

Configuração de algoritmos de aprendizado de máquina na modelagem florestal: um estudo de caso na modelagem da relação hipsométrica

Tunning machine learning algorithms for forestry modeling: a case study in the height-diameter relationship

Sérgio Vinícius Serejo da Costa Filho^I, Julio Eduardo Arce^{II},
Razer Nizer Rojas Montaña^{III}, Allan Libanio Pelissari^{II}

Resumo

No presente estudo foram aplicados quatro algoritmos de aprendizado de máquina na tarefa de modelagem da relação hipsométrica de povoamentos de *Pinus taeda* L. em diferentes idades. Centenas de combinações de parâmetros foram testadas para os algoritmos k-vizinhos mais próximos, florestas aleatórias, máquinas de vetores de suporte e redes neurais artificiais. Para seleção do melhor modelo para cada algoritmo, utilizou-se o método de busca em grade combinada ao método de validação cruzada *k-fold*. Os modelos selecionados foram utilizados para predição da altura total de indivíduos pertencentes a um conjunto de dados independente, e os resultados foram comparados aos obtidos por modelos de regressão linear. Os modelos de aprendizado de máquina apresentaram indicadores estatísticos similares aos modelos de regressão linear, no entanto, tiveram dispersão de resíduos menos tendenciosa, principalmente na análise estratificada por povoamento. A máquina de vetores de suporte e a rede neural artificial foram os modelos mais satisfatórios em precisão e dispersão dos resíduos.

Palavras-chave: Inteligência artificial; Busca em grade; Redes neurais artificiais; Validação cruzada

Abstract

In the present study, four machine learning algorithms were applied in the task of modeling the height-diameter relationship of *Pinus taeda* L. stands at different ages. Hundreds of parameter combinations were tested for the k-nearest neighbors, random forests, support vector machines, and artificial neural networks algorithms. In order to select the best model for each algorithm, the grid search and the k-fold cross validation methods were applied. The selected models were used to predict the total height of individuals in an independent data set, and the results were compared to those obtained by linear regression models. The machine learning models presented similar statistical indicators to the linear regression models. However, they had less biased dispersion of residues, especially in the stratified analysis by age. The support vector machine and the artificial neural network were the most satisfactory models in precision and dispersion of residues.

Keywords: Artificial intelligence; Grid search; Artificial neural networks; Cross validation

Introdução

Na mensuração florestal, os modelos estatísticos são amplamente utilizados para a predição ou explicação de fenômenos físicos. As observações de um fenômeno geralmente são consideradas representáveis por uma função probabilística, na qual uma ou mais variáveis explicativas são empregadas, visando prever o comportamento de uma variável dependente (ROBINSON *et al.*, 2014). Com o avanço dos métodos computacionais, uma nova categoria

^I Engenheiro Florestal, MSc., Pesquisador Autônomo, Universidade Federal do Paraná, Setor de Ciências Agrárias, Universidade Federal do Paraná, Av. Prof. Lothário Meissner, 632, Jardim Botânico, CEP 80210-170, Curitiba (PR), Brasil. sergio.vscf@gmail.com (ORCID: 0000-0001-5432-317X)

^{II} Engenheiro Florestal, Dr., Professor do Departamento de Ciências Florestais, Setor de Ciências Agrárias, Universidade Federal do Paraná, Av. Prof. Lothário Meissner, 632, Jardim Botânico, CEP 80210-170, Curitiba (PR), Brasil. jarce@ufpr.br (ORCID: 0000-0002-4777-2310) / allanpelissari@gmail.com (ORCID: 0000-0002-0915-0238)

^{III} Cientista da Computação, Dr., Professor do Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas, CEP 81520-260, Curitiba (PR), Brasil. razer.anthom@gmail.com (ORCID: 0000-0003-1589-7124)



de modelos, conhecida como modelos ou algoritmos de aprendizado de máquina, tem sido objeto de estudos para modelagem de atributos florestais (BINOTI *et al.*, 2013; GORGENS *et al.*, 2014; CORDEIRO *et al.*, 2015; VENDRUSCOLO *et al.*, 2016; VENDRUSCOLO *et al.*, 2017; MENDONÇA, 2018).

O aprendizado de máquina (AM) é uma área da inteligência artificial voltada à construção de sistemas capazes de induzir hipóteses ou aproximar funções a partir da experiência acumulada em problemas anteriores (FACELI *et al.*, 2011). As decisões tomadas por algoritmos de AM são baseadas no aprendizado indutivo, que pode ser subdividido em supervisionado, quando o objetivo é a resolução de problemas de classificação ou de regressão, e não supervisionado, quando a tarefa é de agrupamento ou associação (MONARD e BARANAUSKAS, 2003).

Estudos na área de mensuração florestal têm utilizado métodos de AM principalmente em tarefas de regressão, sendo as redes neurais artificiais (RNA) os mais estudados, apresentando em diversos casos *performance* superior aos modelos de regressão tradicionais na modelagem de atributos, como volume individual (BINOTI *et al.*, 2014; GORGENS *et al.*, 2014;) e altura total (MARTINS *et al.*, 2016; VENDRUSCOLO *et al.*, 2016).

Outros modelos de AM promissores ainda são pouco investigados quanto às aplicações na modelagem de atributos florestais, no entanto, têm sido amplamente estudados na área de sensoriamento remoto. Nessa situação, citam-se as máquinas de vetores de suporte (MVS) (ABDOLLAHNEJAD *et al.*, 2017), a floresta aleatória (RF, *random forest*) (GORGENS *et al.*, 2015) e o k-vizinhos mais próximos (k-NN, *k-nearest neighbors*) (BLANCHETTE *et al.*, 2015).

Os algoritmos de AM possuem parâmetros específicos, também conhecidos como hiperparâmetros, que devem ser configurados para a realização da etapa ajuste aos dados, denominada treinamento (BERGSTRA e BENGIO, 2012). De modo geral, os modelos de AM não produzem resultados ótimos sem que estes parâmetros sejam apropriadamente configurados, necessitando de métodos de busca que minimizam o erro do modelo (BERGSTRA e BENGIO, 2012). O método mais popular para otimização dos parâmetros é a busca em grade (grid-search), a qual consiste em um processo de pesquisa exaustiva em um espaço de busca definido pelo usuário (BERGSTRA e BENGIO, 2012).

Deste modo, utilizando uma base de dados para ajuste de modelos hipsométricos para *Pinus taeda* L., este trabalho teve como objetivos: apresentar o método e gerar diretrizes para a busca de parâmetros ótimos no treinamento de modelos de AM; e avaliar a *performance* de quatro modelos de aprendizado de máquina em comparação a modelos de regressão. Esperou-se com esse estudo evidenciar diferenças de precisão nas estimativas realizadas por modelos estatísticos de regressão e modelos de inteligência artificial devidamente configurados por meio de métodos de seleção de parâmetros e, deste modo, contribuir para utilização dos mesmos na modelagem florestal.

Material e métodos

Base de dados

Os dados utilizados neste estudo foram obtidos em plantios comerciais de *Pinus taeda*, localizados na região de Itapeva, estado de São Paulo. Na ocasião do inventário florestal, foram amostrados povoamentos de quatro idades diferentes por meio de parcelas temporárias circulares de 201 m² de área (8 m de raio). Na ocasião do inventário florestal, em cada parcela foram mensurados todos os diâmetros a altura do peito (DAP) e a altura total de cinco árvores normais e mais duas dominantes. A Tabela 1 apresenta o resumo estatístico dos dados utilizados para ajuste dos modelos hipsométricos.

Para a avaliação da capacidade de generalização dos modelos ajustados, realizou-se o particionamento dos dados empregando o método *hold-out*, o qual consiste na divisão da base de dados em dois conjuntos independentes, sendo um para treinamento e outro para teste dos modelos (SHALEV-SHWARTZ e BEN-DAVID, 2014). A razão utilizada para a divisão foi de 1/3 dos dados para teste e dois terços para treinamento dos algoritmos e ajuste dos modelos de

regressão. Para que a separação resultasse em dois grupos de amplitude e variação similares, os diâmetros a altura do peito foram ordenados dentro de cada povoamento, com a posterior divisão realizada sistematicamente ao separar a última árvore a cada três.

Tabela 1 – Características da base de dados dos povoamentos de *Pinus taeda*.

Table 1 – Characteristics of the *Pinus taeda* stands database.

Povoamento	Idade	N. de medições	DAP (cm)		H _{dom} (m)		H (m)	
			Média	CV%	Média	CV%	Média	CV%
A	6,7	99	12,7	26,0	10,1	9,9	8,9	18,0
B	7,9	460	19,1	23,6	16,6	10,2	15,0	14,0
C	9,7	213	17,0	27,1	15,6	12,8	13,9	18,7
D	11,8	215	20,5	23,9	18,9	10,1	16,8	15,5
Total		987	18,3	27,3	16,2	17,9	14,6	21,2

Em que: DAP = diâmetro a altura do peito medido a 1,3 m do solo; H_{dom} = altura média dominante das 100 árvores de maior diâmetro obtido por parcela; H = altura total; e CV% = Coeficiente de variação.

Modelos de regressão

Foram selecionados três modelos de regressão listados na literatura científica para o ajuste ao conjunto de treinamento e um quarto foi gerado a partir do procedimento backward. Os modelos foram ajustados no *software* R (R CORE TEAM, 2016) e são apresentados na Tabela 2.

Os modelos de regressão selecionados são denominados genéricos, uma vez que geram equações que permitem a estimativa para conjuntos de dados originados de povoamentos florestais com características silviculturais diferentes. Esses modelos tendem a apresentar precisão superior aos modelos hipsométricos tradicionais, apresentando características específicas de cada povoamento como variáveis independentes, como idade e altura dominante, as quais influenciam na relação altura/diâmetro (BARROS *et al.*, 2002).

Tabela 2 – Modelos de regressão selecionados para a relação hipsométrica de povoamentos de *Pinus taeda*.

Table 2 – Selected regression models for height-diameter relationship of *Pinus taeda* stands.

N.	Modelo
1	$\ln(ht_{ij}) = \beta_0 + \beta_1 \cdot \frac{1}{d_{ij}} + \beta_2 \cdot \frac{1}{I} + \beta_3 \cdot \frac{1}{d_{ij} \cdot I_j} + \varepsilon_i$
2	$ht_{ij} = \beta_0 + \beta_1 \cdot d_{ij} + \beta_2 \cdot d_{ij}^2 + \beta_3 \cdot h_{domj} + \varepsilon_i$
3	$ht_{ij} = \beta_0 + \beta_1 \cdot d_{ij} + \beta_2 \cdot h_{domj} + \beta_3 \cdot I_j + \beta_4 \cdot d_{ij}^2 + \beta_5 \cdot (I_j^2) + \varepsilon_i$
4	$\ln(ht_{ij}) = \beta_0 + \beta_1 \cdot \frac{1}{d_{ij}} + \beta_2 \cdot \ln(h_{domj}) + \beta_3 \cdot I_j + \beta_4 \cdot I_j^2 + \varepsilon_i$

Em que: ln = logaritmo neperiano; ht_i = altura total da árvore i na parcela j; d_{ij} = diâmetro a 1,3 m do solo, h_{domj} = altura média das árvores dominantes da parcela j; I_j = idade do povoamento da parcela j; ε_i = erro aleatório inerente à regressão.

Algoritmos de aprendizado de máquina

Foram utilizados quatro algoritmos de aprendizado de máquina: k-vizinhos mais próximos (k-NN), Floresta aleatória (RF), Máquinas de vetores de suporte (MVS) e Redes neurais artificiais (RNA), cujos treinamentos foram realizados no *software* R Studio 1.0.136. Para a realização dos treinamentos, os dados foram padronizados através de procedimentos anteriores inerentes à cada pacote. As variáveis de entrada utilizadas em todos os algoritmos foram DAP, idade e altura dominante.

K-vizinhos mais próximos (k-NN)

O k-NN é um método que consiste em memorizar os dados de treinamento para realizar previsões para uma nova instância desconhecida, a partir dos valores observados para uma quantidade k de vizinhos mais próximos (SHALEV-SHWARTZ e BEN-DAVID, 2014). O pacote R utilizado para treinamento de k-NN neste trabalho (*kknn*) permite a ponderação das distâncias entre as instâncias através de funções *kernel*. As funções *kernel*, quando aplicadas a uma distância calculada por uma métrica definida, estabelecem o peso de cada vizinho na determinação do valor de uma nova instância (MITCHEL, 1997). A descrição detalhada do método pode ser encontrada em Hechenbichler e Schliep (2004).

O número de vizinhos mais próximos (k) a serem utilizados para previsão de valores desconhecidos foi testado variando de 1 a 50. Também foram testadas duas métricas de distância: euclidiana e Manhattan, dadas pelas seguintes funções:

$$d_{euclidiana}(x_i, x_j) = \sqrt{\sum_{s=1}^n (x_{is} - x_{js})^2} \quad d_{Manhattan}(x_i, x_j) = \sum_{s=1}^n |x_{is} - x_{js}|$$

Em que: d = distância entre as instâncias $x_i (x_{i1}, x_{i2}, \dots, x_{in})$ e $x_j (x_{j1}, x_{j2}, \dots, x_{jn})$; e x_i e x_j = instâncias diferentes com n variáveis explicativas.

Foram testadas também diversas funções *kernel* para ponderação das distâncias, a saber: retangular, triangular, Epanechnikov, quadrática, cúbica, cosseno, Gauss e inversa, representadas pelas seguintes expressões:

$$K_{retangular} = \frac{1}{2} \quad K_{triangular} = (1 - |d|) \quad K_{Epanechnikov} = \frac{3}{4}(1 - d^2) \quad K_{quadrático} = \frac{15}{16}(1 - d^2)^2$$

$$K_{cúbica} = \frac{35}{32}(1 - d^2)^3 \quad K_{cosseno} = \frac{\pi}{4} \cos\left(\frac{\pi}{2}d\right) \quad K_{Gauss} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2} \quad K_{inversa} = \frac{1}{|d|}$$

Em que: K = função kernel; e d = distância entre duas instâncias.

Floresta aleatória (RF)

O algoritmo floresta aleatória (RF) consiste em uma técnica de AM que agrupa diversas árvores de decisão criadas a partir de uma base de treinamento, de modo que o modelo resultante consolida os resultados de todas as árvores para previsões (BREIMAN *et al.*, 2001). As árvores de decisão são modelos hierárquicos de aprendizado supervisionado compostos internamente por nós de decisão e folhas terminais. Cada nó de decisão aplica um teste aos dados de entrada, de modo a seguir por um dos ramos que levará a um novo nó de decisão ou a uma folha terminal, a qual determina o valor da variável de interesse (ALPAYDIN, 2010).

O treinamento foi realizado empregando o método *bagging*, que consiste na utilização de subconjuntos de dados da base de treinamento para geração das árvores. Além do particionamento de instâncias, o número de variáveis independentes aleatoriamente selecionadas para formação

de cada nó de decisão também é um parâmetro a ser definido antes do treinamento. A definição da quantidade de árvores e de variáveis a serem utilizadas em cada árvore tem como objetivo criar árvores de baixa correlação umas com as outras, melhorando a robustez e a capacidade de generalização do modelo final (BREIMAN *et al.*, 2001).

Neste estudo, o número de árvores de regressão (*Ntree*) e o número de variáveis selecionadas em cada nó (*Mtry*) corresponderam aos parâmetros testados no treinamento dos modelos de RF. Testaram-se valores de 10 a 500 para o parâmetro *Ntree*, com saltos de 10 unidades a cada treinamento, enquanto os valores 2 e 3 foram avaliados para *Mtry*. O número mínimo de cinco folhas terminais foi empregado por árvore (valor padrão do pacote R utilizado), com ausência de valor máximo definido, permitindo ao algoritmo gerar quantas folhas possíveis para as predições.

Máquinas de vetores de suporte (MVS)

As máquinas de vetores de suporte (MVS) são técnicas de AM baseadas na teoria de aprendizado estatístico desenvolvida por Vapnik (1995). Assim como os outros métodos de AM, as MVS possuem parâmetros que devem ser configurados para realização do treinamento, como o custo (*C*), que controla a relação entre a acuracidade e a complexidade do modelo, e a função *kernel*, que é utilizada para mapear as variáveis em um espaço de características multidimensional, de modo a elevar a probabilidade dos dados a serem modelados linearmente (CRISTIANINI e SHAWE-TAYLOR, 2000).

A função *kernel* empregada neste estudo foi a de base radial, a qual possui o parâmetro *gamma*, responsável pelo controle da forma das porções mais agudas da curva de ajuste. Para o treinamento das MVS, foi utilizada a função de erro do tipo II, denominada regressão-*v* (SMOLA, 1996), a qual possui o parâmetro *v* utilizado para determinar a proporção entre o número de vetores de suporte e o número total de instâncias na base de treinamento. Para o treinamento, foi utilizado $v = 0,5$ como valor padrão do pacote R (*e1071*).

A busca pelos parâmetros ótimos para treinamento do algoritmo MVS foi realizada conforme o método sugerido por Hsu *et al.* (2003). A primeira busca, denominada exploratória, consistiu na utilização de uma sequência de valores intervalados para os parâmetros *C* e *gamma*. Após a identificação de uma região promissora em que os erros tendiam a valores mais baixos, foi realizada a segunda busca, denominada refinada, utilizando sequências de valores com intervalos menores para cada parâmetro.

Redes neurais artificiais (RNA)

As redes neurais artificiais (RNA) são sistemas computacionais formados por unidades de processamento simples interligadas e paralelamente dispostas, denominadas neurônios, que armazenam o conhecimento e o generalizam para um conjunto de dados desconhecidos (HAYKIN, 2001). Durante o treinamento das RNA, um algoritmo de aprendizado ajusta os pesos sinápticos que ponderam os valores das variáveis de entrada e dos neurônios que compõem as camadas ocultas (HAYKIN, 2001). Os pesos sinápticos que conectam os neurônios, os neurônios e o *bias* das camadas ocultas e de entrada determinam as saídas da rede (CANDEL *et al.*, 2017). Considerando-se o vetor de entradas $x = []$, o vetor de pesos sinápticos $w = []$, e o *bias* b , a saída de um neurônio se dá por uma função de ativação em que:

$$\alpha = \sum_i w_i x_i + b$$

Utilizaram-se neste trabalho redes neurais artificiais multicamadas anteroalimentadas (*feedforward*), com algoritmo de treinamento de retropropagação dos erros, o qual utiliza

o método do gradiente descendente estocástico para atualização dos pesos sinápticos (CANDEL *et al.*, 2017). O número de neurônios testados na camada intermediária foi de 1 a 15 e o limite de ciclos de treinamento foi 3000. Ainda, o treinamento foi automaticamente interrompido pelo algoritmo quando a precisão do modelo não foi incrementada por 20 ciclos. A função de ativação padrão utilizada na camada de saída pelos modelos treinados pelo pacote R utilizado (*h2o*) é a linear, ao passo que, na camada oculta, foram testadas as funções tangente hiperbólica, linear retificada e *maxout*, cujas expressões matemáticas são apresentadas na Tabela 3.

Tabela 3 – Funções de ativação testadas no treinamento das redes neurais artificiais para estimativa da altura total.

Table 3 – Activation functions tested to estimate the total height through artificial neural networks.

Função de ativação	Fórmula
Tangente hiperbólica	$f(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$
Linear retificada	$f(\alpha) = \max(0, \alpha)$
Maxout	$f(\alpha_1, \alpha_2) = \max(\alpha_1, \alpha_2)$

O parâmetro taxa de aprendizado, que determina a taxa de atualização dos pesos sinápticos a cada nova iteração, e o termo momentum, que auxilia na tarefa de evitar mínimos locais e instabilidades relacionadas, não foram testados neste estudo, uma vez que o pacote *h2o* permite a utilização de aprendizado adaptativo, o qual combina o método de arrefecimento da taxa de aprendizado (*rate annealing*) e de treinamento do termo *momentum*, as quais adaptam o valor desses parâmetros durante o treinamento (CANDEL *et al.*, 2017).

Configuração de parâmetros

Para obtenção dos parâmetros que minimizam o erro das predições em cada algoritmo de AM, utilizou-se o método de busca em grade. Este método requer a definição de um conjunto ou sequência de valores para cada parâmetro, para que a busca seja realizada por meio de ajustes sucessivos, utilizando todas as combinações de parâmetros possíveis (BERGSTRA e BENGIO, 2012).

Para evitar a seleção de modelos sobreajustados aos dados de treinamento e de baixa capacidade de generalização, o treinamento dos modelos de AM foi realizado por meio da aplicação do método de validação cruzada *k-fold*, combinado ao método de busca em grade. Segundo Shalev-Shwartz e Ben-David (2014), o método *k-fold* consiste em particionar a base de dados aleatoriamente em *k* subconjuntos de dimensões similares e, para cada configuração do algoritmo, realizar *k* treinamentos, separando um subconjunto por vez para validação. Desse modo, todas as instâncias são utilizadas *k-1* vezes para treinamento e uma vez para validação. A medida de *performance* de cada configuração é obtida pela média do erro quadrático médio das *k* validações. Após a etapa de validação cruzada, o algoritmo efetua o treinamento de um modelo final para a configuração testada, utilizando toda a base de dados de treinamento.

Para cada algoritmo de AM, a melhor configuração foi a que conferiu ao modelo o menor erro quadrático médio obtido por validação cruzada. Os pacotes R utilizados

realizam a validação cruzada *k-fold* automaticamente a partir da determinação do número de subconjuntos (*folds*) pelo usuário. Neste estudo, utilizou-se o método *5-fold*. Adicionalmente, o tempo de busca para os parâmetros por meio do método *5-fold* foi computado para cada algoritmo.

Avaliação da performance dos modelos

A pré-avaliação dos modelos de regressão ajustados foi realizada pelo cálculo dos indicadores usualmente utilizados para comparação de modelos estatísticos: o coeficiente de determinação ajustado (R^2_{aj}) e o erro padrão da estimativa ($S_{yx\%}$), expressos pelas seguintes fórmulas:

$$R^2 = 1 - \frac{\sum(Diff)^2}{\sum(Y_i - \bar{Y})^2} \quad R^2_{aj} = 1 - [(1 - R^2) * \left(\frac{n-1}{n-p-1}\right)] \quad S_{yx} = \sqrt{\frac{\sum(Diff)}{n-p}} \quad S_{yx\%} = \frac{S_{yx}}{\bar{Y}} * 100$$

Em que: \bar{Y} = média dos valores observados; n = número de observações; e p = número de parâmetros do modelo.

A discrepância logarítmica gerada pelo ajuste dos modelos logarítmicos (1 e 4) foi corrigida por meio da aplicação do Fator de Correção de Meyer (FCM) em todos as predições resultantes desses modelos. Após a correção da discrepância logarítmica dos modelos, os parâmetros R^2_{aj} e $S_{yx\%}$ foram recalculados. O fator de correção é representado matematicamente pela seguinte fórmula:

$$FCM = e^{0,5 * S_{yx}^2}$$

Em que: e = constante de Euler.

Os modelos de AM e estatísticos foram avaliados conjuntamente quanto à *performance* de ajuste aos dados de treinamento e teste. A qualidade dos ajustes foi avaliada a partir dos critérios: correlação entre os valores estimados e os observados (r), raiz quadrada do erro médio percentual (RMSE%) e o Desvio (D), representados matematicamente pelas seguintes fórmulas:

$$r = \frac{cov(Y_i, \hat{Y}_i)}{\sqrt{(var(Y_i) * var(\hat{Y}_i))}} \quad RMSE\% = \frac{100}{\bar{Y}} \sqrt{\frac{\sum_{i=1}^n (Diff)^2}{n}} \quad D = \frac{(\sum Diff_i)}{n} \quad Diff = (Y_i - \hat{Y}_i)$$

Em que: Y_i = valor observado; \hat{Y}_i = valor estimado; n = número de observações; *Diff* = diferença.

Adicionalmente, os modelos foram avaliados por meio da distribuição gráfica de resíduos, para verificação de tendências ao longo da linha de ajuste.

Resultados e discussão

Ajuste dos modelos de regressão

A Tabela 4 apresenta os coeficientes ajustados para os modelos de regressão. Os indicadores de ajuste sugerem que todos os modelos foram adequados para uso na estimativa da altura dos povoamentos, sendo o modelo 3 o que apresentou o menor erro padrão da estimativa ($S_{yx\%}$) e o maior coeficiente de determinação ajustado (R^2_{aj}).

Tabela 4 – Coeficientes de regressão e indicadores estatísticos dos modelos de hipsométricos ajustados.

Table 4 – Regression coefficients and statistical indicators of adjusted height-diameter models.

Modelo	b0	b1	b2	b3	b4	b5	R^2_{aj}	$S_{yx\%}$
1	3,696**	-10,391**	-4,953**	21,509*			0,726	9,69
2	-3,701**	0,770**	-0,011**	0,496**			0,861	6,90
3	-7,290**	0,753**	0,448**	0,858*	-0,010**	-0,038*	0,864	6,82
4	1,021**	-6,067**	0,576**	0,081**	-0,004*		0,861	6,91

Em que: ** significativo a 1%; * significativo a 5%.

Resultados da busca em grade

K-vizinhos mais próximos

A melhor configuração obtida para k-NN foi $k = 16$, função de ponderação quadrática e métrica de distância de Manhattan. A métrica de distância e a função de ponderação tiveram pouca influência na *performance* dos resultados, sendo o número de vizinhos mais próximos o parâmetro determinante para qualidade das estimativas. A quadrática combinada com a métrica de distância de Manhattan foi a função que minimizou o erro quadrático, no entanto, a diferença entre o MSE obtido pelo modelo de melhor *performance* e o modelo que ocupou a 800ª posição foi inferior a 1%.

A função inversa de ponderação é comumente utilizada nos estudos aplicados à mensuração florestal (HAARA *et al.*, 1997; TOMMOLA, *et al.*, 1999; SANQUETTA *et al.*, 2013). Contudo, no presente estudo, o melhor modelo gerado pela função inversa obteve *performance* inferior aos melhores modelos das demais funções, porém, com diferença máxima de 0,23%. Esse resultado corrobora a afirmação de Hechenbichler e Schliep (2004) de que a escolha da função *kernel* não é crucial e que maiores diferenças são obtidas somente quando é escolhida a função retangular, a qual aplica o mesmo peso para todos os vizinhos, independentemente da distância.

Sanquetta *et al.* (2013), ao analisarem a *performance* de k-NN para estimar o estoque de carbono em indivíduos de *Araucaria angustifolia* (Bertol.) Kuntze, observaram similaridade entre os resultados obtidos por três configurações distintas, nas quais variaram o número de vizinhos (1 ou 3) e a métrica de distância (euclidiana e euclidiana ao quadrado). No presente estudo, também foi observada elevada similaridade entre diversos modelos gerados com diferentes configurações, no entanto, os modelos treinados com $k \leq 5$ e com elevado número de vizinhos apresentaram precisão inferior aos demais, independentemente da métrica de distância e da função de ponderação escolhida.

De modo geral, o valor ótimo para k variou entre 10 e 20, considerando-se as diferentes funções *kernel* e as métricas de distância. De modo similar ao observado neste estudo, Tommola *et al.* (1999) afirmaram que um baixo valor para k resulta em sobreajuste aos dados de treinamento, enquanto um elevado valor provoca um subajuste, criando curvas mais suaves e fazendo com que as estimativas tendam aos valores próximos à média do conjunto de dados de ajuste. Esses autores observaram aumento no percentual de erro na estimativa da razão entre sortimentos de um povoamento florestal na Finlândia quando k foi superior a 20, tal como ocorrido no presente estudo.

Florestas aleatórias

Para o algoritmo RF, os melhores valores para os parâmetros $Mtry$ e $Ntree$ foram 390 e 2, respectivamente. Os treinamentos realizados utilizando duas variáveis por nó foram superiores, de modo geral, aos realizados utilizando três. O valor recomendado por Breiman (2001) para o parâmetro $Mtry$ é $p/3$, onde p é o número de variáveis de entrada. Esse valor não deve ser considerado como regra, uma vez que no presente estudo p foi igual a 3 e os melhores resultados foram obtidos com $Mtry = 2$.

Não foram observadas tendências ou regiões ótimas para o parâmetro n , uma vez que a diferença entre o melhor modelo ($n = 390$ e $Mtry = 2$) e o pior ($n = 10$ e $Mtry = 3$) foi de 0,5%. No entanto, foi observado que os modelos treinados com $Mtry \leq 100$ foram mais instáveis e frequentemente apresentaram maiores valores de MSE, indicando que os treinamentos realizados com poucas árvores tendem a apresentar *performance* de generalização inferior, embora um elevado número de árvores não necessariamente implique melhores resultados.

Segundo Hastie (2003), o algoritmo RF estabiliza a partir de 200 árvores e o valor ótimo de $Mtry$ depende da tarefa em questão, devendo ser tratado como um parâmetro a ser ajustado. No presente estudo, a estabilização do erro foi observada a partir de aproximadamente 50 árvores.

Máquinas de vetores de suporte

A busca exploratória de uma região promissora para a configuração do algoritmo MVS obteve como melhor resultado $C = 10$ e $gamma = 0.1$. A partir desse resultado, foi realizada uma busca mais restrita, na qual o espaço de possibilidades compreendeu os valores entre 1 e 20 para C e entre 0.01 e 0.2 para $gamma$, sendo encontrados os valores 17 e 0,06 como os ótimos, respectivamente.

A grade utilizada na busca exploratória pelo parâmetro C foi ampla o suficiente para abranger o valor ótimo, o qual foi encontrado na segunda busca. Por outro lado, a grade exploratória não foi capaz de abranger o valor ótimo para $gamma$, uma vez que o valor obtido na primeira busca foi igual ao limite inferior (0,1). Por essa razão, a segunda busca foi realizada empregando um limite inferior ainda menor (0,01), tendo sido o valor adequado para encontrar o ótimo de 0,06.

Redes neurais artificiais

A rede que apresentou os melhores resultados foi treinada utilizando 13 neurônios na camada oculta e função de ativação *maxout*. As funções *maxout* e tangente hiperbólica apresentaram resultados similares e foram superiores à função linear retificada em todos os treinamentos. As redes treinadas com apenas um neurônio apresentaram os piores resultados em comparação às redes treinadas com mais neurônios e, de modo geral, maior precisão foi observada em redes treinadas com pelo menos 3 neurônios.

No estudo de Özçelik *et al.* (2013), que testaram RNA para a estimativa de altura de árvores de zimbros na Criméia, empregando a função de ativação sigmoide combinada às diferentes configurações de parâmetros, os resultados mais satisfatórios foram obtidos por redes treinadas com um ou dois neurônios na camada oculta. Por outro lado, Vendruscolo *et al.* (2016), utilizando a mesma função de ativação para estimativa de alturas individuais em povoamentos de *Tectona grandis*, observaram melhor *performance* de generalização em redes treinadas com seis a dez neurônios na camada oculta. Outros autores, ao testarem diferentes configurações de RNA, obtiveram resultados que demonstram a ausência de uma configuração mais indicada para utilização de redes neurais artificiais na modelagem da relação hipsométrica, ou mesmo de outros atributos florestais (VENDRUSCOLO, 2010; BINOTI *et al.*, 2014; MARTINS *et al.*, 2016) e

que, desse modo, diferentes parâmetros devem ser testados, a fim de se obter a configuração de melhor *performance* para cada caso.

Avaliação dos modelos

A Tabela 5 apresenta os indicadores de avaliação dos modelos ajustados para os conjuntos de treinamento e teste. Excetuando-se o modelo genérico 1, o qual apresentou estatísticas inferiores aos demais, os modelos apresentaram indicadores favoráveis e similares entre si, com RMSE% inferiores a 8,3% e baixos valores para o desvio.

O modelo de MVS apresentou o menor RMSE% (7,85%), enquanto a RNA obteve o maior valor para a correlação entre as alturas estimadas e observadas (0,935) no conjunto de treinamento. Ambos apresentaram distribuição de resíduos equilibrada (Figuras 1 e 2), assim como o modelo de RF, o qual apresentou o menor desvio. Os algoritmos k-NN e RF se diferenciaram na etapa de treinamento por apresentarem as melhores *performances* de ajuste, denotadas pelas maiores correlações e menores RMSE%, no entanto, foram equivalentes aos demais modelos testados na etapa de teste, quanto aos indicadores estatísticos.

Apesar de terem apresentado indicadores estatísticos similares aos modelos de AM na etapa de teste, todos os modelos estatísticos tiveram *performance* inferior aos modelos de AM na dispersão de resíduos, apresentando tendências à subestimativa da altura das árvores mais altas e mais baixas. (Figura 1). O modelo genérico 1 foi o menos acurado, possivelmente devido à ausência da variável altura dominante, a qual demonstrou ter sido decisiva na *performance* dos modelos testados. Ademais, todos os algoritmos de AM apresentaram dispersão de resíduos satisfatória.

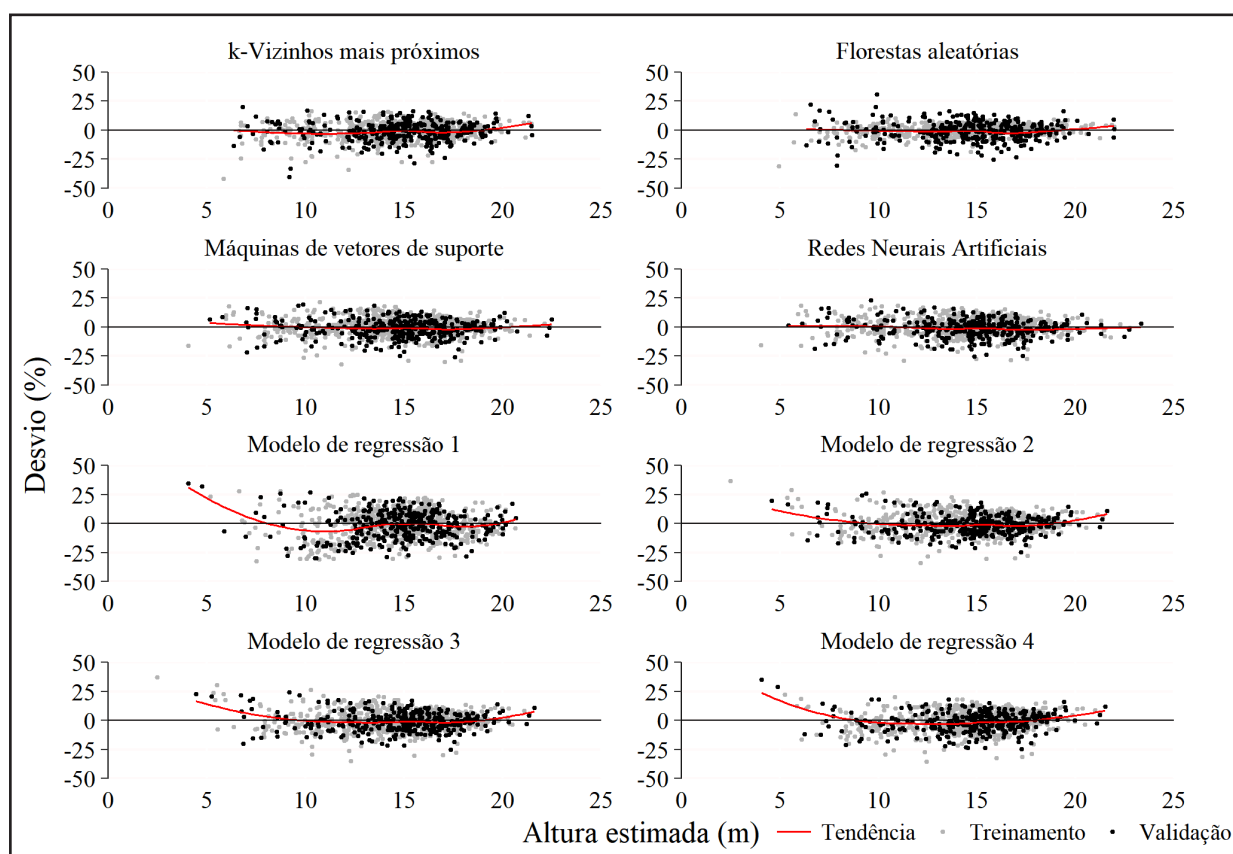
Tabela 5 – Indicadores de *performance* dos modelos estatísticos e de aprendizado de máquina na estimativa da altura total.

Table 5 – Performance indicators of statistical and machine learning models to estimate total height.

Modelo	Treinamento			Teste		
	R	RMSE%	D (m)	R	RMSE%	D (m)
RNA	0,942	7,17	-0,032	0,935	7,88	-0,218
k-NN	0,952	6,61	-0,021	0,927	8,22	-0,186
RF	0,983	3,96	-0,001	0,929	8,08	-0,156
SVM	0,939	7,36	-0,010	0,934	7,85	-0,184
Modelo 1	0,854	11,15	-0,110	0,859	11,20	-0,237
Modelo 2	0,928	7,94	$-3 \cdot 10^{-14}$	0,931	8,00	-0,162
Modelo 3	0,930	7,83	$-8 \cdot 10^{-14}$	0,932	7,95	-0,163
Modelo 4	0,930	7,95	-0,047	0,927	8,27	-0,200

Figura 1 – Dispersão dos resíduos relativos nas estimativas realizadas pelos modelos estatísticos de regressão e modelos de aprendizado de máquina.

Figure 1 – Dispersion of residuals in the estimates made by statistical regression models and machine learning models.

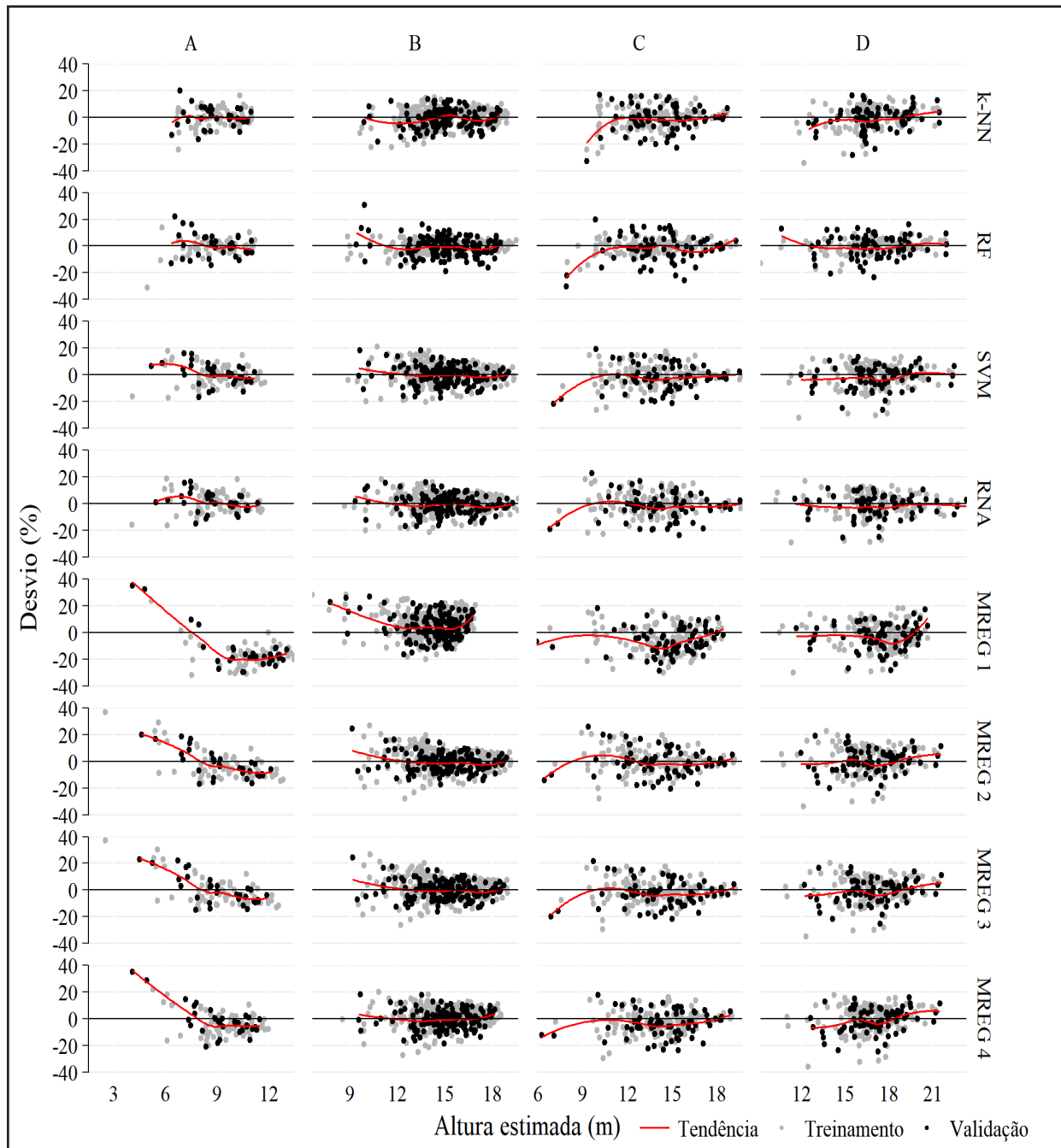


Fonte: Autor (2019)

Os modelos de aprendizado de máquina apresentaram melhor distribuição de resíduos ao longo da linha de ajuste em comparação aos modelos estatísticos, considerando a estratificação por povoamento (Figura 2). Os modelos de MVS e RNA podem ser considerados os que apresentaram os melhores resultados, uma vez que proporcionaram distribuição de resíduos satisfatoriamente homogênea para todos os povoamentos, apresentando tendência somente para as menores árvores do povoamento C, o que ocorreu também para os demais modelos ajustados. Todos os modelos estatísticos de regressão apresentaram elevado viés nas estimativas do povoamento A, além de tendências em ao menos uma das extremidades das nuvens de pontos dos demais povoamentos, o que confirmou a superioridade dos algoritmos de AM na estimativa da altura dos povoamentos estudados.

Figura 2 – Dispersão dos resíduos relativos estimados pelos algoritmos de aprendizado de máquina e modelos de regressão (MREG) com estratificação por idade: A (6,7 anos), B (7,9 anos), C (9,7 anos), D (11,8 anos).

Figure 2 – Dispersion of residuals in the estimates made by statistical regression models (MREG) and machine learning models stratified by age: A (6.7 years), B (7.9 years), C (9.7 years) e D (11.8 years).



Fonte: Autor (2019)

A partir dos resultados expostos acima, observou-se que os modelos de AM podem apresentar vantagens quanto à dispersão dos resíduos em bases de dados não estratificadas, em

comparação aos modelos estatísticos.

Dentre os modelos de inteligência artificial empregados neste estudo, as máquinas de vetores de suporte e redes neurais artificiais merecem destaque por apresentarem os menores valores de RMSE% e correlação entre as alturas estimadas e observadas, além da distribuição de resíduos mais satisfatória na etapa de generalização. Deste modo, é possível afirmar que a *performance* destes modelos foi equivalente no presente estudo.

O resultado aqui obtido difere-se do observado por Cordeiro *et al.* (2015), em que a RNA treinada superou a MVS e o modelo de Schumacher & Hall na estimativa do volume de *Acacia mangium* Willd. A divergência entre resultados indica que não há superioridade absoluta de um dos modelos na estimativa de alturas em povoamentos florestais, mas que ambos tendem a apresentar elevada *performance* de ajuste, por vezes, superior aos métodos estatísticos de regressão. Além disso, os algoritmos k-NN e RF também apresentaram *performance* superior aos modelos estatísticos, tendo apresentado como vantagem em relação aos outros dois algoritmos menor tempo para treinamento e geração do modelo preditor.

Diretrizes para a utilização de algoritmos de AM na modelagem da relação hipsométrica

A realização da busca em grade para otimização dos parâmetros dos algoritmos de AM foi importante para seleção de modelos que superaram a *performance* dos modelos de regressão. A obtenção dos parâmetros que minimizam os erros das estimativas requer um esforço computacional que, em alguns casos, pode inviabilizar a utilização de alguns desses métodos quando há necessidade de resultados imediatos. Contudo, o custo computacional em termos de tempo e de capacidade de processamento pode ser reduzido a uma escala aceitável quando a grade de busca apresenta menores dimensões (BERGSTRA e BENGIO, 2012). Para isso, são necessárias diretrizes para realização da busca, as quais foram objeto de estudo neste trabalho.

O tempo necessário para realização da busca em grade para cada algoritmo, utilizando o método de validação cruzada 5-*fold* em uma máquina Intel® Core™ i7-5500U 2.40GHz com dois núcleos e 8Gb de RAM, foi: k-NN = 6,7 segundos; RF = 5,4 minutos; MVS = 3,4 minutos (busca exploratória) e 5,5 minutos (busca refinada); RNA = 12,8 minutos. Para o treinamento do algoritmo RF, o tempo de busca de parâmetros pode ser reduzido drasticamente, uma vez que o erro se estabilizou a partir da geração de 50 árvores. A título de exemplo, o tempo de processamento da busca em grade para treinamento do algoritmo com 500 árvores, e visando buscar o valor mais adequado para *Mtry* em um espaço de busca de 2 valores (2 ou 3), foi de 17,2 segundos.

O tempo gasto para treinamento de uma MVS com $C = 1$ e $\gamma = 0.1$, empregando o conjunto de treinamento utilizado neste estudo como entrada, foi de 1,7 segundos, enquanto para o mesmo valor de γ e com $C = 100$, o tempo de treinamento foi de 7,19 segundos, o que demonstrou que, para este caso, quanto maior o valor de C maior o tempo de treinamento. O conjunto de valores utilizado para busca de C na grade de busca exploratória foi amplo o suficiente para localizar uma região promissora (1 a 100).

Para o treinamento de RNA, o teste de diferentes funções de ativação se demonstrou importante, uma vez que a precisão obtida pelas redes treinadas com função linear retificada foi inferior às treinadas com as outras funções. Candel *et al.* (2017) afirmam que é difícil determinar a melhor função de ativação a ser usada, e que uma pode ser superior às demais em cenários diferentes, sendo adequada a aplicação da busca em grade para seleção da melhor função em cada problema.

É importante ressaltar que a avaliação de algoritmos de aprendizado de máquina deve ser realizada em um ou mais conjuntos de teste, e que a avaliação em um conjunto de treinamento não avalia a capacidade de generalização de um algoritmo treinado, especialmente quando são empregados k-vizinhos mais próximos e florestas aleatórias, uma vez que esses modelos tenderam a sobreajustar os dados de treinamento no presente estudo. Para efeitos práticos, o

particionamento da base de dados em conjuntos de treinamento e de teste deve ser realizada apenas para comparação entre modelos, devendo, após a escolha do melhor método, ser realizado um treinamento final com a base completa (ALPAYDIN, 2010).

Ainda, a avaliação estratificada dos resíduos demonstrou ser de suma importância para o detalhamento dos erros obtidos quando se utiliza uma base de dados heterogênea, uma vez que tendências dentro dos povoamentos podem não ser identificadas por meio da análise dos resíduos como um todo. Esse método deve ser utilizado para identificação de modelos com maior capacidade de generalização para todos os povoamentos, quando tratados de maneira independente.

Conclusões

Os modelos de aprendizado de máquina apresentam distribuição de resíduos menos tendenciosa que os modelos de regressão.

A avaliação estratificada dos resíduos relativos pode ser realizada para identificação de modelos com maior capacidade de generalização quando são utilizadas bases de dados não estratificadas para ajuste.

O treinamento de algoritmos de aprendizado de máquina deve ser tratado como uma tarefa de otimização com vistas à identificação dos parâmetros que minimizam o erro das previsões;

A melhor configuração para o algoritmo k-vizinhos mais próximos pode ser obtida utilizando somente a busca pelo número ótimo de vizinhos mais próximos.

A melhor configuração para o algoritmo RF pode ser obtida buscando o valor ótimo para o parâmetro *Mtry*, com um número de árvores (*Ntree*) superior a 50.

Redes neurais artificiais apresentam elevada precisão e baixa tendenciosidade nas estimativas de altura de árvores. O treinamento destas no pacote *h2o* deve ser realizado de modo a identificar a melhor função de ativação, além da quantidade ideal de neurônios na camada oculta.

Assim como as RNA, as máquinas de vetores de suporte com *kernel* função de base radial geram modelos de elevada precisão e baixa tendenciosidade na estimativa de alturas de povoamentos florestais, sendo importante a utilização da busca em grade para identificação dos valores para os parâmetros *C* e *gamma* que conferem maior precisão ao modelo.

Os modelos de máquina de vetores de suporte e rede neural artificial obtiveram *performances* de generalização superiores aos demais modelos quanto aos indicadores estatísticos e dispersão de resíduos.

Referências

ABDOLLAHNEJAD, A. *et al.* Prediction of Dominant Forest Tree Species Using QuickBird and Environmental Data. **Forests**, [s.l.], v. 8, n. 2, p.42-60, fev. 2017.

BARROS, D. A. *et al.* Comportamento de modelos hipsométricos tradicionais e genéricos para plantações de *Pinus oocarpa* em diferentes tratamentos. **Boletim de Pesquisa Florestal**, Colombo, n. 45, p. 03-28, jul./dez. 2002.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **The Journal Of Machine Learning Research**, Montréal, Qc, Canada, v. 13, n. 1, p.281-305, jan. 2012.

BINOTI, D.H.B.; BINOTI, M.L.M.S.; LEITE, H.G. Configuração de Redes Neurais Artificiais para Estimação do Volume de Árvores. **Revista Ciência da Madeira**, Pelotas, v. 5, n. 1, p.58-67, 31 maio 2014.

BLANCHETTE, D. *et al.* Predicting wood fiber attributes using local-scale metrics from terrestrial LiDAR data: A case study of Newfoundland conifer species. **Forest Ecology and Management**,

[s.l.], v. 347, p.116-129, jul. 2015.

BREIMAN, L. Random Forests. **Machine Learning**, [s.l.], v. 45, n. 1, p.05-32, out. 2001.

CANDEL, A. *et al.* **Deep Learning with H2O**. Disponível em: <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf>. Acesso em: 02 jul. 2017.

CORDEIRO, M. A., *et al.* 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, 255-261, set. 2015.

CRISTIANINI, N.; SHAWE-TAYLOR, J. An introduction to support vector machines and other kernel-based learning methods. New York: Cambridge University Press, 2000.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011. 378p.

GORGENS, E. B. *et al.* Influência da arquitetura na estimativa de volume de árvores individuais por meio de redes neurais artificiais. **Revista Árvore**, Viçosa-MG, v. 38, n. 2, p.289-295, abr. 2014.

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

HAARA, A.; MALTAMO, M.; TOKOLA, T. The K-nearest-neighbour method for estimating basal-area diameter distribution. **Scandinavian Journal Of Forest Research**, Uppsala, v. 12, n. 2, p.200-208, maio, 1997. HECHENBICHLER, K; SCHLIEP, K. **Weighted k-Nearest-Neighbor Techniques and Ordinal Classification**. Discussion paper 399, Ludwig-Maximilians University Munich, Munich, 2004. Disponível em: https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf. Acesso em: 20 jun. 2017.

MARTINS, E. R. *et al.* Configuração de redes neurais artificiais para estimação da altura total de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias (Agrária)**, Recife, v. 11, n. 2, p. 117-123, 2016.

ÖZÇELIK, R. *et al.* Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. **Forest Ecology And Management**, [s.l.], v. 306, p.52-60, out. 2013.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2016. Disponível em: <http://www.R-project.org/>. Acesso em: 04 ago. 2017.

ROBINSON, A. P.; LANE, S. E.; THÉRIEN, G. Fitting forestry models using generalized additive models: a taper model example. **Canadian Journal Of Forest Research**, [s.l.], v. 41, n. 10, p.1909-1916, out. 2011.

SANQUETTA, C. R. *et al.* On the use of data mining for estimating carbon storage in the trees. **Carbon Balance And Management**, [s.l.], v. 8, n. 1, p.6-14, 2013.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning: From Theory to Algorithms**. New York: Cambridge University Press, 1. ed., 2014.

SMOLA, A. Regression Estimation with Support Vector Learning Machines. 1996. 78 p. Master's thesis (Physics) - Technische Universitat At Munchen, Munchen, 1996.

TOMMOLA, M., *et al.* Estimating the characteristics of a marked stand using k-nearest-neighbour regression. **Journal of Forest Engineering**, [s.l.], v. 10, p.75-81, 1999.

VAPNIK, V N. **The nature of statistical learning theory**. 2. ed. New York: Springer-Verlag, 2000. 314 p.

VENDRUSCOLO *et al.* Height prediction of *Tectona grandis* trees by mixed effects modelling and artificial neural networks. **International Journal of Current Research**, [s.l.], v. 8, n. 12, p.43189-43195, dez. 2016.