Nativa, Sinop, v. 9, n. 1, p. p. 54-61, jan./fev. 2021.
Pesquisas Agrárias e Ambientais
DOI: https://doi.org/10.31413/nativa.v9i1.11202 ISSN: 2318-7670

Performance da modelagem para classificação de sítios florestais em bases de dados com outliers

Pábulo Diogo de SOUZA[1]*, Carlos Alberto ARAÚJO JÚNIOR[2], Christian Dias CABACINHA[2], Leandro Silva de OLIVEIRA[2], Celso Dotta LOPES JUNIOR[3], Wellington de ALMEIDA[4]

[1]Programa de Pós-Graduação em Engenharia Florestal, Universidade Federal de Santa Maria, Santa Maria, RS, Brasil.
[2]Instituto de Ciências Agrárias, Universidade Federal de Minas Gerais, Montes Claros, MG, Brasil.
[3]AcelorMittal Bioflorestas, Belo Horizonte, MG, Brasil.
[4]GELF Siderurgia S/A, Itacambira, MG, Brasil.
*E-mail: pabulodiogo@gmail.com
(Orcid: 0000-0002-2446-8041; 0000-0003-0909-8633; 0000-0002-8148-083X; 0000-0003-0800-5001; 0000-0003-4268-4458; 0000-0002-1259-3961)

Recebido em 29/09/2020; Aceito em 11/12/2020; Publicado em 10/02/2021.

RESUMO: As informações utilizadas para estimativa da capacidade produtiva de sítios florestais provêm de bases de dados de inventário florestal que podem conter observações discrepantes (*outliers*). Assim, torna-se necessário a análise de consistência para exclusão destes. Porém, os *outliers* podem representar determinado padrão de crescimento existente na floresta, logo a exclusão destes pode ser uma ação equivocada. Objetivou-se comparar a performance de diferentes técnicas de modelagem para classificação de sítios florestais, considerando uma base de dados com a presença de *outliers*. Utilizou-se pares de dados de idade e altura dominante (HD) de parcelas permanentes de *Eucalyptus urophyla* x *Eucalyptus grandis* localizadas no norte de Minas Gerais. Foi simulado um *outlier* de HD. A base de dados foi modelada, com e sem presença de outliers, por regressão linear (RL) e redes neurais artificiais Multilayer Perceptron (MLP) e Radial Basis Function (RBF). Os métodos foram analisados por meio dos critérios estatísticos de precisão: bias, raiz quadrada do erro médio, correlação de Pearson, erro médio percentual e gráfico de dispersão residual. A MLP foi superior para estimativa do índice de sítio. Portanto, a MLP é indicada para classificação de sítios florestais quando há presença de *outliers* na base de dados.
Palavras-chave: índice de sítio; inventário florestal; dados discrepantes.

Performance of modeling for classification of forest sites in databases with outliers

ABSTRACT: The information used to estimate the productive capacity of forest sites comes from forest inventory databases that may contain discrepant observations (outliers). Thus, consistency analysis is required to exclude these. However, the outliers may represent a certain growth pattern existing in the forest, so their exclusion may be a mistaken action. The objective was to compare the performance of different modeling techniques for forest site classification, considering a database with the presence of outliers. We used pairs of data of age and dominant height (HD) of permanent parcels of Eucalyptus urophila x Eucalyptus grandis located in the north of Minas Gerais. A HD outlier was simulated. The database was modeled, with and without the presence of outliers, by linear regression (RL) and artificial neural networks Multilayer Perceptron (MLP) and Radial Basis Function (RBF). The methods were analyzed by means of precision statistical criteria: bias, square root of mean error, Pearson correlation, mean percentage error and residual scatter plot. The MLP was superior for site index estimation. Therefore, the MLP is indicated for forest site classification when there are outliers in the database.
Keywords: site index; forest inventory; discrepant data.

1. INTRODUÇÃO

O entendimento acerca dos processos de crescimento e produção de determinado povoamento florestal é indispensável ao gerenciamento racional das florestas plantadas (COSENZA et al., 2015). Neste sentido, técnicas de modelagem matemática tornam-se ferramentas importantes para obtenção de informações que possam subsidiar as tomadas de decisão no planejamento florestal. Dentre tais informações, a classificação de sítios, a partir das estimativas de capacidade produtiva da terra, se destaca por sua contribuição na elaboração de estratégias de manejo silvicultural e formulação de estudos relacionados à prognose da produção florestal (CAMPOS; LEITE, 2013).

As informações utilizadas para tais estimativas são oriundas de bases de dados de inventário florestal contínuo (IFC), as quais estão sujeitas à existência de observações discrepantes (*outliers*). Tais dados são aqueles que fogem do padrão geral apresentado pela maioria das medições e podem ser causados tanto pela ocorrência de erros operacionais quanto pela existência de um singular padrão existente na população. Estes dados podem interferir negativamente na modelagem, principalmente quando se aplicam técnicas

clássicas como a regressão linear (GUJARATI; PORTER, 2011).

A simples exclusão destes dados, sem uma profunda verificação de sua origem, torna-se uma ação equivocada e implica em aumento no tempo de processamento e análise dos dados coletados. Isso porque, em se tratando de medições de IFC, tal tarefa tende a ser onerosa para o gestor que, na maioria das vezes, trabalha com extensas bases de dados e não dispõe de tempo suficiente para uma análise minuciosa (ARAÚJO JÚNIOR et al., 2016). Dessa forma, a aplicação de técnicas de modelagem tolerantes à ruídos, ou seja, observações que fogem do padrão dos dados, pode ser uma alternativa que viabiliza a manutenção de *outliers* na base de dados sem que os mesmos prejudiquem as estimativas.

Nesse sentido, redes neurais artificiais (RNA) podem ser entendidas como uma técnica de modelagem matemática com características que podem favorecer o tratamento de dados com presença de *outliers*. Isso se justifica devido ao seu elevado número de conexões entre neurônios artificiais, o que confere, dentre outras características, a tolerância a ruídos (HAYKIN, 2002).

As RNA são modelos computacionais que fazem imitações grosseiras do funcionamento do cérebro humano (OZÇELIK et al., 2013). Dentre as diferentes estruturas de RNA, as redes dos tipos *multilayer perceptron* (MLP) e *radial basis function* (RBF) são frequentemente utilizadas como aproximadores universais de funções, o que as confere vasta aplicação no campo da mensuração florestal. Uma rede neural RBF consiste em três camadas, uma camada de entrada de nós de origem, uma camada oculta que aplica uma transformação não linear no espaço de entrada com um grande número de neurônios e uma camada de saída que é linear e fornece a resposta da rede. Já uma MLP contém uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, que podem ser compostas por diferentes funções de ativação (ARTERO, 2009).

Diversos trabalhos têm sido realizados na área florestal com aplicação de RNA, tais como, estimativa de diâmetros de *Tectona grandis* (LEITE et al., 2009), estimativa de alturas (BINOTI et al., 2013) e estimativa de volume (RIBEIRO et al., 2016). Estudos que abordam a classificação de sítios florestais quando há presença de *outliers* na base de dados podem ser benéficos no âmbito da aplicação de RNA para processamento de dados de mensuração florestal. Além disso podem servir de trabalhos preliminares para a aprimoração do processamento de extensas bases de dados de inventário florestal contínuo.

Diante do exposto, objetivou-se, comparar a performance de diferentes técnicas de modelagem matemática para a classificação de sítios florestais, considerando uma base de dados com a presença de *outliers*.

2. MATERIAL E MÉTODOS

2.1. Dados

Utilizou-se pares de dados de idade (ID) e altura dominante de (HD) de 38 parcelas permanentes medidas em 4 ocasiões, aos 30, 42, 54 e 66 meses, em um maciço florestal de *Eucalyptus urograndis* localizado na região norte do estado de Minas Gerais. Para cada parcela e em cada medição a altura dominante foi obtida conforme o critério de Assmann (1970).

A base de dados foi classificada em função da idade de medição e em seguida foi simulado a presença de um *outlier* para o valor de HD aos 54 meses, a partir da alteração do valor de uma observação de 14,38 m para 7,38 m. Em seguida a base de dados foi submetida à análise de *boxplot*, conforme sugerido por Schneider et al. (2009), para comprovação da presença do dado discrepante. Este procedimento foi realizado com o auxílio *boxplot* contido no pacote *stats* do *software* R (R CORE TEAM, 2014). As bases de dados, com e sem a presença de *outliers*, foram submetidas a modelagem por regressão linear (RL) e por redes neurais artificiais do tipo *multilayer perceptron* (MLP) e *radial basis function* (RBF).

2.2. Processamento por regressão linear

Para estimar a capacidade produtiva a partir da modelagem por regressão linear, utilizou-se modelo proposto por Schumacher (1939) em sua forma linearizada (1):

$$\mathrm{Ln(HD)} = \beta_0 + \beta_1 \frac{1}{\mathrm{ID}} + \varepsilon \qquad (01)$$

em que: *HD* é a altura dominante da parcela (m); *ID* é a idade (anos); *Ln* é o logaritmo neperiano; β_0 e β_1 são os parâmetros a serem estimados pelo modelo de regressão; e ε é o erro aleatório com $\varepsilon \sim NID(0,\sigma^2)$.

As estimativas dos parâmetros para o modelo (1) foram obtidas pelo método dos mínimos quadrados ordinários (MQO) com o auxílio da função *lm* contida na biblioteca *stats* do software R (R CORE TEAM, 2014).

Após a obtenção dos parâmetros do modelo, realizou-se a estimativa do índice de local (S) para cada observação da base de dados. Para isso o modelo (1) foi rearranjado (2), conforme sugerido por Campos e Leite (2013).

$$\mathrm{Ln(S)} = \mathrm{Ln(HD)} - \beta_1 \left[\frac{1}{\mathrm{ID}} - \frac{1}{\mathrm{I_{ref}}}\right] + \varepsilon \qquad (02)$$

em que: S é o índice de sítio (m); I_{ref} é a idade de referência (anos); e os demais já definidos anteriormente.

Para avaliação da qualidade da modelagem por RL com e sem a presença de *outliers* foram avaliados o erro padrão da estimativa, análise gráfica dos resíduos, a significância dos parâmetros e o coeficiente de determinação (R^2), conforme sugerido por Schneider (2009).

2.3. Processamento por redes neurais artificiais

O processamento dos dados por redes neurais artificiais foi realizado no *software Neuroforest* (versão 4.0), em dois momentos, que consistiram, respectivamente, em treinar 30 RNA do tipo MLP e 30 RBF. As redes MLP foram treinadas com algoritmo *Resilient Backpropagation* e função de ativação do tipo sigmoidal, conforme Cordeiro el al. (2015).

Para o treinamento de ambos os tipos de RNA foram considerados, aleatoriamente, 70% dos dados para o treinamento e 30% para validação, com 8 neurônios na camada intermediária. O critério de parada do processamento foi um número de ciclos igual à 5.000 ou erro de 0,001 (BINOTI et al., 2013).

Ao fim do processamento, as RNA's MLP e RBF de melhor desempenho foram armazenadas. Neste caso, a avaliação das redes neurais treinadas foi realizada com auxílio do histograma de resíduos e das estatísticas de *bias* (3), raiz quadrada do erro quadrático médio (*RQME*) (4), Variância (5) e correlação de Pearson ($r_{y,\hat{y}}$) (6), conforme Binoti et al. (2015) e Diamantopoulou et al. (2015), cujas as equações são apresentadas a seguir:

$$\text{Bias} = \frac{1}{n}\sum e \quad (03)$$

$$\text{RQEM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad (04)$$

$$\text{Variância} = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_j - \bar{y}\right)^2 \quad (05)$$

$$r_{y,\hat{y}} = \frac{\sum(y\hat{y}) - \frac{(\sum y)*(\sum \hat{y})}{n}}{\sqrt{\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}*\sqrt{\left[\sum \hat{y}^2 - \frac{(\sum \hat{y})^2}{n}\right]}} \quad (06)$$

em que: *e* é o erro associado a cada estimativa da RNA; *n* é o número de padrões processados pela RNA; y_j é o valor real de saída esperado para o padrão *j*, $\hat{y}_j$ é o valor estimado de saída obtido para o padrão *j* com a aplicação da RNA e $\bar{y}$ a média dos valores reais dos *j* padrões de dados.

Para compor a estrutura de dados de treinamento das RNA's, as informações de IFC foram organizadas de forma que cada medição fosse pareada com a medição realizada no ano posterior, ou seja, IFC_1-IFC_2, IFC_2-IFC_3, ..., IFC_{n-1}-IFC_n e de forma que cada medição fosse pareada com a medição realizada no ano anterior, ou seja, IFC_n-IFC_{n-1}, ..., IFC_3-IFC_2, ..., IFC_2-IFC_1.

As redes neurais foram treinadas com três variáveis de entrada: Idade 1 (ID_1), Idade 2 (ID_2) e Altura dominante 1 (HD_1). A variável de saída foi a Altura dominante 2 (HD_2) para cada parcela de IFC. Após o treinamento, a RNA armazenada foi aplicada a uma base dados contendo as medições de cada parcela, obtendo-se assim, a altura dominante na idade índice (72 meses), consequentemente, o índice de sítio (S).

2.4. Avaliação da qualidade das técnicas de modelagem

As estimativas de índice de sítio obtidas a partir da modelagem por RL, RQ e pelas redes MLP e RBF foram classificadas a partir da Equação (7), que retorna o valor central da classe com base em uma amplitude pré-estabelecida.

$$\text{Classe} = \text{int}\left(\frac{S}{a}\right)a + \left(\frac{1}{2}\,a\right) \quad (07)$$

em que: *int*(.) é a função que retorna o valor inteiro de um número real, *S* é o índice de sítio estimado (em metros), *a* é a amplitude de cada classe de sítio considerada.

Os resultados da classificação da capacidade produtiva de sítios florestais foram submetidos à análise de estabilidade, conforme sugerido por Scolforo (2006) e Chaves et al. (2016). Neste caso, considerou-se três intervalos de últimas medições, ou seja: (1) IFC1 a IFC4; (2) IFC2 a IFC4 e (3) IFC3 a IFC4.

Para avaliação da precisão dos resultados obtidos em cada técnica de modelagem, foram realizadas estimativas da altura dominante (HD) para a última medição em cada parcela de IFC. Em seguida, os resultados foram comparados pelo bias, RQEM, coeficiente de correlação de Pearson, erro médio percentual (8), e gráfico de dispersão do resíduo em função da idade de medição de IFC.

$$e = \sum\left(\frac{y_j - \hat{y}_j}{y_j} * 100\right) \quad (08)$$

em que: e é o erro médio percentual, y_j é a altura média dominante real e $\hat{y}_j$ é a altura média dominante observada.

3. RESULTADOS

A classificação dos dados do inventário florestal resultou em quatro classes de idade (30, 42, 54 e 66 meses). Com a aplicação da análise por *boxplot* em cada uma das classes observou-se a presença de *outliers* na base de dados, de modo que um destes resultou da alteração do valor de HD de 14,38 para 7,38 m aos 54 meses de idade (Figura 1).

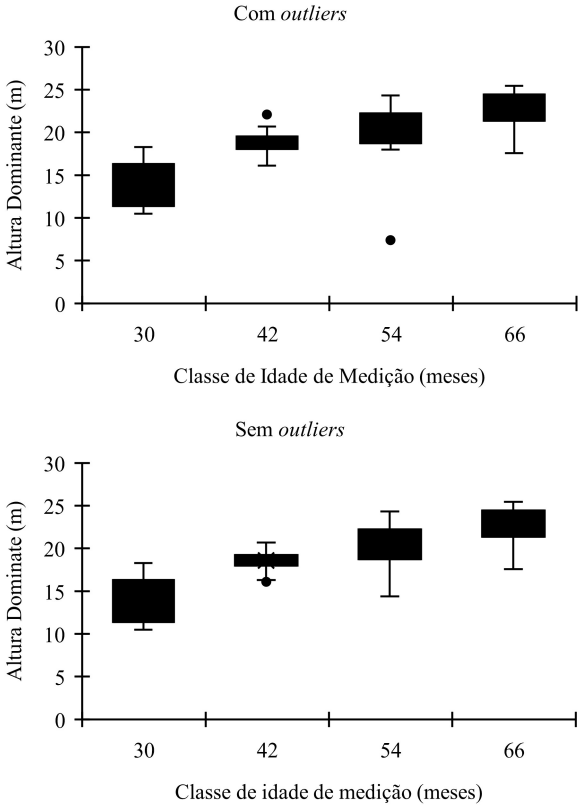


Figura 1. Análise de boxplot aplicada a base de dados com e sem a presença de *outliers*.
Figure 2. Boxplot analysis applied to the database with and without the presence of outliers.

A estatística descritiva para o conjunto de dados com e sem *outliers* mostrou-se próxima, pois os valores para médias e medianas de altura dominante por classe de idade pouco se alteraram. Após o pré-processamento dos dados observaram-se apenas ligeiras alterações nos valores de média e mediana de HD das classes onde procedeu-se a exclusão de *outliers*, sobretudo na classe de 54 meses (Tabela 1).

Tabela 1. Média e mediana das alturas dominantes em cada classe de idade.
Table 1. Mean and median of dominant heights in each age class.

Classe de idade	Sem *outliers*		Com *outliers*	
(meses)	Média	Mediana	Média	Mediana
30	13,78	13,50	13,78	13,50
42	18,81	18,98	18,79	18,93
54	20,42	20,59	20,24	20,62
66	22,73	22,58	22,73	22,58

A modelagem por RL resultou em parâmetros significativos (estatisticamente diferentes de zero pelo teste t), para ambas as bases de dados. No entanto, observou-se que a presença de *outliers* interferiu na qualidade do ajuste.

Pois houve alteração na dispersão residual (Figura 2), diminuição do valor de R^2 e aumento do erro padrão em comparação com os resultados obtidos após a retirada dos mesmos (Tabela 2).

Tabela 2. Estimativas dos parâmetros e estatísticas das equações ajustadas por regressão linear considerando a base de dados com e sem *outliers*.
Table 2. Estimates of the parameters and statistics of the equations adjusted by linear regression considering the database with and without outliers.

Base de dados	Parâmetros		Erro	R^2	Erro padrão
Com *outliers*	β0	3,5949***	0,0355	0,73	0,127
	β1	-30,727***	1,5325		
Sem *outliers*	β0	3,6166***	0,0261	0,85	0,0932
	β1	-31,376***	1,1238		

Legenda: *** significância de 1% e R^2 = coeficiente de determinação.

Foi observada uma leve alteração no comportamento da curva gerada pelo ajuste por RL, pois, na presença do *outlier* na idade de 54 meses, tal curva se deslocou ligeiramente para baixo (Figura 3).

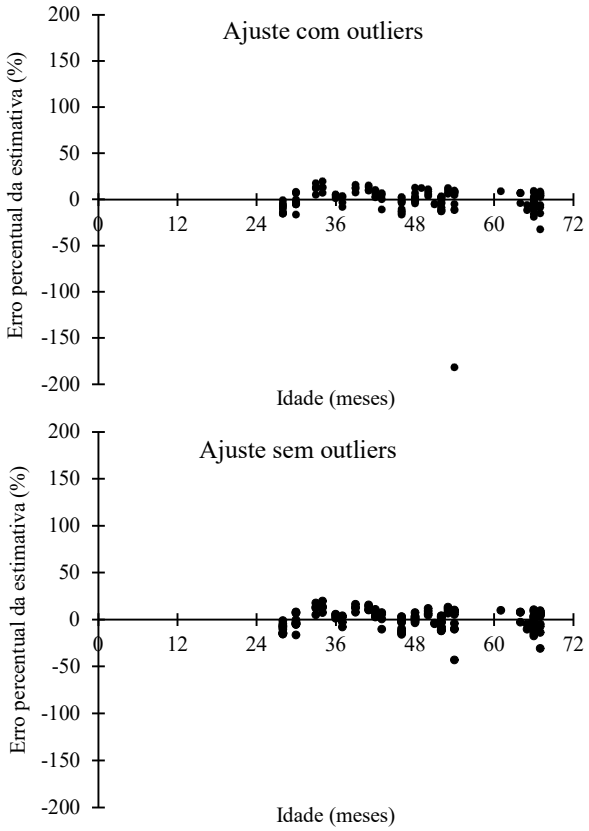


Figura 3. Dispersão residual para o ajuste do modelo de regressão linear com e sem a presença de outliers na base de dados.
Figure 3. Residual dispersion for the adjustment of the linear regression model with and without the presence of outliers in the database.

Para ambas as bases de dados as etapas de treinamento e validação das redes MLP e RBF com e sem *outliers* produziram resultados satisfatórios, uma vez que foram obtidos critérios de qualidade que indicaram boa capacidade de generalização, como por exemplo baixos valores de bias nas etapas de treino e validação (Tabela 3).

Através dos histogramas de distribuição dos resíduos é possível observar que os erros gerados pelas estimativas das RNA's MLP e RBF (Figuras 4 e 5), tendem a ser normalmente distribuídos. Contudo, observou-se que, na presença de *outliers*, a RBF apresentou subestimativas que ocasionaram erros em torno de 40 %, que não foram observados após o pré-processamento.

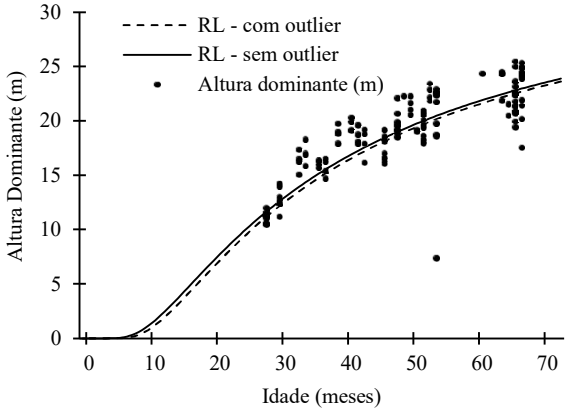


Figura 3. Tendência de crescimento em altura dominante, conforme o ajuste de equações de regressão linear (RL) para a base de dados com e sem presença de *outliers*.
Figure 3. Growth trend in dominant height, according to the adjustment of linear regression equations (RL) for the database with and without outliers.

Para a análise de precisão das estimativas da altura dominante geradas pelos três métodos de modelagem, observou-se que a rede MLP apresentou resultados ligeiramente melhores para os critérios estatísticos de precisão analisados (Tabela 4). O somatório dos escores confirma a superioridade da modelagem por MLP na estimativa do índice de sítio (Tabela 5).

Quanto ao gráfico de distribuição dos resíduos, na presença de *outliers* a rede MLP mostrou-se ligeiramente superior às demais técnicas de modelagem. Já após o pré-processamento praticamente não houve diferença entre a dispersão dos resíduos para redes MLP e RBF (Figura 6). Deste modo, os resíduos obtidos pelas estimativas feitas com a MLP ficaram melhor distribuídos independentemente da presença de *outliers* no processamento.

As estimativas de índice de sítio obtidas através da modelagem por redes MLP mostraram-se superior às demais técnicas, independentemente do número de últimas medições de IFC consideradas e da presença ou ausência de *outliers* na base de dados (Figura 7).

4. DISCUSSÃO

O comportamento da curva média para estimar a altura dominante com a modelagem por RL com e sem a presença de *outliers* apresentou tendências semelhantes. No entanto, observou-se uma ligeira alteração no valor do índice de sítio médio estimado, sendo inicialmente igual a 23,76 m, passando a ser de 24,10 m após o pré-processamento (retirada dos dados discrepantes). Araújo Júnior et al. (2016), ao compararem a estimativa do índice de sítio com RL, evidenciaram interferência negativa na estimativa quando há presença de *outliers*. De acordo com Schneider et. al (2009), a presença de *outliers* em bases de dados submetidas ao ajuste de modelos de regressão pode comprometer o pressuposto de normalidade dos resíduos, sobretudo em função do viés na estimativa dos coeficientes.

Tabela 3. Treino e validação das RNA's obtidas para estimativa da altura dominante na idade 2 (HD_2) na base de dados com e sem a presença de *outliers*.
Table 3. Training and validation of the RNA's obtained to estimate the dominant height at age 2 (HD2) in the database with and without the presence of outliers.

RNA	Base de dados	Tipo de Dados	Bias	RQME	Variância	$r_{y,\hat{y}}$
MLP	Com *outliers*	Treino	-0,07618	1,30045	1,71015	0,92897
		Validação	0,01457	1,30190	1,70499	0,92556
	Sem *outliers*	Treino	-0,02342	0,90836	0,83025	0,96118
		Validação	-0,03341	1,27451	1,64787	0,90635
RBF	Com *outliers*	Treino	0,00719	0,97586	0,95823	0,95285
		Validação	0,01039	1,50817	2,30522	0,92515
	Sem *outliers*	Treino	0,01264	0,92451	0,86037	0,95935
		Validação	0,04612	1,10680	1,24199	0,93205

Legenda: RNA = rede neural artificial, MLP = *multilayer perceptron*, RBF = *radial basis function*, RQME = raiz quadrada do erro médio; $r_{y,\hat{y}}$ = correlação de Pearson.

Tabela 4. Análise da precisão das estimativas de altura das dominantes obtidas com aplicação das técnicas modelagens ajustadas a base de dados com e sem *outliers*.
Table 4. Analysis of the accuracy of the height estimates of the dominants obtained with the application of modeling techniques adjusted to the database with and without outliers.

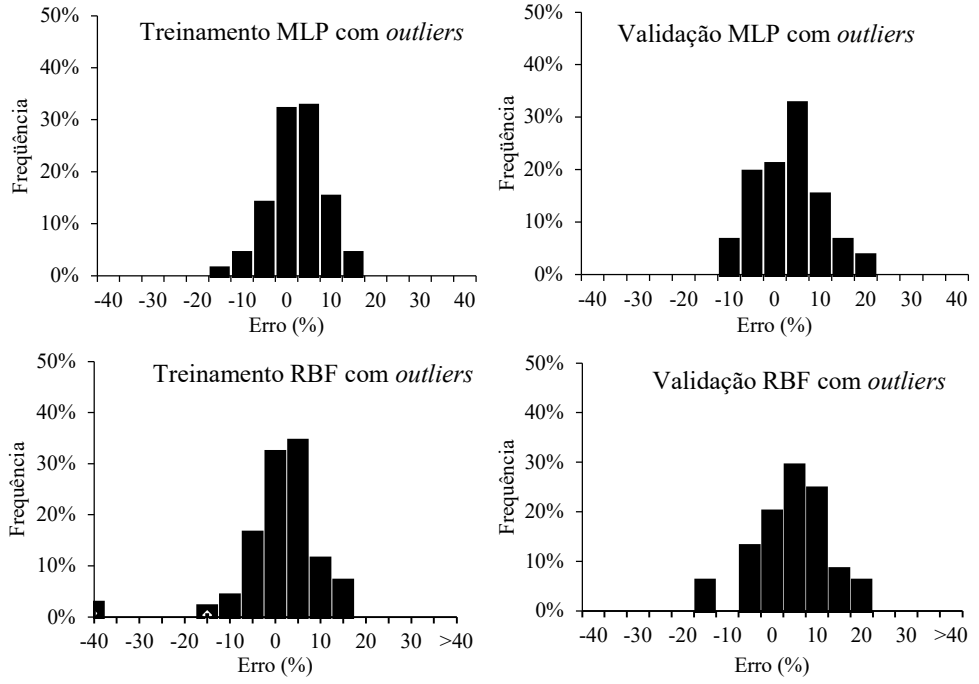
Tipo de modelagem	Com *outliers*				Sem *outliers*			
	Bias	RQEM	$r_{y,\hat{y}}$	e %	Bias	RQEM	$r_{y,\hat{y}}$	e %
RL	-0,3944	22,709	0,59	-1,52	-0,6045	21,316	0,40	-3,03
MLP	0,0013	14,734	0,65	-1,40	0,0211	14,799	0,61	-0,05
RBF	0,0144	29,217	0,44	2,35	0,0293	17,662	0,46	-0,15

Legenda: RL = regressão linear, MLP = *multilayer perceptron*, RBF = *radial basis function.*

Tabela 5. Avaliação de *scores* para os critérios estatísticos analisados para as técnicas de estimativa do índice de sítio com e sem a presença de *outliers*.
Tabela 5. Evaluation of scores for the statistical criteria analyzed for the techniques of estimating the site index with and without the presence of outliers.

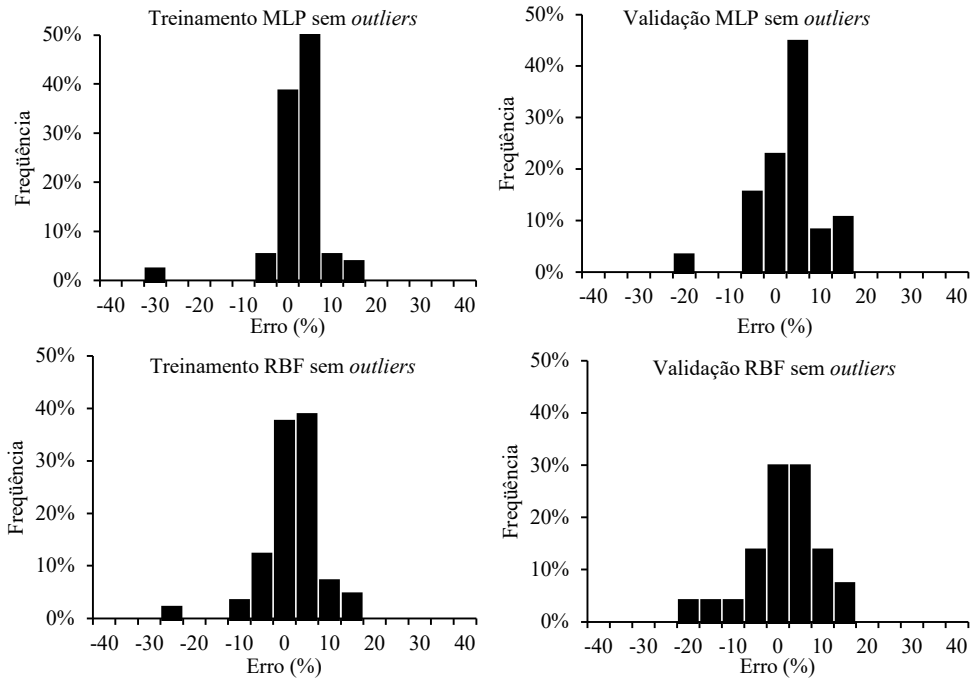
Tipo de modelagem	Com *outliers*					Sem *outliers*				
	Bias	RQEM	$r_{y,\hat{y}}$	e %	Total	Bias	RQEM	$r_{y,\hat{y}}$	e %	Total
RL	3	3	2	2	10	3	3	3	3	12
MLP	1	1	1	1	4	1	1	1	1	4
RBF	2	2	3	3	10	2	2	2	2	8

Legenda: RL = regressão linear, MLP = *multilayer perceptron*, RBF = *radial basis function.*



Legenda: MLP = *multilayer perceptron*; RBF = *radial basis function*.

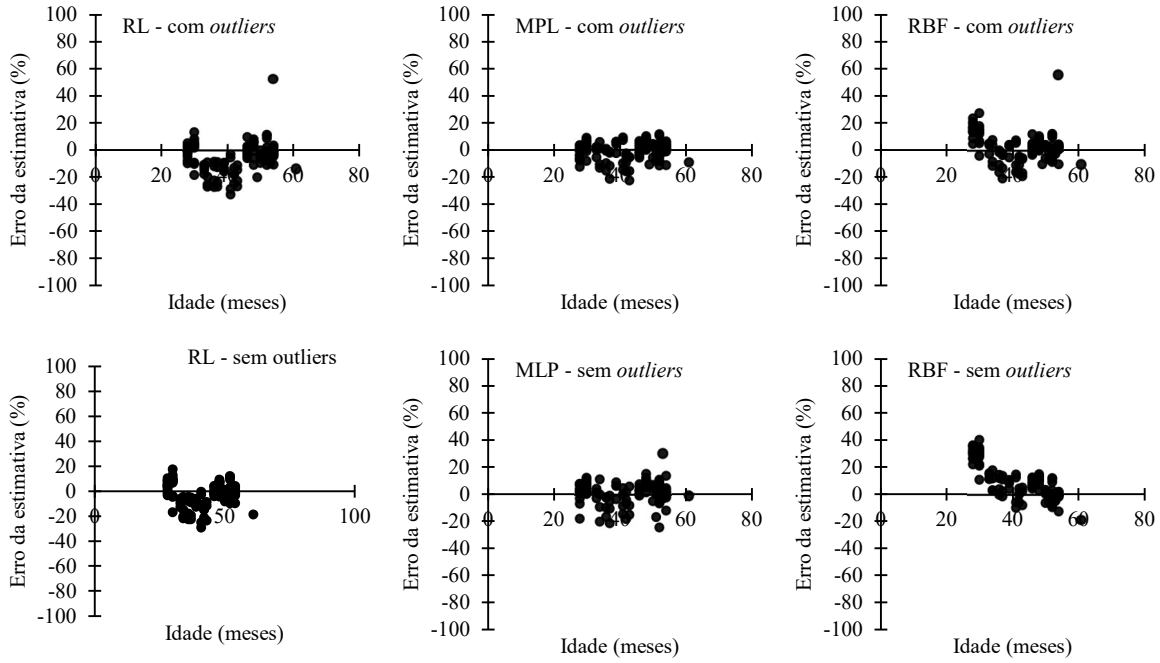
Figura 4. Histograma de frequência percentual dos erros a partir das estimativas feitas pelas redes MLP e RBF treinadas com a base de dados com *outliers*.
Figure 4. Histogram of percentage frequency of errors based on estimates made by MLP and RBF networks trained with the database with outliers.



Legenda: MLP = *multilayer perceptron*; RBF = *radial basis function*.

Figura 5. Histograma de frequência percentual dos erros a partir das estimativas feitas pelas redes MLP e RBF treinadas com a base de dados sem *outliers*.

Figure 5. Histogram of percentage frequency of errors based on estimates made by MLP and RBF networks trained with the database without outliers.



Legenda: RL = regressão linear, MLP = *multilayer perceptron*, RBF = *radial basis function.*

Figura 6. Dispersão percentual dos erros gerados pela estimativa da modelagem por RL, MLP e RBF aplicada a base de dados com e sem a presença de *outliers*.

Figure 6. Percentage dispersion of errors generated by the estimation of the modeling by RL, MLP and RBF applied to the database with and without the presence of outliers.

Neste trabalho apesar da presença de poucos *outliers* em uma base de dados não extensa, a modelagem por regressão mostrou-se sensível aos dados discrepantes. Pois os critérios de qualidade de ajuste da mesma foram melhores após o pré-processamento. Em relação à análise de precisão das técnicas de modelagem, pôde-se perceber que na presença de *outliers* a modelagem por RL tendeu a subestimar as estimativas de altura dominante em idades mais avançadas. Observou-se ainda que, sujeita às mesmas condições, a modelagem por redes neurais artificiais do tipo RBF apresentou tendência de erro de superestimativa para as idades mais jovens.

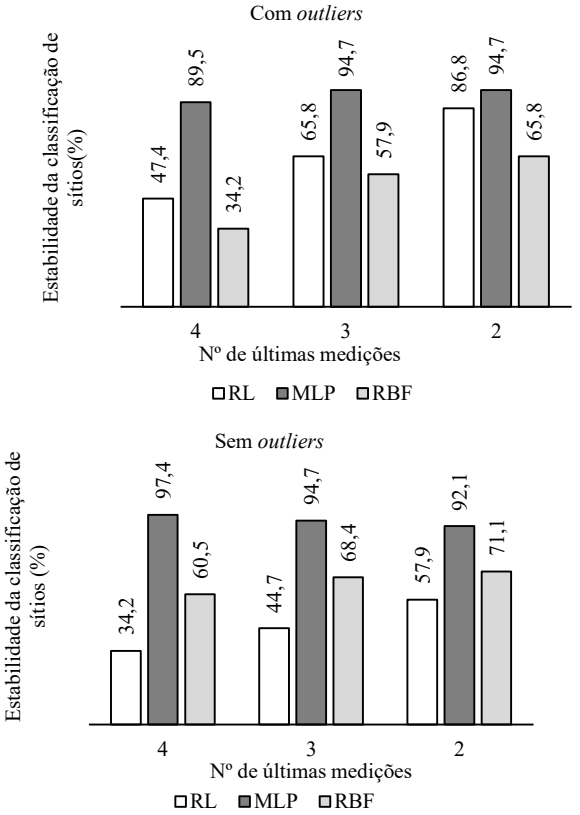


Figura 7. Estabilidade da classificação de sítios obtida pela modelagem por mínimo quadrados ordinários (RL), *Multilayer Perceptron* (MLP) e *Radial Basic Function* (RBF).
Figure 7. Stability of the classification of sites obtained by modeling by least ordinary squares (RL), Multilayer Perceptron (MLP) and Radial Basic Function (RBF).

Já a ligeira superioridade nas estatísticas de precisão obtidas com as redes MLP, inclusive com a presença de *outliers*, certamente estão associadas ao fato destas serem consideradas aproximadores universais de funções, resultante de seu poder computacional, que às conferem a aproximação de qualquer função matemática linear ou não linear (BRAGA et al., 2014).

Binoti et al. (2013) corroboram com esses resultados ao obterem boas estimativas de altura com a aplicação de redes MLP em dados de povoamentos equiâneos de eucalipto, sendo inclusive superiores a resultados obtidos com a equações hipsométricas. Vendruscolo et al. (2017) também obtiveram com a aplicação de redes neurais artificiais estimativas de altura em povoamentos de teca mais precisas que a modelagem realizada com regressão.

A performance da estabilidade da classificação obtida com as modelagens foi pertinente com os resultados obtidos nas análises de precisão da estimativa de HD, visto que, em ambas as avaliações a modelagem com MLP mostrou-se superior às demais. Ercanli et al. (2018), ao predizer o índice de área foliar em florestas de Pinus com diferentes técnicas de modelagem, dentre elas as redes MLP e RBF, também obtiveram melhores resultados com a primeira (ERCANLI; GÜNLÜ; KELEŞ, 2018).

Binoti et al. (2014) ao terem obtido bons resultados no treinamento e validação de redes neurais para estimativa de volume com e sem casca em povoamentos de eucalipto no Sul da Bahia, corroboram com os resultados de performance do treinamento das RNA's obtidos no presente trabalho. De acordo com Braga et al. (2014), a capacidade de generalização das RNA's é indispensável para o sucesso da modelagem, pois garante a precisão na resolução dos problemas aos quais o sistema é submetido.

Ademais, considerando as três técnicas de medição, a estabilidade foi maior quando foram consideradas apenas as 2 últimas medições, o que indica instabilidade na fase inicial de crescimento. Resultados semelhantes também foram observados por Chaves et al. (2016), ao estimar o índice de sítio em um povoamento de *Tectona grandis*. Machado et al. (1997), ressaltam que tal comportamento quanto a estabilidade da classificação de sítios em decorrência do fato de maciços florestais apresentarem padrões de crescimento em altura mais definidos a medida que se aproximam da idade técnica de corte.

A performance superior das redes MLP, tanto na estabilidade da classificação de sítio quanto na precisão de suas estimativas, evidenciam sua capacidade de generalização associada à sua tolerância a ruídos, ou seja, capacidade de ter resultados precisos na modelagem devido a captação do padrão de dados, independentemente de *outliers*. Silva et al. (2009) e Araújo Júnior et al. (2019) corroboram com os resultados alcançados, ao obterem melhores resultados em modelagens com redes MLP comparados com modelos clássicos de regressão.

Dentre as possibilidades de ocorrência de dados discrepantes em uma base de dados de inventário florestal estão os erros operacionais de campo e escritório. Em contra partida, os *outliers* podem não se tratar de padrões de crescimento decorrentes da interferência de fenômenos bióticos ou abióticos e, portanto, podem descrever o crescimento florestal em determinado sítio.

Neste sentido resultados obtidos com a aplicação de técnicas de modelagem robustas à presença de *outliers*, como RNA, demostram sua utilidade para o processamento de bases de dados de inventário florestal contínuo. Pois tais técnicas permitem a possibilidade de dispensar o pré-processamento em conjuntos de dados para classificação de sítios, e consequentemente a exclusão de dados que não se tem clareza quanto a natureza ou causa destes. No entanto, sobre tudo é crucial que seja realizada uma análise minuciosa dos critérios estatísticos de precisão para as etapas de treino e validação, pois é indispensável que a RNA apresente uma boa generalização.

Apesar deste estudo ter sido estruturado com a simulação de *outlier*, o mesmo pode preliminarmente contribuir para a melhoria da eficiência de processamento em dados de inventário florestal contínuo, as quais estão sujeitas a frequente ocorrência de erros de medição.

5. CONCLUSÕES

A modelagem por redes MPL apresentou a melhor performance estatística para classificação de sítios florestais em bases de dados com presença de *outliers*. Portanto, o uso de RNA do tipo MLP é uma alternativa interessante para modelar bases de dados de inventário florestal com dados discrepantes que não se tem clareza acerca das causas.

6. REFERÊNCIAS

ARAÚJO JÚNIOR, C. A.; SOARES, C. P. B.; LEITE H. G. Regressão quantílica para gerar curvas de índice de sítio em povoamentos de eucalipto no Brasil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 6, p. 720-727,

2016. DOI: http://dx.doi.org/10.1590/S0100-204X2016000600003

ARAÚJO JÚNIOR, C. A.; SOUZA, P. D.; ASSIS, A. L.; CABACINHA, C. D.; LEITE, H. G.; SOARES, C. P. B.; SILVA, A. A. L.; CASTRO R. V. O. Artificial neural networks, quantile regression, and linear regression for site index prediction in the presence of outliers. **Pesquisa agropecuária Brasileira**, Brasília, v. 54, n. 1, p. 1-8, 2019. DOI: 10.1590/S1678-3921.pab2019.v54.00078

ARTERO, A. O. **Inteligência Artificial:** teoria e prática. 1 ed. São Paulo: Livraria da Física, 2009. 230p.

ASSMAN E. **The principles of forest yield study**. 5.ed. Oxford: Pergamon Press, 1970. 506p.

BINOTI, M. L. M. S.; BINOTI, D. H. B.; LEITE, H. G. Aplicação de redes neurais artificiais para estimação da altura de povoamentos equiâneos de eucalipto. **Revista Árvore**, Viçosa, v. 37, n. 4, p. 639-645, 2013. DOI: http://dx.doi.org/10.1590/S0100-67622013000400007

BINOTI, M. L. M. S.; LEITE, H. G.; BINOTI, D. H. B.; GLERIANE, J. M. Prognose Em Nível De Povoamento de clones de eucalipto empregando redes neurais artificiais. **Cerne**, Lavras, v. 21, n. 1, p. 97-105, 2015. DOI: https://doi.org/10.1590/01047760201521011153

BRAGA, A. P.; CARVALHO, A. P. L. P.; LUDEMIR, T. B. **Redes Neurais Artificiais:** Teoria e Aplicações. 2 ed. Rio de Janeiro: LTC, 2014. 226p.

CAMPOS J. C. C.; LEITE H. G. **Mensuração florestal: perguntas e respostas**. 4 ed. Viçosa: UFV, 2013. 605p.

CHAVES, A. G. S.; DRESHER, R.; CALDEIRA, S. F.; MARTINEZ, D. T.; VENDRUSCOLO, D. G. S. Capacidade produtiva de *Tectona grandis* L. f. no Sudoeste de Mato Grosso. **Scientia Florestalis**, Piracicaba, v. 44 n. 110, p. 415-416, jun. 2016. DOI: http://dx.doi.org/10.18671/scifor.v44n110.14

COSENZA, D. N.; LEITE, H. G.; MARCATTI, E. G.; BINOTI, D. H. B.; ALCÂNTARA, A. E. M.; RODE R. Classificação da capacidade produtiva de sítios florestais utilizando máquina de vetor de suporte e rede neural artificial. **Scientia Forestalis**, Piracicaba, v. 43, n. 108, p. 955-963, dez. 2015. DOI: http://dx.doi.org/10.18671/scifor.v43n108.19

CORDEIRO, M. A.; PEREIRA, N. N. J.; BINOTI, D. H. B.; BINOTI, M. L. M. S. Estimativa do volume de *Acacia mangium* utilizando técnicas de redes neurais artificiais e máquinas vetor de suporte. **Pesquisa Florestal Brasileira**, Colombo, v. 35, n. 83, p. 255-261, 2015. DOI: 10.4336/2015.pfb.35.83.596

DIAMANTOPOULOU, M. J.; ÖZÇELIK, R.; CRECENTE-CAMPO, F.; ELER, U. Estimation of Weibull function parameters for modelling tree diameter distribution using least squares and artificial neural networks methods. **Biosystems Engineering**, Northumberland, v. 133, p. 33-45, mar. 2015. DOI: http://dx.doi.org/10.1016/j.biosystemseng.2015.02.013

ERCANLI, İ.; GÜNLÜ, A.; KELEŞ, S. Artificial neural network models predicting the leaf area index: a case study in pure even-aged Crimean pine forests from Turkey. **Forests Ecosystems**, v. 5, n. 29, p. 2-12, 2018. DOI: https://doi.org/10.1186/s40663-018-0149-8.

HAYKIN, S. **Redes Neurais:** princípios e prática. 2.ed. Porto Alegre: Bookman, 2002. 900p.

GUJARATI, D. N.; PORTER, D. C. **Basic Econometric**. 5 ed. New York: AMGH, 2011. 924p.

LEITE, H. G.; SILVA, M. L. M.; BINOTI, D. H. B.; FARDIN, L.; TAKIZAWA, F. H. Estimation of inside-bark diameter and heartwood diameter for *Tectona grandis* Linn. trees using artificial neural networks. **European Journal of Forest Research**, München, v. 130, n. 2, p. 263-269, ago. 2010, DOI: http://dx.doi.org/10.1007/s10342-010-0427-7

MACHADO, S. A.; OLIVEIRA, E. B.; CARPANEZZI, A. A.; BARTOSZECK, A. C. P. S. Classificação de sítio para bracatingais na região metropolitana de Curitiba. **Boletim de Pesquisa Florestal**, Colombo, v. 35, n. 1, p. 21-37, jul./dez. 1997.

ÖZÇELIK, R.; DIAMANTOPOULOU, M. J.; CRECENTE-CAMPO, F.; ELER, U. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. **Forest Ecology and Management**, v. 306, p. 52-60, 2013. DOI: https://doi.org/10.1016/j.foreco.2013.06.009

R Core Team. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2014. Disponível em: <http://www.R-project.org>. Acesso em: 10 jul. 2019.

RIBEIRO, R. B. S.; GAMA, J. R. V.; SOUZA, A. L.; LEITE, H. G.; SOARES, C. P.; SILVA, G. F. Métodos para estimar o volume de fustes e galhos na floresta nacional do Tapajós. **Revista Árvore**, Viçosa, v. 40, n. 1, p. 81-88, jan./fev. 2016. DOI: https://doi.org/10.1590/0100-67622016000100009

SCHNEIDER, P. R.; SCHNEIDER, P. S. P.; SOUZA C. A. M. **Análise de regressão aplicada à engenharia florestal**. Santa Maria: FACOS-UFSM, 2009. 294p.

SCOLFORO, J. R. S. **Biometria Florestal** – Modelos de Crescimento e Produção. Lavras: UFLA-Fundação de apoio ao Ensino, Pesquisa e Extensão, 2006. 355p.

SILVA, M. L. M.; BINOTI, D. H. B.; GLERIANI, J. M. Ajuste do modelo de Schumacher e Hall e aplicação de redes neurais artificiais para estimar volume de árvores de eucalipto. **Revista Árvore**, Viçosa, v. 33, n. 6, p. 1133-1139, 2009. DOI: http://dx.doi.org/10.1590/S0100-67622009000600015

VENDRUSCOLO, D. G. S.; CHAVES, A. G. S.; MEDEIROS, R. A.; SILVA, R. S.; SOUZA, H. S.; DRESCHER, R.; LEITE, H. G. Estimativa da altura de árvores de *Tectona grandis* L.f. utilizando regressão e redes neurais artificiais. **Nativa**, Sinop, v. 5, n. 1, p. 52-58, jan./fev. 2017. DOI: http://dx.doi.org/10.5935/2318-7670.v05n01a09