UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

CÂMPUS DE CHAPADÃO DO SUL

JOÃO VICTOR SAMPAIO DA SILVA

APRENDIZADO DE MÁQUINAS POR “RANDOM FOREST” PARA A MODELAGEM DA ALTURA DE ÁRVORES DE SERINGUEIRA

Chapadão do Sul – MS

2023

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

CÂMPUS DE CHAPADÃO DO SUL

APRENDIZADO DE MÁQUINAS POR “RANDOM FOREST” PARA A MODELAGEM DA ALTURA DE ÁRVORES DE SERINGUEIRA

Trabalho de Conclusão de Curso submetido à Universidade Federal de Mato Grosso do Sul para obtenção do Grau de Bacharel em Engenharia Florestal, sob orientação do docente Dr. Gileno Brito de Azevedo.

Chapadão do Sul

2023

Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

CERTIFICADO DE APROVAÇÃO

AUTOR: **JOÃO VICTOR SAMPAIO DA SILVA.**

ORIENTADOR: **Prof. Dr. Gileno Brito de Azevedo.**

Aprovado pela Banca Examinadora como parte das exigências do Componente Curricular Não Disciplinar TCC, para obtenção do grau de BACHAREL EM ENGENHARIA FLORESTAL, pelo curso de Bacharelado em Engenharia Florestal da Universidade Federal de Mato Grosso do Sul, Câmpus de Chapadão do Sul.

Prof. Dr. Gileno Brito de Azevedo
Presidente da Banca Examinadora e Orientador

Profa. Dra. Glauce Taís de Oliveira Sousa Azevedo
Membro da Banca Examinadora

Eng. Florestal MSc. Marcos Vinícius Vieira Borges
Membro da Banca Examinadora

Chapadão do Sul, 15 de dezembro de 2023.

Documento assinado eletronicamente por **Gileno Brito de Azevedo, Professor do Magisterio Superior**, em 15/12/2023, às 09:26, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Glauce Taís de Oliveira Sousa Azevedo, Professora do Magistério Superior**, em 15/12/2023, às 09:30, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcus Vinicius Vieira Borges, Usuário Externo**, em 15/12/2023, às 09:37, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufms.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4546993** e o código CRC **4A00EF4D**.

AGRADECIMENTOS

Primeiramente agradeço a Deus por me dar a oportunidade de viver essa experiência em minha vida, onde me ensinou a ter paciência e resiliência.

Aos meus pais Julio Cesar da Silva e Luciana Martins Sampaio da Silva que me apoiaram durante a caminhada. Meu irmão Pedro Henrique Sampaio da Silva que sempre esteve comigo.

Ao meu orientador Dr. Gileno de Brito Azevedo, onde me orientou e compartilhou todo seu conhecimento e experiência.

A UFMS, CPCS e o CNPq pela estrutura de ensino proporcionada e suporte dado ao longo desses anos de graduação.

Aos meus amigos que me acompanharam durante os percalços Mateus Dias, Vitória Piccinin, Mateus Ferra, Lucas Silva, Luca Nunes, Marcus Borges, Natalia de Jesus, Guilherme de Oliveira e aos demais.

A minha banca examinadora Dr. Gileno de Brito Azevedo; Dra. Glauce Taís de Oliveira Sousa Azevedo e; Me. Marcus Vinícius Vieira Borges.

APRENDIZADO DE MÁQUINAS POR “RANDOM FOREST” PARA A MODELAGEM DA ALTURA DE ÁRVORES DE SERINGUEIRA

Resumo: A altura total das árvores (H) é uma variável fundamental, porém de difícil mensuração nos inventários florestais. Geralmente, é mensurada em apenas algumas uma amostra, enquanto para as demais árvores são estimadas por técnicas de regressão. Recentemente, os algoritmos de *machine learning*, como exemplo o Random Forest (RF), tem se mostrado promissores para gerar estimativas precisas de H. Portanto, este estudo objetivou avaliar a precisão das estimativas de H em árvores de seringueira utilizando modelos de RF e Regressão Linear (RL), treinados a partir de diferentes relações funcionais. A base de dados foi proveniente da medição de 28 parcelas permanentes distribuídas em quatro talhões de seringueira (clones RRIM600 e RRIM937), localizadas no município de Paraíso das Águas/MS. Em cada parcela, foram mensuradas as variáveis: H, em metros; diâmetro a 1,3 m do nível do solo (D), em centímetros; e altura dominante (Hd), em metros; em cinco a sete idades (I), compreendidas entre 4,3 e 16,5 anos após o plantio. O banco de dados foi subdividido de forma aleatória em: treinamento (80%) e validação (20%). Os modelos de RF e RL foram treinados considerando quatro relações funcionais: 1) $H = f(D, Hd, I)$; 2) $H = f(D, Hd)$; 3) $H = f(D, I)$ e; 4) $H = f(D)$, e posteriormente foram utilizados para gerar as estimativas de H nos dados de validação. Foi avaliada a precisão das estimativas. As relações funcionais 1 e 2 proporcionam a obtenção de estimativas precisas da altura das árvores e são superiores a relação funcional 4. Os modelos de RF, de forma geral, apresentam desempenho superior na predição de H em árvores de seringueira, indicando ser uma boa alternativa para a modelagem da relação hipsométrica.

Palavras-chave: hipsometria, inteligência artificial, relação funcional.

RANDOM FOREST MACHINE LEARNING FOR MODELING THE HEIGHT OF RUBBER TREES

Abstract: Total tree height (H) is a fundamental variable, but one that is difficult to measure in forest inventories. It is usually measured in just a few samples, while the rest of the trees are estimated using regression techniques. Recently, machine learning algorithms, such as Random Forest (RF), have shown promise in generating accurate estimates of H. Therefore, this study aimed to evaluate the accuracy of H estimates in rubber trees using RF and Linear Regression (LR) models, trained from different functional relationships. The database came from the measurement of 28 permanent plots distributed over four rubber tree plots (clones RRIM600 and RRIM937), located in the municipality of Paraíso das Águas/MS. In each plot, the following variables were measured: H, in meters; diameter at 1.3 m from ground level (D), in centimeters; and dominant height (Hd), in meters; at five to seven ages (I), between 4.3 and 16.5 years after planting. The database was randomly subdivided into training (80%) and validation (20%). The RF and RL models were trained considering four functional relationships: 1) H = f(D, Hd, I); 2) H = f(D, Hd); 3) H = f(D, I) and; 4) H = f (D), and were then used to generate estimates of H in the validation data. The accuracy of the estimates was assessed. Functional relationships 1 and 2 provide accurate estimates of tree height and are superior to functional relationship 4. The RF models generally perform better in predicting H in rubber trees, indicating that they are a good alternative for modeling the hypsometric relationship.

Keywords: hypsometry, artificial intelligence, functional relationship.

SUMÁRIO

1. INTRODUÇÃO

A *Hevea brasiliensis* (Willd. ex A. Juss.) Müll. Arg., também conhecida como seringueira, é uma espécie pertencente à família Euphorbiaceae, originária da região amazônica, com porte considerado elevado atingindo altura de até 30 metros (PIZETTA et al., 2021). A seringueira é cultivada em diversas partes do globo, principalmente no continente asiático. O produto principal de suas florestas é o látex, o qual é a matéria-prima para fabricação da borracha natural (HAN et al., 2022). Os sistemas de produção desta cultura variam desde o tradicional extrativista, até com padrões mais tecnificado visando aplicação de técnicas silviculturais e práticas de manejo afim de aumentar a produtividade (CAVALCANTE FILHO et al., 2019).

O acompanhamento do crescimento da floresta de seringueira pode fornecer informações para auxílio no manejo e aplicação de técnicas silvicuturais, visando maximizar a produtividade da área (SILVA et al., 2022). Para isso, há necessidade de estabelecimento de parcelas permanentes e o inventário florestal contínuo. No meio florestal, os inventários computam características quali e quantitativas da floresta, afim de caracterização do povoamento (ALMEIDA et al., 2021). Algumas variáveis apresentam maior complexidade de obtenção com a necessidade de adoção de técnicas de modelagem para sua estimativa.

Uma importante variável mensurada em inventários florestais é a altura total, contudo esta variável está relacionada a medidas de forma indireta, erros de medição e atraso da atividade. A altura total é indispensável para auxiliar na estimativa de outros caracteres de interesse, como o volume e classificação de índice local, bem como para acompanhamento do crescimento e auxilio de manejo (MIRANDA et al., 2021; SILVA et al., 2020; ZHOU et al., 2021). O estabelecimento de relações hisométricas, são capazes de estimar altura com base em outras variáveis como o diâmetro do tronco (NICOLETTI et al., 2020).

As equações hisométricas são amplamente estudadas, de forma que estabeleça funções que melhor predizem a variável, aumentando a confiança dos inventários (LEAL; LEAL; SILVA, 2020). Essas dependem da relação entre variáveis obtidas nos inventários e informações da floresta, como idade, e a potencialidade de explicação da altura total por estas variáveis. A inclusão de caracteres potenciais pode melhorar a precisão da estimativa, pois aumenta o nível de informação disponibilizada

e melhor classifica a variável resposta, acarretando na estimativa mais confiável (NIE; LIU, 2023).

A definição das técnicas de modelagem se torna essencial para a precisão das estimativas obtidas, uma vez que para o melhor posicionamento de decisões a avaliação quantitativa deve ser confiável. Os modelos de regressão linear são amplamente utilizados no meio florestal para a estimativa da altura (LEITE et al., 2020; VALVERDE et al., 2022). Contudo, os dados florestais são complexos e devem ser bem trabalhados afim de boas orientações, e a adoção de técnicas mais elaboradas visa auxiliar em maior confiabilidade das estimativas (LIMA et al., 2022). Portanto, os modelos de inteligencia artificial têm ganhado espaço, pois apresentam maior capacidade de generalização dos dados e relacionar os caracteres de interesse (CANTERAL et al., 2023; DE OLIVEIRA NETO et al., 2022).

Os modelos de Random Forest apresentam característica de captar o comportamento de dados complexos, melhorando a capacidade de modelagem e aumentando a precisão das estimativas (ANTONIADIS; LAMBERT-LACROIX; POGGI, 2021; COSENZA et al., 2021; GENUER; POGGI, 2020). Estes algoritmos apresentam alta capacidade de aplicação na área da engenharia florestal, para obtenção de informações mais acuradas (NUNES MIRANDA et al., 2022). Diante do exposto, este estudo teve como objetivo realizar a modelagem de predição da altura total, considerando modelos lineares e de random forest, a partir de diferentes relações funcionais em plantações de seringueira na região de Paraíso das Águas – MS.

2. MATERIAL E MÉTODOS

2.1 ÁREA DE ESTUDO

Os dados utilizados neste estudo foram obtidos em quatro talhões de seringueira, clones RRIM 600 e RIMM 937, cultivados em áreas adjacentes e condições edafoclimáticas semelhantes. As plantações encontram-se localizadas na Fazenda Promissão, no município de Paraíso das Águas, no estado de Mato Grosso do Sul, em coordenadas geográficas de 19°03'08" S de latitude e 52°58'06" N de longitude, com uma altitude média de 600 metros.

O clima da região, conforme a classificação de Köppen e Geiger, é tropical do tipo Aw, caracterizado por estações bem definidas, com um período chuvoso durante o verão e seco no inverno (ALVARES et al., 2013). A área apresenta temperatura média anual de 23,6º, e pluviosidade média anual de 1.549 mm. O solo do local é classificado como Neossolo quartzarênico (SANTOS et al., 2014), com as seguintes características químicas na profundidade de 0-20 cm: pH ($CaCl_2$) = 4,7; Al (cmolc dm^{-3}) = 0,24; Ca (cmolc dm^{-3}) = 0,75; Mg (cmolc dm^{-3}) = 0,10; P (mg dm^{-3}) = 5,6; K (mg dm^{-3}) = 14; Capacidade de troca catiônica (ou CTC) (cmolc) = 2,9; Saturação por bases (%) = 30,7 (OLIVEIRA et al., 2018).

2.2 BASE DE DADOS

Em cada talhão, foram estabelecidas aleatoriamente sete parcelas permanentes (total de 28 parcelas), compostas pela área útil de 36 árvores (3 linhas de plantio x 12 árvores em cada linha). As unidade de amostra foram remedidas em 5 ou sete ocasiões, em idades que variaram de 4,3 a 16,5 anos (Tabela 1).

Tabela 1. Características silviculturais e idades de medição (anos após o plantio) em cada um dos talhões de seringueirae m Paraíso das Águas/MS.

Talhão	Clone	Espaçamento de plantio	Área parcela (m^2)	Nº de medições	Primeira medição	Última medição
T1	RRIM 600	7,0 x 2,5 m	630	5	13,3	16,5
T2	RRIM 600	7,0 x 2,5 m	630	5	11,3	14,5
T3	RRIM 600	7,0 x 2,7 m	680,4	7	4,3	9,5
T4	RRIM 937	7,0 x 2,7 m	680,4	7	4,3	9,5

Em cada medição, foram mensuradas as variáveis: 1) circunferência na altura de 1,3 m do nível do solo (C), em centimetros, em todas as ávores da parcela, com auxílio de uma fita métrica; 2) a variável C foi convertida em diâmetro (D), a partir da divisão de seu valor por π; 3) altura das árvores (H), em metros, nas árvores localizadas na linha central e para as árvores dominantes, utilizando um clinômetro Haglof. Foram consideradas como árvores dominantes as sete árvores de maior C em cada parcela-idade, e altura dominante (Hd), em metros, foi obtida a partir da média aritimética de H destas árvores, conforme o conceito de Assman (1970).

2.2 ANÁLISE DE DADOS

Inicialmente, com objetivo de compreender as relações entre a variável H e as variáveis explicativas a serem utilizadas nos modelos de RF e RL, foi analisada a relação entre pares de variáveis formadas por H, D, Hd e I, em cada um dos talhões. Foi gerada uma matriz gráfica com boxplots, gráficos de dispersão, gráficos densidade e correlação de Pearson entre as variáveis, com auxílio da função “ggpairs” do pacote “*GGally*” do software R (SCHLOERKE et al., 2021).

Para realizar a modelagem, o banco de dados com o total de 2.904 observações foi subividido, de forma aleatória, em dois conjuntos: o primeiro com 80% das observações foi destinado para a etapa de treinamento dos modelos de RF e RL, e os 20% restantes para a etapa de validação.

Para o treinamento dos modelos de RF e RL foram utilizadas diferentes relações funcionais, que configuram quatro combinações de variáveis explicativas nos modelos para a predição de H. As relações funcionais utilizadas foram: 1) *H = f(D, Hd, I)*; 2) *H = f(D, Hd)*; 3) *H = f(D, I)* e; 4) *H = f (D)*; em que H = altura total (m); D = diâmetro à altura de 1,3 m do nível do solo (cm); Hd = altura dominante (m); I = idade (anos).

O treinamento dos modelos de RF foi realizado com auxílio do pacote "*randomForest*" do software R (LIAW e WIENER, 2002). Todos os modelos de RF treinados foram do tipo “*regression*”, com “*number of trees* = 1000” e “*number of variables tried at each split* = 1”. Para realizar o treinamento dos modelos de RL foi realizada a transformação logarítmica das variáveis a serem utilizadas no modelo e foram ajustados modelos de regressão linear do tipo $Ln(H) = \beta_0 + \beta_1 * Ln(X_1) + ... + \beta_n * Ln(X_n) + \varepsilon$, em que: Ln: Logaritmo neperiano; β_0 a β_n = coeficientes dos modelos; H = altura total (m); X_1 a X_n = variáveis explicativas que compõe as relações funcionais avaliadas e; ε = erro aleatório. O ajuste dos modelos de RL foi realizado com auxílio da função “*lm*”, disponível no software R (R CORE TEAM, 2023).

Os modelos de RF e RL treinados foram utilizados para gerar as estimativas de H nos dados de validação. Com base nos vetores formados pelos dados de H observados e estimados, foram obtidas métricas que permitem inferir sobre a capacidade preditiva e precisão das estimativas geradas por cada um dos modelos, permitindo indicar os modelos de melhor desempenho. As métricas utilizadas foram: coeficiente de correlação de Pearson (r); raiz quadrada do erro médio (rmse); erro médio absoluto (mae); e bias. Os valores de rmse, mae e bias foram obtidos com

auxílio do pacote "*Metrics*" do *software* R (HAMNER e FRASCO, 2018) e foram convertidos em porcentagem a partir da divisão dos valores obtidos pela média observada de H e multiplicação por 100. Também foi realizada a análise gráfica dos erros (H observada – H estimada) em porcentagem, através de gráficos de dispersão e boxplots, com auxílio do pacote "ggplot2" do *software* R (WICKHAM, 2016).

De forma complementar, com o objetivo de avaliar a estabilidade nas estimativas de H geradas pelos modelos de RF e RL a partir das diferentes relações funcionais, o banco de dados utilizado no treinamento (T), com 2303 observações, foi subdividido de forma aleatória em 5 subconjuntos com 464/465 observações cada (S_1 a S_5). Estes foram utilizados para realizar um novo ajuste dos modelos de RF e RL, com as diferentes relações funcionais, perfazendo um total de cinco repetições. Na sequência, os modelos obtidos foram utilizados para estimar os valores de H nos dados de validação e calcular as métricas: r, rmse, mae e bias. Os valores dessas métricas obtidos para cada uma das cinco repetições foram submetidos a ANOVA ($\alpha = 0,05$), considerando o esquema fatorial 2 x 4 (modelos x relações funcionais), com o delineamento inteiramente casualizados, com cinco repetições. Em caso de diferenças significativas, as médias foram comparadas pelo teste de Tukey ($\alpha = 0,05$).

3. RESULTADOS E DISCUSSÃO

A altura das árvores (H) de seringueira apresentou correlação linear positiva e significativa com as variáveis diâmetro (D), altura dominante (Hd) e idade (I) (Figura 1). A correlação entre as variáveis foi influenciada pelos talhões e os maiores valores de correlação de H com as demais variáveis foram registrados nos talhões mais jovens (T3 e T4). As maiores correlações com H também foram diferentes entre os talhões: T1: Hd; T2: D; T3 e T4: D e Hd. Por outro lado, se considerado o banco de dados como um todo, composto por talhões com diferentes idades de plantio, as maiores correlações foram observadas na seguinte ordem: Hd > I > D.

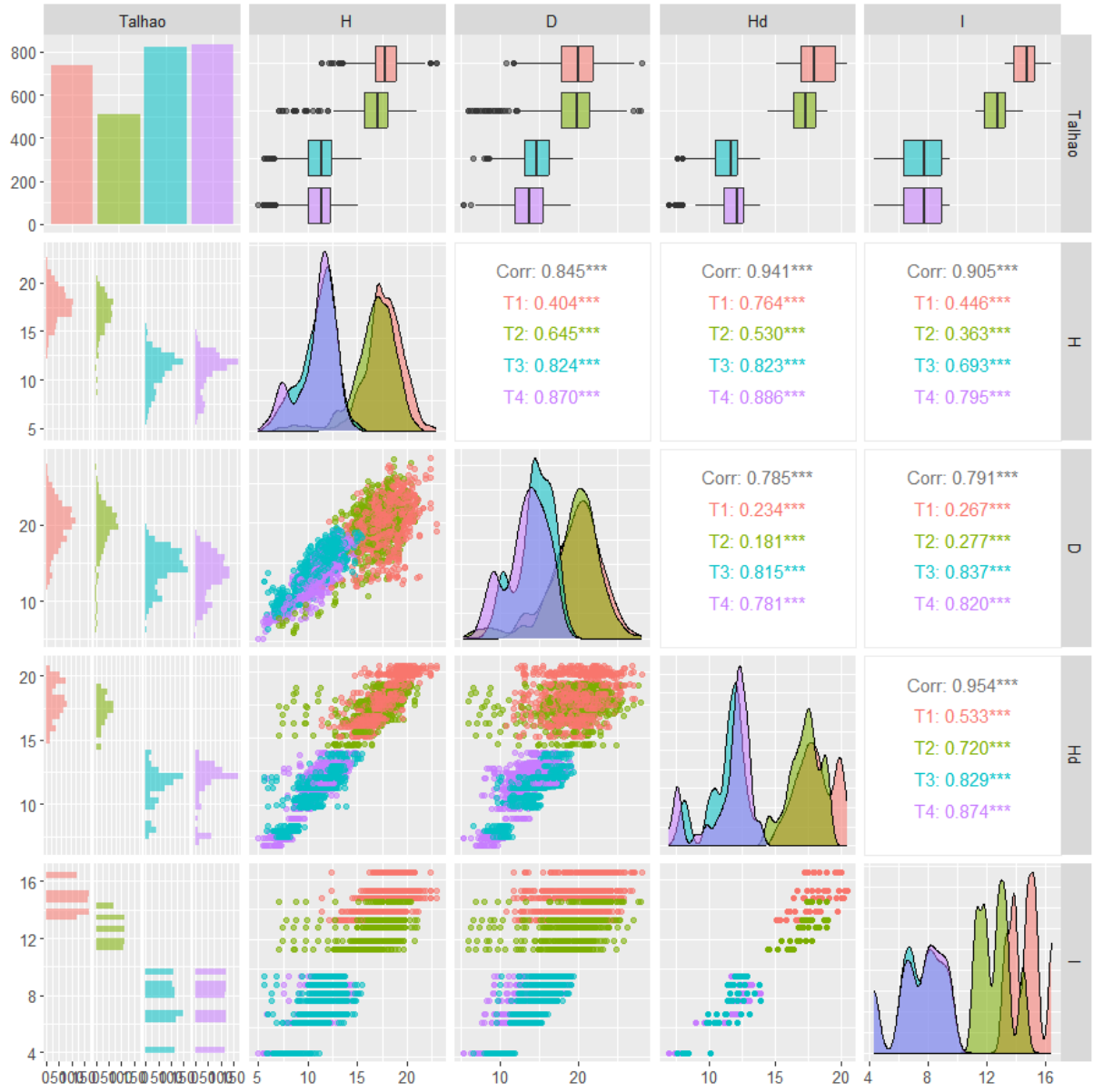

Figura 1. Correlação entre as variáveis dendrométricas em plantações de seringueira. H = altura total (m); D = diâmetro à 1,30 m de altura (cm); Hd = altura dominante (m); I = idade (anos). T1 a T4 = talhões.

A forte correlação das variáveis independentes (Hd, D e I) com a dependente (H), indica que estas podem contribuir para explicar a variação H, com a obtenção de modelos que proporcionam estimativas precisas dessa variável (SILVA et al., 2017). O D é a variável mais comumente utilizada nas relações hipsométricas, sendo muitas vezes a única variável incluída no modelo (BINOTI et al., 2018; CASAS et al., 2022; SOARES et al., 2021).

A inclusão das variáveis Hd e I nos modelos contribuiu para a melhoria da precisão nas estimativas de H (Tabelas 2 e 3). Esses resultados podem explicados pelas altas correlações de Hd e I com H (Figura 1), indicando que a inclusão dessas variáveis melhora a capacidade preditiva dos modelos.

Tabela 2. Métricas de precisão obtidas na validação dos modelos de Random Forest (F) para a projeção da altura total (H – em metros) de árvores de seringueira em Paraíso das Águas/MS.

Modelo	Métricas de precisão			
	r	rmse (%)	bias (%)	mae (%)
RF1	0,9698	6,53	0,22	4,93
RF2	0,9685	6,64	0,18	5,07
RF3	0,9478	8,52	0,35	6,53
RF4	0,8154	15,64	1,35	11,36

Em que: RF1: *H = f(D, Hd, I)*; RF2: *H = f(D, Hd)*; RF3: *H = f(D, I)*; RF4: *H = f (D)*; H = altura total (m); D = diâmetro à altura de 1,3 m do nível do solo (cm); Hd = altura dominante (m); I = idade (anos); r = coeficiente de correlação de Pearson (r); rmse = raiz quadrada do erro médio; mae = erro médio absoluto; bias = bias ou viés.

Tabela 3. Métricas de precisão obtidas na validação dos modelos de regressão linear (RL) para a projeção da altura total (H – em metros) de árvores seringueira em Paraíso das Águas/MS.

Modelo	Coeficientes				Métricas de precisão			
	β_0	β_1	β_2	β_3	r	rmse (%)	bias (%)	mae (%)
RL1	-0,4617	0,3842	0,8350	-0,0914	0,9574	7,77	0,82	5,77
RL2	-0,3268	0,3686	0,7210	-	0,9576	7,75	0,82	5,80
RL3	0,3897	0,4353	0,4362	-	0,9270	10,08	1,08	7,82
RL4	-0,0284	0,9418	-	-	0,8251	15,23	1,96	11,23

Em que: RL1: $Ln(H) = \beta_0 + \beta_1 * Ln(D) + \beta_2 * Ln(Hd) + \beta_3 * Ln(I) + \varepsilon$; RL2: $Ln(H) = \beta_0 + \beta_1 * Ln(D) + \beta_2 * Ln(Hd) + \varepsilon$; RL3: $Ln(H) = \beta_0 + \beta_1 * Ln(D) + \beta_3 * Ln(I) + \varepsilon$; RL4: $Ln(H) = \beta_0 + \beta_1 * Ln(D) + \varepsilon$; Ln = Logaritmo neperiano; β_0, β_1, β_2, β_3 = coeficientes dos modelos; H = altura total (m); D = diâmetro à altura do peito (cm); Hd = altura dominante (m); I = Idade (anos); ε = Erro residual; r = coeficiente de correlação de Pearson (r); rmse =

raiz quadrada do erro médio; mae = erro médio absoluto; bias = Bias ou viés.

De maneira geral, os modelos de RF e RL treinados com mais de uma variável explicativa proporcionaram melhor desempenho nas estimativas de H (r>0,90; RMSE<10%; MAE<10%; BIAS<1%), enquanto os modelos que utilizaram a relação funcional *H = f (D)* proporcionaram menor precisão nas estimativas. Essas estatísticas são capazes de mensurar a precisão dos modelos, e são amplamente utilizadas na literatura voltada a área de ciência florestal (BATISTA et al., 2022; DE AZEVEDO et al., 2020; SANTANA et al., 2023). Assim, os resultados do presente estudo ressaltam a importância de incluir variáveis para melhorar a capacidade preditiva dos modelos, contribuindo para melhorar a precisão nos inventários florestais. A inclusão dessas variáveis impacta em melhor classificação da variável dependente, acarretando em estimativas mais precisas (ACOSTA et al., 2020; COSTA FILHO et al., 2019; SCHMITT; ANDRADE, 2019).

Quando comparadas as métricas de precisão obtidas pelos modelos de RF e RL (Tabelas 2 e 3), RF proporcionou maior precisão nas estimativas de quando utilizadas as relações funcionais *H = f(D, Hd, I)*; *H = f(D, Hd)* e *H = f(D, I)*, enquanto modelo RL foi ligeiramente superior para a relação *H = f (D).* Os modelos de regressão linear tradicionais apresentam uma boa aceitação para as relações hipsométricas (MARTINS et al., 2020; SANTOS et al., 2019). Contudo, de acordo com Carvalho et al. (2019) os modelos de RF são capazes de sintetizar melhor os dados e com maior capacidade de predição. As funções de inteligência artificial têm ganhado espaço na modelagem florestal, pois apresentam boa capacidade de relacionar os caracteres dependentes e independentes (BORGES et al., 2022; SOARES et al., 2013; VIEIRA et al., 2022).

Os modelos testados proporcionaram estimativas precisas de H, sem a presença padrões de tendências nas estimativas, com a maior distribuição dos erros em torno de zero para os valores observados de H no banco de dados como um todo (Figura 3). Por outro lado, quando os erros foram avaliados separadamente para os talhões (Figura 4), verifica-se que alguns modelos proporcionaram estimativas ligeiramente tendenciosas, com aumento da distância da média e mediana dos erros em relação ao erro zero. De forma geral, o melhor desempenho das estimativas foi obtido para as relações funcionais *H = f(D, Hd, I)* e *H = f(D, Hd)*, enquanto os modelos

de RF proporcionaram estimativas mais precisas dos que os RL.

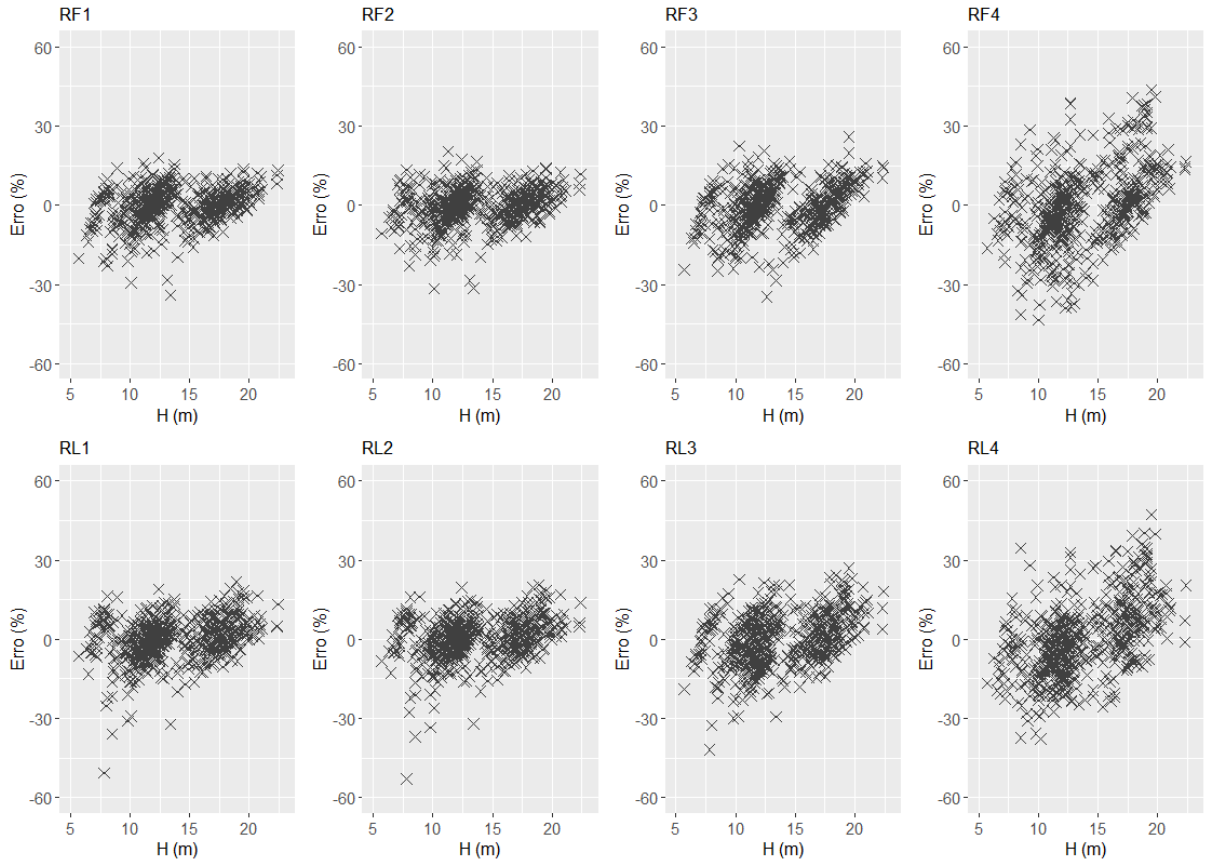

Figura 2. Distribuição dos erros (%) na estimativa da altura de árvores de seringueira a partir de modelos de Random Forest (RF) e Regressão Linear (RL). Os números de 1 a 4 que acompanham os prefixos RF e RL indicam as relações funcionais utilizadas no treinamento dos modelos: 1: *H = f(D, Hd, I)*; 2: *H = f(D, Hd)*; 3: *H = f(D, I)*; 4: *H = f (D)*.

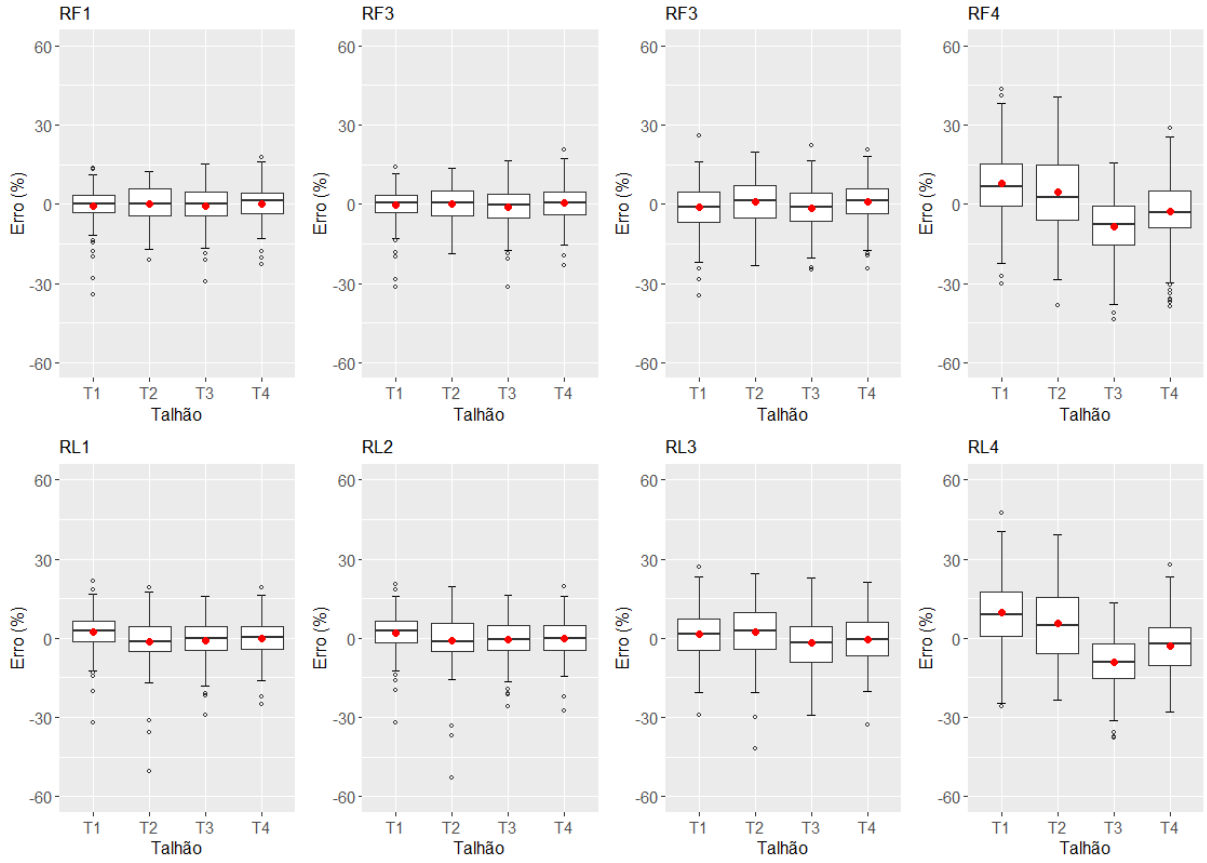

Figura 3. Distribuição dos erros (%) na estimativa da altura de árvores em talhões de seringueira (T1 a T4) a partir de modelos de Random Forest (RF) e Regressão Linear (RL). Os números de 1 a 4 que acompanham os prefixos RF e RL indicam as relações funcionais utilizadas no treinamento dos modelos: 1: *H = f(D, Hd, I)*; 2: *H = f(D, Hd)*; 3: *H = f(D, I)*; 4: *H = f (D)*.

Conforme mencionado por Monteiro et al. (2021), a análise gráfica dos resíduos é primordial para captar o comportamento dos dados estimados em relação aos observados, via erro relativo. A baixa dispersão dos resíduos relacionados aos valores das métricas de precisão em dados de validação alcançadas no presente trabalho, podem ser considerados satisfatórios para predição, pois na área de ciência florestal valores como RMSE<10 e r>0,90 indica boa precisão dos modelos (FIANDINO et al., 2020; HAMIDI et al., 2021; REZAALI et al., 2021).

A tendenciosidade de modelos é um fator limitante à precisão, pois funções tendenciosas podem super ou subestimar variáveis, impactando negativamente na plenitude dos dados (MARTINS et al., 2017). Quando os erros relativos estão mais dispersos a reta zero, a capacidade de estimativa dos modelos treinados é viesada,

impactando na exatidão das variáveis preditas (DAI et al., 2021). O estudo de relações e modelos de predição aplicados a florestas, impactam positivamente na disponibilização de funções com ótimo ajuste para aplicação em campo.

Em acordo com o observado nos resultados anteriores, a ANOVA indicou que a precisão das estimativas de H foi significativamente influenciada pelos modelos e pelas relações funcionais avaliadas. Para os valores de r, rmse e mae houve interação entre os modelos e as relações funcionais ($p < 0,05$). De forma geral, quando comparado os modelos, RF apresentou melhor desempenho quando utilizada as relações funcionais de *H = f(D, Hd, I)*, *H = f(D, Hd)* e *H = f(D, I)*, enquanto o modelo de RL foi superior para *H = f(D)*. Quando comparadas as relações funcionais, tanto RF quanto RL proporcionaram as melhores estimativas para *H = f(D, Hd, I)* e *H = f(D, Hd)* (Figura 4). Para a métrica bias não houve interação significativa entre os modelos e relações funcionais. A análise dos fatores isolados indica os maiores valores dessa métrica para os modelos de RL e para a relação funcional *H = f(D)*.

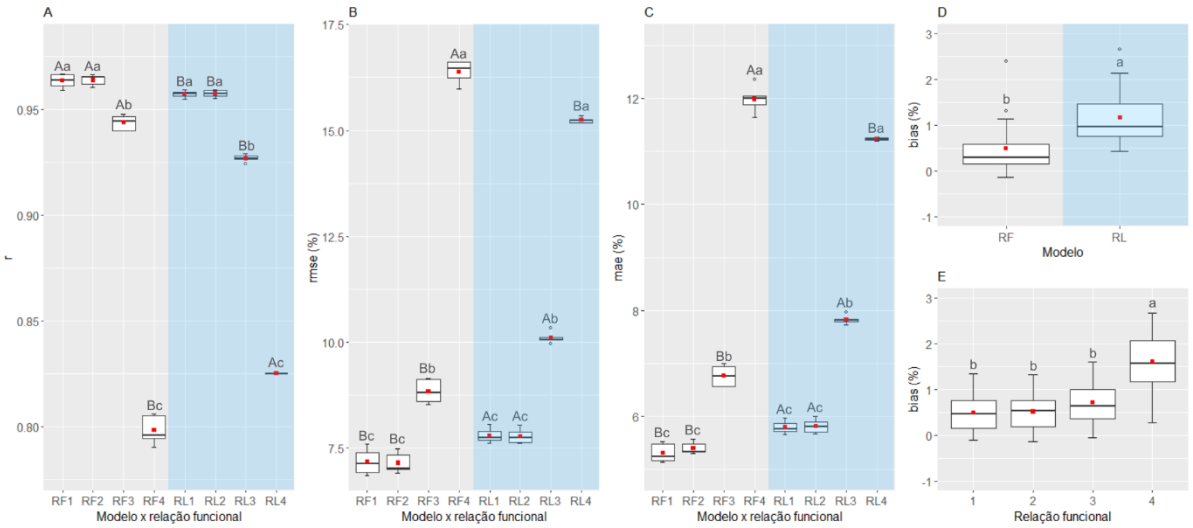

Figura 4. Métricas de precisão para a estimativas da altura total (H) de árvores de seringueira a partir de modelos de Random Forest (RF) e Regressão Linear (RL). Os números de 1 a 4 que acompanham os prefixos RF e RL indicam as relações funcionais utilizadas no treinamento dos modelos: 1: *H = f(D, Hd, I)*; 2: *H = f(D, Hd)*; 3: *H = f(D, I)*; 4: *H = f (D)*. r = coeficiente de correlação de Pearson (r); rmse = raiz quadrada do erro médio; mae = erro médio absoluto; bias = bias ou viés. Letras maiúsculas comparam os modelos RF e RL para uma mesma relação funcional e letras minúsculas comparam as relações funcionais em cada modelo.

Os modelos treinados neste trabalho, de forma geral apresentaram boas estatísticas de precisão e a inclusão de variáveis explicativas nos modelos contribuiu para aumentar a capacidade preditiva. Quanto ao tipo de modelo, RF proporcionou desempenho superior na predição de altura total de seringueira, indicando ser uma boa alternativa para a modelagem da relação hipsométrica. Quanto as relações funcionais, a utilização de outras variáveis explicativas em conjunto com o diâmetro, contribuiu com a melhoria da capacidade preditiva dos modelos. Apesar do desempenho semelhante das relações funcionais *H = f(D, Hd, I)* e *H = f(D, Hd)*, a relação *H = f(D, Hd)* apresenta vantagens do ponto de vista prático, uma vez que necessita de uma variável a menos, contribuindo para a maior facilidade na obtenção dos dados, além de tornar os modelos menos complexos.

Em inventários florestais a altura total de plantas é uma variável com maior complexidade de mensuração, relacionada a maiores erros. Com isso, a necessidade de estudos que captam característica de modelos capazes de estimar com maior capacidade de predição essa variável, auxiliam na precisão dos inventários florestais. As relações hipsométricas é de grande importância para os projetos de quantificação da produtividade, desta forma as funções devem ser melhores escolhidas e testadas no conjunto de dados para maior sucesso (DANTAS et al., 2020; DA SILVA et al., 2021).

4. CONCLUSÕES

Os modelos de Random Forest, de forma geral, apresentam desempenho superior na predição de altura total de árvores de seringueira, indicando ser uma boa alternativa para a modelagem da relação hipsométrica.

As relações funcionais *H = f(D, Hd, I)* e *H = f(D, Hd)* proporcionam a obtenção de estimativas precisas da altura das árvores e são superiores a relação funcional *H = f(D).*

5. REFERÊNCIAS

ACOSTA, H. A. B. et al. Identidade de modelos hipsométricos para clones de eucalipto na região oriental do Paraguai. **Scientia Forestalis**, v. 48, n. 125, 31 mar. 2020.

ALMEIDA, A. et al. Individual Tree Detection and Qualitative Inventory of a Eucalyptus sp. Stand Using UAV Photogrammetry Data. **Remote Sensing**, v. 13, n. 18, p. 3655, 13 set. 2021.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 1 dez. 2013.

ANTONIADIS, A.; LAMBERT-LACROIX, S.; POGGI, J.-M. Random forests for global sensitivity analysis: A selective review. **Reliability Engineering & System Safety**, v. 206, p. 107312, fev. 2021.

ASSMANN, E. **The principles of forest yield study**. New York: Perfamon Press, p. 506, 1970.

BATISTA, T. S. et al. Artificial neural networks and non-linear regression for quantifying the wood volume in Eucalyptus species. **Southern Forests: a Journal of Forest Science**, v. 84, n. 1, p. 1–7, 2 jan. 2022.

BINOTI, D. H. B. et al. ESTIMATION OF HEIGHT OF EUCALYPTUS TREES WITH NEUROEVOLUTION OF AUGMENTING TOPOLOGIES (NEAT). **Revista Árvore**, v. 41, n. 3, 22 fev. 2018.

BORGES, M. V. V. et al. High-throughput phenotyping of two plant-size traits of Eucalyptus species using neural networks. **Journal of Forestry Research**, v. 33, n. 2, p. 591–599, 3 abr. 2022.

CANTERAL, K. F. F. et al. Machine learning for prediction of soil CO2 emission in tropical forests in the Brazilian Cerrado. **Environmental Science and Pollution Research**, v. 30, n. 21, p. 61052–61071, 12 abr. 2023.

CARVALHO, M. C. et al. ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA MODELAGEM DA DISTRIBUIÇÃO POTENCIAL DE HABITATS DE ESPÉCIES ARBÓREAS. **Nativa**, v. 7, n. 5, p. 600, 12 set. 2019.

CASAS, G. G. et al. Configuration of the Deep Neural Network Hyperparameters for the Hypsometric Modeling of the Guazuma crinita Mart. in the Peruvian Amazon. **Forests**, v. 13, n. 5, p. 697, 29 abr. 2022.

CAVALCANTE FILHO, P. G. et al. Dinâmica inovativa e investimento na reserva extrativista Chico Mendes. **Brazilian Journal of Development**, v. 5, n. 8, p. 13358–13382, 2019.

COSENZA, D. N. et al. Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. **Forestry: An International Journal of Forest Research**, v. 94, n. 2, p. 311–323, 4 mar. 2021.

COSTA FILHO, S. V. S. DA et al. Configuração de algoritmos de aprendizado de máquina na modelagem florestal: um estudo de caso na modelagem da relação hipsométrica. **Ciência Florestal**, v. 29, n. 4, p. 1501–1515, 10 dez. 2019.

DAI, P. V. S. et al. Estimativa de volume de madeira baseada em índices de vegetação. **Scientia Forestalis**, v. 49, n. 129, 1 mar. 2021.

DANTAS, D. et al. Reduction of sampling intensity in forest inventories to estimate the total height of eucalyptus trees. **Bosque (Valdivia)**, v. 41, n. 3, p. 353–364, dez. 2020.

DA SILVA, A. K. V. et al. Predicting Eucalyptus Diameter at Breast Height and Total Height with UAV-Based Spectral Indices and Machine Learning. **Forests**, v. 12, n. 5, p. 582, 7 maio 2021.

DE AZEVEDO, G. B. et al. Multi-volume modeling of Eucalyptus trees using regression and artificial neural networks. **PLOS ONE**, v. 15, n. 9, p. e0238703, 11 set. 2020.

DE OLIVEIRA NETO, R. R. et al. Estimation of Eucalyptus productivity using efficient artificial neural network. **European Journal of Forest Research**, v. 141, n. 1, p. 129–151, 27 fev. 2022.

FIANDINO, S. et al. Modeling forest site productivity using climate data and topographic imagery in Pinus elliottii plantations of central Argentina. **Annals of Forest Science**, v. 77, n. 4, p. 95, 7 dez. 2020.

GENUER, R.; POGGI, J.-M; TULEAU, C. **Random Forests: some methodological insights**, v.1, n. 6729 p. 32, 2008.

HAMIDI, S. K. et al. Analysis of plot-level volume increment models developed from machine learning methods applied to an uneven-aged mixed forest. **Annals of Forest Science**, v. 78, n. 1, p. 4, 12 mar. 2021.

HAMNER, B.; FRASCO, M. **Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4**, 2018.

HAN, Q. et al. Development and characterization of microsatellite markers for the rubber tree powdery mildew pathogen Oidium heveae. **European Journal of Plant Pathology**, v. 164, n. 2, p. 253–262, 18 out. 2022.

LEAL, F. A.; LEAL, G. DA S. A.; SILVA, T. C. DA. Redes neurais artificiais e modelos alométricos aplicados para estimativa de volume e altura em Eucalyptus urophylla S.T.Blacke. **Advances in Forestry Science**, v. 7, n. 3, p. 1181–1188, 3 nov. 2020.

LEITE, R. V. et al. Estimating Stem Volume in Eucalyptus Plantations Using Airborne LiDAR: A Comparison of Area- and Individual Tree-Based Approaches. **Remote Sensing**, v. 12, n. 9, p. 1513, 9 maio 2020.

LIAW. A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 3, n. 3, p. 18-22, 2022.

LIMA, E. DE S. et al. Random forest model to predict the height of eucalyptus. **Engenharia Agrícola**, v. 42, 2022.

MARTINS, A. P. M. et al. Estimativa do Afilamento do Fuste de Araucária Utilizando Técnicas de Inteligência Artificial. **Floresta e Ambiente**, v. 24, n. 0, 2017.

MARTINS, M. et al. RELAÇÃO HIPSOMÉTRICA DE TRÊS ESPÉCIES DA CAATINGA, SEMIÁRIDO PERNAMBUCO. **Agrarian Academy**, v. 7, n. 13, 30 jul. 2020.

MIRANDA, R. O. V. DE et al. Métodos da curva guia e equação das diferenças na classificação de sítio e sua relação na descrição da altura em Pinus taeda L. **Scientia Forestalis**, v. 49, n. 131, 1 set. 2021.

MONTEIRO, B. C. et al. Uso de modelos mistos para estimativa do volume de árvores individuais em tipologias florestais no Estado do Amapá. **BIOTA AMAZÔNIA**, v. 11, p. 1–4, 2021.

NICOLETTI, M. F. et al. Equações hipsométricas, volumétricas e funções de afilamento para Pinus spp. **Revista de Ciências Agroveterinárias**, v. 19, n. 4, p. 474–482, 14 dez. 2020.

NIE, J.; LIU, S. Incorporated neighborhood and environmental effects to model individual-tree height using random forest regression. **Scandinavian Journal of Forest Research**, v. 38, n. 4, p. 221–231, 19 maio 2023.

NUNES MIRANDA, E. et al. Variable selection for estimating individual tree height using genetic algorithm and random forest. **Forest Ecology and Management**, v. 504, p. 119828, jan. 2022.

OLIVEIRA, V. H. S. DE et al. Initial development and sample dimensioning of rubber tree clones. **Bioscience Journal**, v. 34, n.5, p. 1225–1231, 2018.

PIZETTA, M. et al. Fusariosis in rubber tree: pathogenic, morphological, and molecular characterization of the causal agent. **European Journal of Plant Pathology**, v. 161, n. 4, p. 769–782, 23 dez. 2021.

R Core Team. R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2023.

REZAALI, M. et al. A wavelet-based random forest approach for indoor BTEX spatiotemporal modeling and health risk assessment. **Environmental Science and Pollution Research**, v. 28, n. 18, p. 22522–22535, 9 maio 2021.

SANTANA, D. C. et al. Machine Learning Methods for Woody Volume Prediction in Eucalyptus. **Sustainability**, v. 15, n. 14, p. 10968, 13 jul. 2023.

SANTOS, H. G. DOS. **Sistema brasileiro de classificação de solos**. [s.l: s.n.].

SANTOS, F. M. et al. Modeling the height–diameter relationship and volume of young African mahoganies established in successional agroforestry systems in northeastern Brazil. **New Forests**, v. 50, n. 3, p. 389–407, 21 maio 2019.

SCHLOERKE, B.; COOK, D.; LARMARANGE, J.; BRIATTE, F.; MARBACH, M.; THOEN, E.; ELBERG, A.; CROWLEY, J. **GGally: Extension to 'ggplot2'**. R package version 2.1.2, 2021.

SCHMITT, T.; ANDRADE, V. C. L. DE. Identidade de modelos hipsométricos para um plantio de eucalipto clonal. **Advances in Forestry Science**, v. 6, n. 2, 24 jul. 2019.

SILVA, A. V. DOS S. et al. Classificação de sítio em plantio florestal de eucalipto no estado do Amapá. **Revista Arquivos Científicos (IMMES)**, v. 3, n. 1, p. 106–110, 23 jun. 2020.

SILVA, C. A. et al. Predição da biomassa aérea em plantações de Pinus taeda L. por meio de dados LiDAR aerotransportado. **Scientia Forestalis**, v. 45, n. 115, 1 set. 2017.

SILVA, E. V. DA et al. Crescimento de clones de Hevea brasiliensis; sob doses crescentes de nitrogênio, fósforo e potássio. **Ciência Florestal**, v. 32, n. 4, p. 1964–1979, 23 nov. 2022.

SOARES, F. A. A. M. N. et al. Recursive diameter prediction for calculating merchantable volume of eucalyptus clones using Multilayer Perceptron. **Neural Computing and Applications**, v. 22, n. 7–8, p. 1407–1418, 19 jun. 2013.

SOARES, G. M. et al. ARTIFICIAL NEURAL NETWORKS (ANN) FOR HEIGHT ESTIMATION IN A MIXED-SPECIES PLANTATION OF Eucalyptus globulus LABILL AND Acacia mearnsii DE WILD. **Revista Árvore**, v. 45, p. 4512 2021.

VALVERDE, J. C. et al. Taper and individual tree volume equations of Eucalyptus varieties under contrasting irrigation regimes. **New Zealand Journal of Forestry Science**, v. 52, 24 maio 2022.

VIEIRA, T. A. S. et al. Production of high-quality forest wood biomass using artificial intelligence to control thermal modification. **Biomass Conversion and Biorefinery**, p. 1-17, 26 abr. 2022.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.

ZHOU, H. et al. Research on volume prediction of single tree canopy based on three-dimensional (3D) LiDAR and clustering segmentation. **International Journal of Remote Sensing**, v. 42, n. 2, p. 738–755, 17 jan. 2021.