ESTIMATIVA DIAMÉTRICA A PARTIR DA ALTURA UTILIZANDO ALGORITMO *RANDOM FOREST*

DIAMETRIC ESTIMATE FROM HEIGHT USING RANDOM FOREST ALGORITHM

Paulo Ricardo Santos Miranda[1] Sthefany Mendes Zuba Thais Sales Gonçalves
Gabriela Letícia Ramos de Carvalho Christian Dias Cabacinha Carlos Alberto Araújo Júnior

RESUMO

O manejo florestal atual demanda o conhecimento da floresta e a sua relação com meio, bem como do seu crescimento e a suas potencialidades. O objetivo deste trabalho foi avaliar a eficiência do processamento do algoritmo *Random Forest* para estimativa de distribuição diamétrica em árvores do gênero *Eucalyptus sp.* Foi aplicado o algoritmo *Random Forest* para predição de valores para diâmetro. Foram avaliados dos modelos de árvore de decisões, sendo eles baseados na variável dependente diâmetro (DAP) e variáveis independentes, altura (h), idade (I) e área de plantio (A) para a estimação de DAP. O tipo 2 se demonstrou mais satisfatório para predição dos valores.
Palavras-chave: Inteligência artificial; Mensuração florestal; Diâmetro.

ABSTRACT

Current forest management demands knowledge of the forest and its relation to the environment, as well as its growth and potential. The objective of this work was to evaluate the efficiency of the Random Forest algorithm to estimate diameter distribution in trees of the genus Eucalyptus sp. The Random Forest algorithm was applied to predict values for diameter. The decision tree models were evaluated, based on the dependent variable diameter (DBH) and independent variables, total height (HT), age (I) and planting area (A) for the estimation of DBH. Type 2 was shown to be more satisfactory for predicting values .
Keywords: Artificial intelligence; Forest mensuration; Diameter.

INTRODUÇÃO

O setor florestal tem se apresentado como grande fornecedor de energia, matéria-prima para indústria de transformação e construção civil. Tal fato é suportado pela alta produtividade das suas áreas de florestas plantadas no Brasil. Em 2016, o país liderou o ranking global de produtividade, alcançando uma média de 35,7 m³/ha para florestas de eucalipto e 30,5 m³/ha ao ano nos plantios de pinus (IBA ,2017) (SNIF,2016).

O manejo florestal atual demanda o conhecimento da floresta e a sua relação com meio, bem como do seu crescimento e a sua potencialidade para produção de madeira para celulose, chapas, serraria e energia, dentre outros fins. A avaliação acerca do povoamento florestal só pode ser feita através de medições representativas da floresta, sendo que as principais variáveis utilizadas são DAP (Diâmetro à altura do peito) e a altura.

A altura é considerada uma importante característica da árvore, a qual pode ser medida ou estimada. Seu valor é importante para o cálculo do volume das árvores, além de servir como indicador da qualidade produtiva (SILVA et al. 2012). Diferentemente do diâmetro, a variável altura não é de fácil medição, e em plantios homogêneos a sua mensuração geralmente é feita por aparelhos ópticos baseados em princípios trigonométricos (MACHADO e FIGUEIREDO FILHO, 2003). Com o avanço da tecnologia e modernização nas atividades de mensuração florestal, novos meios para obtenção das alturas das árvores foram desenvolvidos, dentre eles o uso o LiDAR.

O LiDAR (*Ligth Detection and Ranging)* é uma das maneiras de fornecer informações precisas acerca da estrutura vertical da floresta, como altura e cobertura do dossel. Seu principio é baseado em pulsos de laser que são enviados do sensor, esses são capazes de penetrar no dossel da floresta e disponibilizar essas informações (URBAZAEV et al, 2018). O LiDAR, portanto, tem sido usado como importante fonte de dados para o inventário florestal. A estimativa de altura do povoamento com o LiDAR pode ser trabalhada levando em conta dois tipos de abordagens: abordagens baseadas em árvores, a qual calcula as alturas das árvores de acordo

[1]Estudante de graduação em Engenharia Florestal, Instituto de Ciências Agrárias, Universidade Federal de Minas Gerais, Av. Universitária, 1000- Universitário, CEP:39.404-547, Montes Claros(MG), Brasil. E-mail: ricarddosm@hotmail.com

com altura média das árvores extraídas com base no delineamento individual da copa das árvores, e abordagem baseada em parcela, a qual estima as alturas a partir de estatísticas descritivas em nível de parcela (LEE, 2018).

Devido ao fato de que as medições do LiDAR estão associadas à estrutura vertical da floresta, faz-se necessário estimar a variável diâmetro. Para tais estimativas, se faz necessário o uso de modelos que consigam obter valores estimados próximos dos reais. Nesse sentido, a *Random Forest* se apresenta como alternativa à solução dessa problemática. Tal algoritmo é um método que gera um conjunto de árvores aleatórias treinadas individualmente e então combina os seus resultados levando a uma única previsão (JEUNE, 2018). A *Random Forest* foi desenvolvida por Breiman (2001) para realizar regressão e classificação. As árvores de decisão fazem a divisão do banco de dados de forma interativa, gerando, a partir de um nó pai, diversos subconjuntos chamados de nós filhos (ZAMO et al., 2014). A divisão dos dados é realizada de acordo com alguns critérios (SHAIKHINA et al., 2017), sendo comumente feito por um "limiar de corte das variáveis independentes" atuando na diminuição da variância dos nós filhos formados (JAMES et al*.,* 2013; ZAMO et al., 2014).

O algoritmo é apontado como capaz de superar as abordagens de regressão clássica, principalmente quando há relações complexas e não lineares entre as variáveis. Os resultados também são mais fáceis de interpretar, em relação a outros modelos de regressão (STROBL et al*.,* 2009). Assim, o objetivo deste trabalho foi avaliar a eficiência do processamento do algoritmo *Random Forest* na estimativa diamétrica em árvores do gênero *Eucalyptus sp.*

MATERIAIS E MÉTODOS

Os dados utilizados nesta pesquisa são de inventários florestais realizados em plantios florestais do norte de Minas Gerais entre os anos de 2012 a 2016. Para o processamento, os dados foram separados em validação e treinamento, sendo 98% dos dados para treinamento e 2% para validação. Ainda, foram consideradas duas formas de separação dos dados, sendo o primeiro aquele que considerou como dados de treinamento aqueles medidos nos anos de 2012 a 2015 e os dados de 2016 utilizados para validação (Modelo 1) e o segundo aquele que considerou uma escolha aleatória (Modelo 2) para compor os conjuntos de treinamento e validação. O software para análise dos dados foi o Statistica 13, estabelecendo como parâmetros 500 árvores de decisões com dois números de predições.

Para avaliação dos resultados, foram calculadas as estatísticas dos modelos analisados, sendo BIAS (equação 1), raiz quadrada do erro médio quadrático (RQME) (equação 2), correlação entre valores estimados e valores observados (r) (equação 3) e erro médio percentual (EMP) (equação 4). Além disso, foram criados gráficos de dispersão analisando os valores estimados e observados e histogramas de resíduos. Para obtenção do erro de estimativa, foi utilizada a fórmula $e_i = \hat{y}_i - y_i$.

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) \quad \text{(equação 1)}$$

$$\text{RQME (\%)} = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad \text{(equação 2)}$$

$$r = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right).\ (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(equação 3)}$$

$$\text{EMP} = \frac{1}{n}\sum_{i=1}^{n}\left[100.\frac{\left(y_i - \hat{y}_i\right)}{y_i}\right] \quad \text{(equação 4)}$$

Em que $\hat{y}_i$ é o valor estimado e y_i é o valor observado para o i-ésimo dado, n é o número total de observações, $\bar{y}$ é a média dos valores observados e $\hat{y}$ é a média dos valores estimados.

RESULTADOS E DISCUSSÃO

As estatísticas dos modelos são apresentadas na (Tabela 1). Quando se analisa a correlação (quanto mais próxima de 1, mais as variáveis estão relacionadas), é possível perceber que a modelagem 2, tanto de treinamento quanto validação apresentaram melhor desempenho, portanto uma forte correlação. Em relação à estatística da RQME (quanto menor for o erro do modelo, melhor), esse mesmo modelo se destaca, se apresentando como o mais eficiente, portanto o apresenta menor erro na estimativa dos valores de distribuição de diâmetro. É valido ressaltar que o tipo 1 de validação se apresentou como o menos eficiente, por não haver dados representativos da variável de interesse.

Analisando os gráficos é possível perceber que o treinamento (TIPO1), apresentou uma grande relação entre os valores de diâmetros observados e estimados, isso expresso pela concentração dos pontos, refletindo em uma menor dispersão. O mesmo pode ser observado em relação ao treinamento do Tipo 2. Quando analisado o gráfico de validação do Tipo 1, verifica-se a presença de valores superestimados. O Tipo 2 de validação ele se distribui melhor, mesmo havendo pontos de dispersões (Figura 1).

Tabela 1: Estatísticas de treinamento e validação da árvore de decisões para estimativa de diâmetro
Table 1: Calculated statistics for training and validation of the decision tree for diameter estimation

Modelo de treinamento	Etapa	Bias	RQME	r	RQME%	EMP
1	Treinamento	0.00	1.57	0.93	11.02	1.74
	Validação	0.51	1.97	0.59	11.80	4.83
2	Treinamento	0.00	1.58	0.93	11.02	1.74
	Validação	0.16	1.63	0.91	11.62	2.77

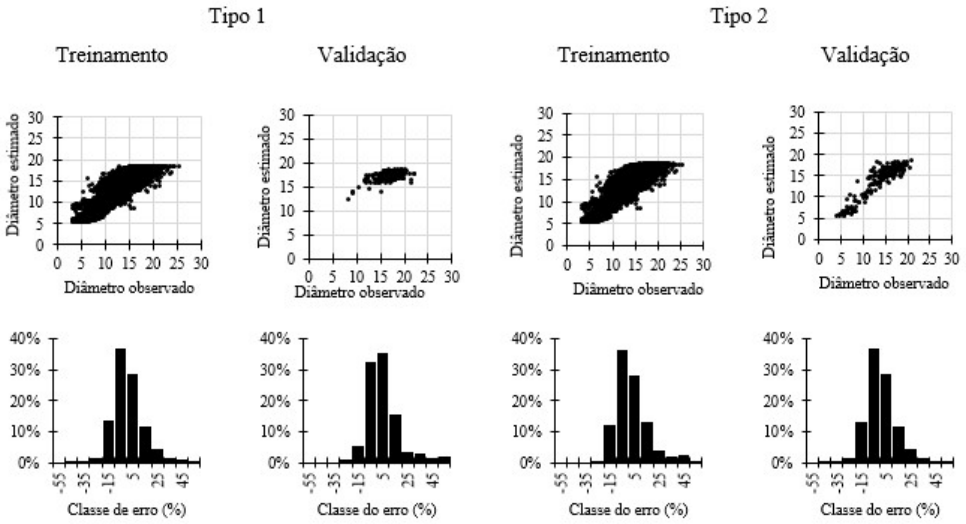

FIGURA 1: Gráficos de resíduos das estimativas diamétricas utilizando *Random Forest*
FIGURA 1**:** Residual plot of the diameter estimates using Random Forest

CONCLUSÕES

O modelo de estimativa permite a obter distribuição diamétrica de um povoamento de *Eucalyptus sp,* através do uso do algoritmo *Random Forest*. O tipo 2 utilizado nas análises se mostrou eficiente na estimação de diâmetro de árvores do povoamento.

AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Minas Gerais e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais pelo apoio técnico e financeiro.

REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAM, L. Random forests. **Mach Learn,** Massachusetts, v. 45, n. 1, p. 5-32, 2001.
INSTITUTO BRASILEIRA DE ÁRVORES. IBA: Indústria Brasileira de Árvores. Brasília, DF, 2017. 80 p. **Relatório Ibá**. 2017.
JAMES, G. et al. **An Introduction to Statistical Learning.** New York, NY: Springer New York, 2013.
JEUNE, W. et al. Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil **Classes in Western Haiti. Revista Brasileira de Ciência do Solo**,Viçosa, v. 42, n. 1, p. 1-20, 2018.

LEE, Junghee. et al. Machine Learning approaches for estimating forest stand height using plot-based observations and airborne lidar data. **Forests**,[s.l], v. 9, n. 5, p. 1-16, 2018.

MACHADO, S. A.; FIGUEIREDO, F, A. **Dendrometria**. Curitiba: Universidade Federal do Paraná, 2003.

SHAIKHINA, T. LOWE. et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. **Biomedical Signal Processing and Control**, [s.l],v. 1, n. 1, p. 1-7, 2017.

SILVA, G. F. Avaliação de métodos de medição de altura em florestas naturais. **Revista Árvore**, Viçosa, v. 36 , n. 2, p. 1-8, 2012.

SNIF - SISTEMA NACIONAL DE INFORMAÇÕES FLORESTAIS. Produção florestal. **Serviço Florestal Brasileiro**, 2016.

STROBL, C.; MALLEY, J.; TUTZ, G. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests**. Psychological Methods**, Washington, DC, v. 14, n. 4, p.323-348, 2009.

URBAZAEV, M. et al. Estimation of forest aboveground biomass and uncertainties by integration of field measurements, airborne LiDAR, and SAR and optical satellite data in Mexico. **Carbon Balance Manage**, Maryland, v. 13, n. 5, p. 1-20, 2018.

ZAMO, M. et al. Benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. **Solar Energy**,[s.1],v. 105, n. 1, p. 792-803, 2014.