RMSE (%)
1	i, dc, hc, hf, ht, dbh	6/4/1	0.959	4.533	0.914	10.309
2	i, dc, hc, ht, dbh	5/5/1	0.948	5.281	0.903	9.525
3	i, hc, ht, dbh	4/6/1	0.950	5.094	0.934	8.304
4	hc, ht, dbh	3/5/1	0.962	4.393	0.917	9.475
5	hc, dbh	2/3/1	0.961	4.343	0.919	9.436
6	dbh	1/5/1	0.959	4.355	0.919	9.388

N: Number of treatment; Architecture: represents the number of neurons in the input layer/hidden/output.

successively for the remaining treatments, until only dbh remained in the composition of the last treatment, in other words, treatment 6.

As shown in Figure 1, dbh presented the largest relative contribution in all treatments, and this fact may be explained by the largest linear correlation of this variable with the individual dry aboveground biomass (r = 0.954), and also by the efficiency of using only the dbh at the MLP that resulted in R^2 = 0.919 and RMSE of 9.388% in testing (treatment 6), in addition to a good graphic distribution of residues (Figure 2), which indicates that in the absence of dendrometric variables hard to obtain, for example, height and age, the individual dry aboveground biomass of native bracatinga trees may be estimated without substantial losses in the accuracy of the estimation using only dbh.

In a general manner, treatments tend to overestimate the individual dry aboveground biomass of trees. However, treatments 3 and 6 present a more homogenous residue distribution around the zero axis, confirming the superiority of these treatments in estimating individual dry aboveground biomass of native bracatinga trees of the Metropolitan Region of Curitiba, Paraná.

Urbano (2007) found lower results for the estimation of individual dry aboveground biomass, for this same dataset, and when he used dbh as the input vector for allometric equations, the value obtained for $R^2_{aj.}$ was 0.909. This author found that the best estimate for biomass resulted from the forward method, which selected dbh, dbh^2, dbh^3, dbh^{-1}, ht, hc, hf and dc and resulted in R^2_{aj}. of 0.972.

Other studies analyzed the effect of the composition of input vectors on estimating biomass (Miranda, 2015) and

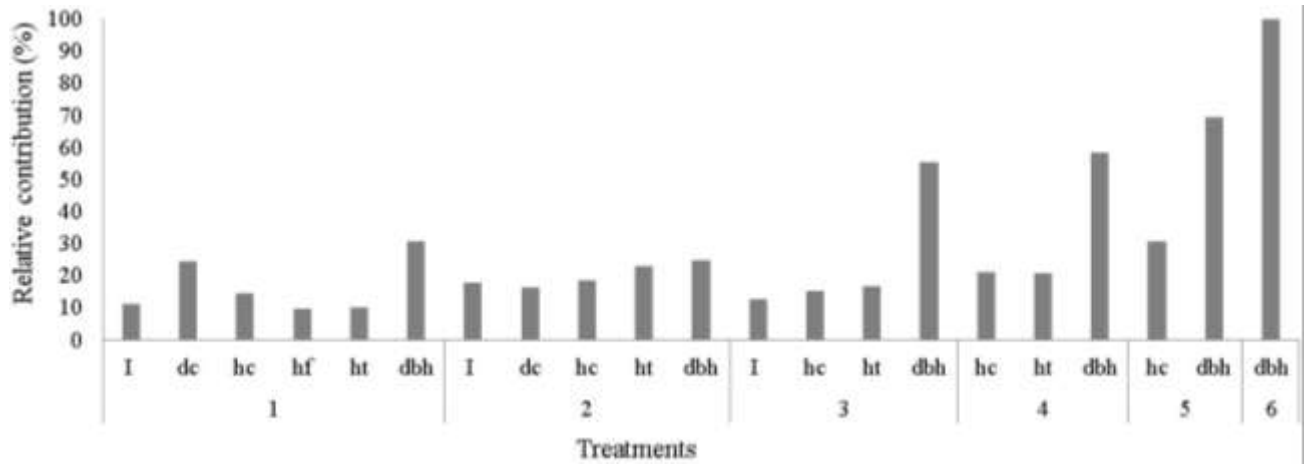

Figure 1. Relative contribution result of the Garson algorithm for the input vectors of MLP for each treatment.

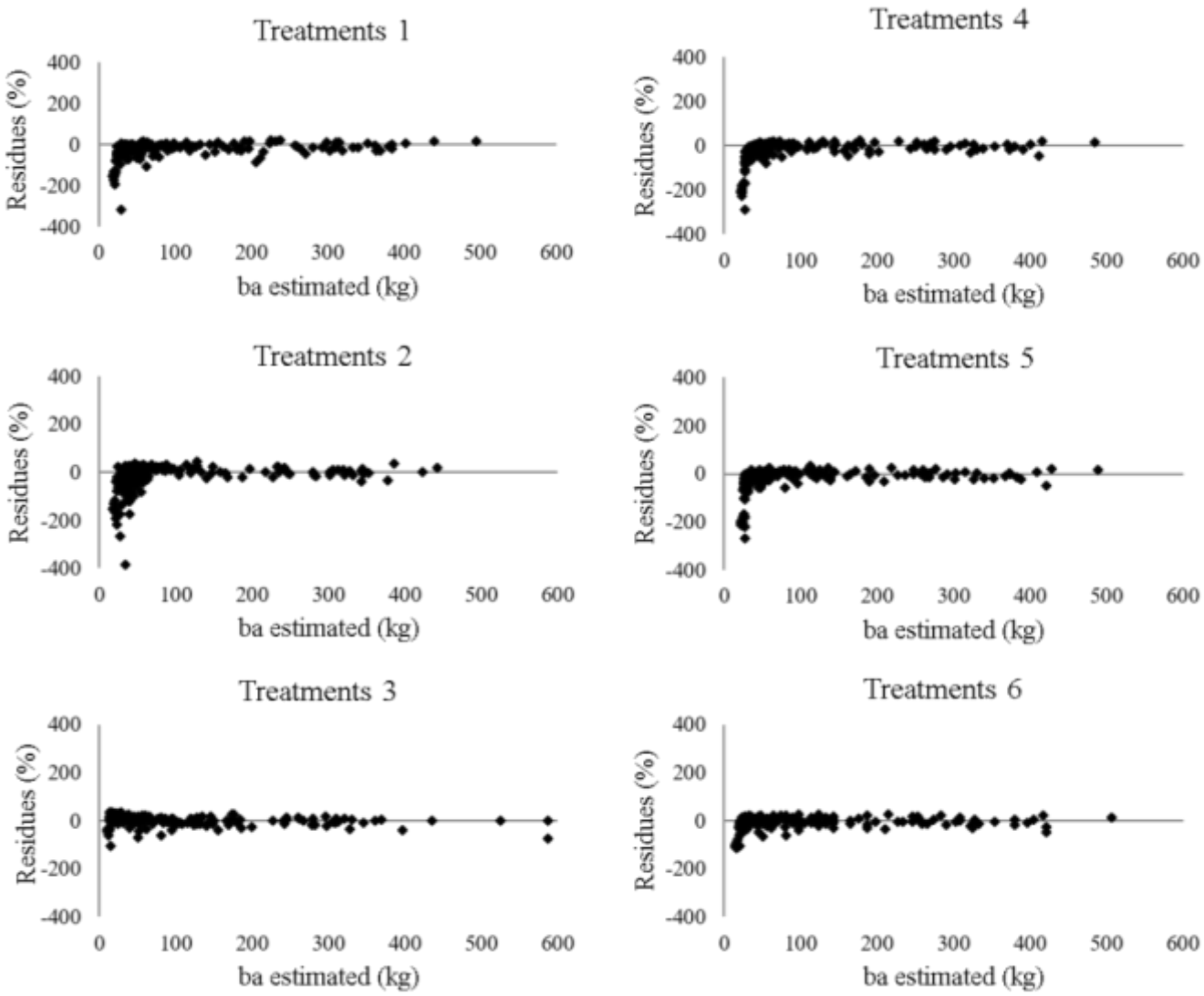

Figure 2. Residue dispersion generated from many treatments for the estimation of individual dry aboveground biomass of *M. scabrella* Benth.

when he utilized a MLP to model the total biomass from a fragment of a Deciduous Seasoned Forest, found R^2 varying from 0.67 to 0.98 when employing input vectors dbh and the ht. Vahedi (2016) used a MLP to estimate the aboveground biomass in northern Iran forests, and used a dbh and ht input layer and found $R^2 = 0.873$ and RMSE of 10.16% in the testing phase.

Other studies that used other machine learning techniques like the study of Sanquetta et al. (2015) that estimated the individual dry aboveground biomass of native trees of Mata Atlântica at Seropédica (RJ), when using variables independent of classification based on instance the dbh, dc, ht, hc, apparent density and basic density, concluded that the use of all variables provided more precise biomass estimates as compared to the reduced number, and the worst performance occurred with the exclusive use of dap.

Conclusion

Composition of input vectors of the MLP that provided the best performance included the variables, age, crown height, total height and the diameter at breast height. The use of only diameter at breast height propitiates consistent estimates when compared with the remaining estimates that used a higher number of input vectors and harder collection, indicating depenence on the desired precision and the resources available; only the dbh propitiates good estimates of individual dry aboveground biomass of native bracatinga trees of the Metropolitan Region of Curitiba, using the referred MLP.

The Garson algorithm presented itself as an interesting tool that assists in the method of attributing selection to determine which input vector have higher relative contribution, providing a better learning of the MLP and selecting the variables essential for the modelling, which may contribute to minimizing costs of forest inventories.

Conflict of Interests

The authors have not declared any conflict of interests.

ACKNOWLEDGEMENT

The authors thank the Coordination for the Improvement of Higher Education Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-CAPES) for their financial support.

REFERENCES

Atkinson PM, Tatnall ARL (1997). Neural networks in remote sensing. Int. J. Remote Sens. 18:699-709.

Binoti DHB, Binoti MLMS, Leite HG (2014). Configuração de redes neurais artificiais para estimação do volume de árvores. Rev. Ciênc. Madeira 5(1):1-6.

Binoti MLMS, Binoti DHB, Leite HG (2013). Aplicação de redes neurais artificiais para estimação da altura de povoamentos equiâneos de eucalipto. Rev. Árvore 37(4):639-645.

Castro RVO, Soares CPB, Martins FB, Leite HG (2013). Crescimento e produção de plantios comerciais de eucalipto estimados por duas categorias de modelos. Pesqui. Agropecu. Bras. 48(3):287-295.

Cordeiro MA, Pereira NNJ, Binoti DHB, Binoti MLMS, Leite HG (2015). Estimativa do volume de *Acacia mangium* utilizando técnicas de redes neurais artificiais e máquinas vetor de suporte. Pesqui. Florest. Bras. 35(83):255-261.

Ferreira JCB, Lafetá BO, Penido TMA, Campos PM, Castro PM (2014). Altura de mudas de Tibouchina granulosa cogn. (melastomataceae) estimada por redes neurais artificiais. Rev. Soc. Bras. Arb. Urbana 9(1):151-160.

Fu L (1994). Neural networks in computer intelligence. New York: McGraw-Hill.

Garson GD (1991). Interpreting neural-network connection weights. Artif. Intell. Expert. 6:47-51.

Gozlan RE, Mastrorillo S, Copp GH, Lek S (1999). Predicting the structure and diversity of young-of-the-year fish assemblages in large rivers. Freshwater Biol. 41:809-820.

Hagan MT, Menhaj M (1994). Training feed-forward networks with the Marquardt algorithm. IEEE T. Neural Networ. 5(6):989-993.

Haykin S (2001). Redes neurais: princípios e prática. Porto Alegre: Bookman P 900.

Heath GE (2010). Training, testing and validating data set in Neural Network. Disponível em: <http://www.mathworks.com/matlabcentral/newsreader/view_thread/295781#917734>. Access: 04 de july 2013.

Kalteh AM (2008). Rainfall-runoff modelling using artificial neural networks (ANNs): modelling and understanding. Caspian J. Environ. Sci. 6(1):53-58.

Martins ER, Binoti MLMS, Leite HG, Binoti DHB, Dutra GC (2016). Configuração de redes neurais artificiais para estimação do afilamento do fuste de árvores de eucalipto. Rev. Bras. Ciênc. Agrár. 11(1):33-38.

Mendonça NP, Carvalho MC, Gomide LR, Ferraz Filho AC, Ferreira MA (2015). Previsão de diâmetros ao longo do fuste de eucalipto via redes neurais artificiais. Rev. Encicl. Biosfera 11(22):2419-2429.

Miranda JFN (2015). Modelos de regressão e de redes neurais artificiais na quantificação de carbono e biomassa lenhosa em floresta estacional decidual no Brasil central. Dissertação (mestrado). Univ. Bras. Bras. 76 p.

Olden JD, Jackson DA (2002). Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154:135-150.

Osborne J, Waters E (2002). Four assumptions of multiple regression that researchers should always test. Pract. Asses. Res. Eval. 8(2):1-8.

Sanquetta CR (2002). Métodos de determinação de biomassa florestal. In. As Florestas e o Carbono. Curitiba pp. 119-140.

Sanquetta CR, Wojciechowski J, Corte APD, Behlinga, Péllico Neto S, Rodrigues AL, Sanquetta MNI (2015). Comparison of data mining and allometric model in estimation of tree biomass. Bioinformatics 16(247):2-9.

Silveira P, Koehler HS, Sanquetta CR, Arce JE (2008). O estado da arte na estimativa de biomassa e carbono em formações florestais. Rev. Flor. 38(1):185-206.

Trautenmüller JW (2015). Quantificação e distribuição do estoque de biomassa acima do solo em floresta estacional decidual. Dissertação (Mestrado) – Universidade Federal de Santa Maria 92 p.

Urbano E (2007). Quantificação e estimativa da biomassa aérea e do carbono fixado em árvores de bracatingais nativos da região metropolitana de Curitiba. Dissertação (Mestrado) – Universidade Federal do Paraná 140 p.

Vahedi AA (2016). Artificial neural network application in comparison with modeling allometric equations for predicting above-ground biomass in the Hyrcanian mixed-beech forest of Iran. Biomass Bioenergy 8(5):66-76.