UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CÂMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

LUIZ FELIPE DOS SANTOS SILVA

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EXPLICATIVAS PARA A ESTIMATIVA DA ALTURA DE ÁRVORES DE EUCALIPTO

CHAPADÃO DO SUL – MS
2022

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CÂMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

LUIZ FELIPE DOS SANTOS SILVA

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EXPLICATIVAS PARA A ESTIMATIVA DA ALTURA DE ÁRVORES DE EUCALIPTO

Orientador: Prof. Dr. Gileno Brito de Azevedo

Dissertação apresentada à Universidade Federal de Mato Grosso do Sul, como parte dos requisitos para obtenção do título de Mestre em Agronomia, área de concentração: Produção Vegetal.

CHAPADÃO DO SUL – MS
2022

Serviço Público Federal

Ministério da Educação

Fundação Universidade Federal de Mato Grosso do Sul

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Luiz Felipe dos Santos Silva

ORIENTADOR: Dr. Gileno Brito de Azevedo

TÍTULO: Métodos de seleção de variáveis explicativas para a estimativa da altura de árvores de eucalipto

AVALIADORES:

Prof. Dr. Gileno Brito de Azevedo

Prof. Dr. Paulo Eduardo Teodoro

Profa. Dra. Glauce Tais de Oliveira Sousa Azevedo

Chapadão do Sul, 01 de novembro de 2022.

Documento assinado eletronicamente por **Paulo Eduardo Teodoro**, **Professor do Magisterio Superior**, em 01/11/2022, às 09:14, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Gileno Brito de Azevedo**, **Professor do Magisterio Superior**, em 01/11/2022, às 09:14, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Glauce Tais de Oliveira Sousa Azevedo**, **Professora do Magistério Superior**, em 01/11/2022, às 09:14, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

DEDICATÓRIA

Aos meus pais, por não terem medido esforços em me apoiar em toda minha trajetória acadêmica e por terem sempre acreditado em mim.

DEDICO

AGRADECIMENTOS

A Deus por abençoar os meus passos e ter me trazido até aqui, sendo minha luz e força nos momentos mais difíceis desta caminhada.

Aos meus pais Anildo da Silva e Eloni dos Santos Silva, por estarem sempre ao meu lado independente de tudo, acreditando nos meus sonhos e fazendo de tudo para que eu possa realizá-los. Muito obrigada por serem meus maiores exemplos de vida, minhas riquezas e a quem devo todo meu amor e respeito.

A minha irmã Fernanda C. dos S. Silva, por ter vindo ao mundo alegrar os meus dias e pela pessoa excepcional que é, exemplo de ser humano, e mãe do sobrinho mais bonito do mundo.

A toda minha família por torcerem por mim e contribuírem de alguma forma com a realização deste sonho.

Ao meu orientador Gileno Brito de Azevedo, por ser um excelente profissional com o qual aprendi ensinamentos valiosos, e um ser humano generoso que me apoiou nessa jornada, acreditando em mim, me fazendo persistir e chegar até aqui.

Aos meus amigos de infância Luciano Alves Rosa e Victor Hugo Barbosa Ribeiro por estarem comigo independente do tempo e da distância.

Aos amigos conquistados ao longo dos anos de vida acadêmica e profissional, em especial a Patrícia Covo Carvalho, Amanda Amorim da Silva, Emanuella Chagas, Cosme O. do Nascimento, Jorgielly Ávila, Victoria Toledo, Osane Alves, Ana Paula Nunes, Gustavo Henrique, Iryana Viana, entre tantos outros que me ajudaram nesta caminhada.

A Universidade Federal do Mato Grosso do Sul, campus Chapadão do Sul, e a todos os professores da graduação e do programa de pós-graduação em Agronomia que acrescentaram em minha formação, em especial aos professores Ana Paula Leite de Lima e Sebastião Ferreira de Lima, que tanto me ajudaram e ensinaram, como pessoa e futuro profissional. Muito obrigado pela paciência, confiança e por ter me acolhido durante a graduação.

Aos prefossores Glauce Taís de Oliveira Sousa Azevedo e Paulo Eduardo Teodoro, por terem aceito o convite para a banca examinadora e por contribuir com a sua experiência na conclusão deste trabalho.

Enfim, agradeço a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

EPÍGRAFE

"... O melhor de mim é aquilo que não sei, aquilo que desconheço, aquilo que me renova."

Mário Sérgio Cortella, 2013.

LISTA DE FIGURAS

LISTA DE TABELAS

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EXPLICATIVAS PARA A ESTIMATIVA DA ALTURA DE ÁRVORES DE EUCALIPTO

RESUMO

Conhecer a altura das árvores dos povoamentos é fundamental para quantificação dos recursos florestais. Contudo, a obtenção desta variável é de difícil operacionalização, realizada por meio de métodos indiretos, através da utilização de hipsômetros, que são passíveis a diversos erros. O objetivo do trabalho foi avaliar o desempenho de diferentes estratégias para seleção de variáveis na precisão das estimativas da altura de árvores de eucalipto. Em cada talhão foram delimitadas cinco parcelas retangulares, com aproximadamente 540 m² de área, sendo mensurado o diâmetro (cm) de todas as árvores, e a altura de total (m) cerca de 33% das árvores das parcelas, também com a identificação e medição da altura das árvores dominantes. O conjunto de dados foram divididos de forma aleatória em dois subconjuntos: ajuste (75%) e validação (25%). Os dados de ajuste foram utilizados para aplicação de diferentes estratégias para a seleção de variáveis a serem incluídas nos modelos hipsométricos, e posteriormente, nos ajustes dos modelos selecionados. Foram utilizados diferentes procedimentos estatísticos para a seleção das variáveis explicativas: modelo empírico, correlação de Pearson, regressão Stepwise (método Forward), análise de trilha e algoritmo Random Forest, sendo executados, com e sem a variável altura dominante (Hd), no conjunto de variáveis. Foram avaliadas as seguintes hipóteses: $H_0(1)$: Os métodos de seleção das variáveis explicativas proporcionam a seleção de variáveis com maior capacidade preditiva, que resultam na melhoria da precisão da estimativa da altura de árvores de eucalipto; e $H_0(2)$: A variável Hd pode ser substituída por outras de mais fácil obtenção, sem a perda da capacidade preditiva dos modelos. A análise de trilha é o procedimento estatístico mais eficiente para a seleção das variáveis explicativas na construção dos modelos hipsométricos, com desempenho superior aos modelos de referência comumente utilizados. Portanto, a hipótese $H_0(1)$ foi aceita. Todos os modelos obtidos sem a inclusão da variável Hd tem perda na qualidade de precisão das estimativas, quando comparados aos modelos cuja altura dominante estava inclusa. Portanto, a hipótese $H_0(2)$ foi rejeitada.

Palavras-chave: Análise de trilha. Inventário Florestal. Modelos hipsométricos. Random Forest, Modelos de Regressão.

METHODS OF SELECTION OF EXPLANATORY VARIABLES FOR ESTIMATING THE HEIGHT OF EUCALYPTUS TREES

ABSTRACT

Knowing the height of trees in stands is fundamental for quantifying forest resources. However, obtaining this variable is difficult to operationalize, performed by means of indirect methods, through the use of hypsometers, which are subject to several errors. The objective of this work was to evaluate the performance of different strategies for selection of variables on the accuracy of estimates of the height of eucalyptus trees. In each plot, five rectangular plots were delimited, with an area of approximately 540m², and the diameter (cm) of all trees was measured, and the total height (m) of about 33% of the trees in the plots, also with the identification and measurement the height of the dominant trees. The dataset was randomly divided into two subsets: fit (75%) and validation (25%). The adjustment data were used to apply different strategies for selecting variables to be included in the hypsometric models, and subsequently, in the adjustments of the selected models. Different statistical procedures were used for the selection of explanatory variables: empirical model, Pearson correlation, Stepwise regression (Forward method), path analysis and Random Forest algorithm, being executed, with and without the dominant height variable (Hd), in the set of variables. The following hypotheses were evaluated: H0(1): The methods of selection of explanatory variables provide the selection of variables with greater predictive capacity, which result in improved precision in estimating the height of eucalyptus trees; and H0(2): The variable Hd can be replaced by others that are easier to obtain, without losing the predictive capacity of the models. Path analysis is the most efficient statistical procedure for selecting explanatory variables in the construction of hypsometric models, with superior performance than commonly used reference models. Therefore, the hypothesis H0(1) was accepted. All models obtained without including the variable Hd have a loss in the quality of precision of the estimates, when compared to the models whose dominant height was included. Therefore, hypothesis H0(2) was rejected.

Keywords: Trail analysis. Forest Inventory. hypsometric models. Random Forest, Regression Models.

SUMÁRIO

1. INTRODUÇÃO

No Brasil, há aproximadamente 9,55 milhões de hectares ocupados com plantios florestais, sendo a maior parte, aproximadamente 78%, destinada à cultura do eucalipto (IBÁ, 2021). Com base na expansão das áreas de cultivo e na busca por altas produtividades, é cada vez mais notória a necessidade de aprimorar a condução dos inventários florestais na mensuração dos povoamentos, principalmente no que diz respeito a quantificação dos estoques de madeira (BINOTI et al., 2013). Nesse sentido, o conhecimento da altura das árvores é fundamental nos processos de quantificação dos recursos florestais (VIBRANS et al., 2015; VENDRUSCOLO et al., 2015). No entanto, a obtenção desta variável é de difícil operacionalização, realizada principalmente por meio de métodos indiretos, como através da utilização de equipamentos específicos (hipsômetros), que, apesar de viabilizarem a medição, se mostram passíveis de erros e tornam o processo lento, cansativo e oneroso (BINOTI et al., 2013; THIERSCH et al., 2013; SANQUETTA et al., 2014; FERRAZ FILHO et al., 2018).

Desta forma, a utilização de modelos hipsométricos surge como uma alternativa eficiente para contornar esses problemas (THIERSCH et al., 2013). Com base nos pares altura-diâmetro obtidos com a medição da altura de apenas algumas árvores das parcelas lançadas, são ajustados modelos de regressão que geram equações para estimar a altura das demais árvores, o que leva a uma redução no tempo de trabalho em campo e, consequente, maior economicidade do processo de inventário como um todo (THIERSCH et al., 2013; FERRAZ FILHO et al., 2018). Embora a maioria dos modelos tradicionais correlacionem a altura apenas com o diâmetro das árvores, essa alternativa pode apresentar problemas em situações em que haja heterogeneidade das parcelas em relação a fatores como: posição sociológica, região, idade, densidade de plantio, silvicultura, entre outros, haja vista a influência desses fatores sobre o crescimento em altura das árvores (THIERSCH et al., 2013; VENDRUSCOLO et al., 2017; ACOSTA et al., 2020).

Assim, uma alternativa frequente consiste em correlacionar a altura individual das árvores com o seu diâmetro e com a altura dominante (Hd) da parcela, uma vez que esta variável é relacionada à capacidade produtiva dos sítios florestais (LEITE et al., 2003; LEITE et al., 2011; CAMPOS et al., 2016). No entanto, considerando que a obtenção desta variável acarreta num maior tempo de trabalho em campo, devido a necessidade adicional de obtenção da altura das árvores dominantes, é importante verificar se sua utilização interfere na precisão das estimativas da altura em comparação aos modelos que não a utilizam, ou se outras variáveis de mais fácil obtenção possam substituir Hd. Leite et al. (2011) verificaram que a variável diâmetro

dominante (Dd) pode ser empregada em modelos hipsométricos em substituição a Hd, sem perda de precisão nas estimativas.

Contudo, dada a vasta gama de informações que podem ser obtidas a partir da mensuração dos povoamentos florestais, modelos com alto número de variáveis selecionadas de forma empírica, acabam sendo menos parcimoniosos, mais complexos e de difícil ajuste (MORAES NETO et al., 2012; VENDRUSCOLO et al., 2015, CERQUEIRA et al., 2019). Portanto, é fundamental buscar por métodos capazes de determinar as variáveis que possuem maior associação com a variável dependente, evitando aquelas que possam ser irrelevantes, ruidosas, correlacionadas entre si e/ou não confiáveis (ANDERSEN; BRO, 2010; TABAKHI et al., 2014).

Nesse sentido, diferentes métodos para seleção de variáveis explicativas vêm sendo amplamente propostos pela literatura em diversas áreas (MASIERO E ANZANELLO, 2011; ANZANELLO, 2013; STEIN et al., 2014; CERVO E ANZANELLO, 2015), sendo uma importante alternativa para redução da complexidade e maior poder de predição dos modelos a serem utilizados (MEHMOOD et al., 2012; YUN, et al., 2019). É possível encontrar diversos métodos para seleção dessas variáveis, a exemplo da correlação de Pearson (STANTON, 2001; FREITAS et al., 2017), regressão Stepwise (HOCKING, 1976; ALVES et al., 2013), análise de trilha (WRIGHT, 1921; ZUFFO et al., 2018; AZEVEDO et al., 2022) e algoritmo Random Forest (BREIMAN, 2001; PEREIRA et al., 2022).

Em suma, a correlação de Pearson (r) mede a direção e o grau da relação linear entre duas variáveis quantitativas. Valores negativos, indicam uma correlação inversa, ou seja, conforme uma variável aumenta a outra diminui, já valores positivos indicam uma correlação de mesmo sentido, onde o aumento de uma variável implica no aumento da outra. Por fim, valores próximos a zero, indicam a inexistência de uma relação linear entre as variáveis (FIGUEIREDO et al., 2014; ZHU ET AL., 2019).

A regressão Stepwise pelo método Forward adiciona sequencialmente ao modelo uma variável preditora a cada etapa, até que a inclusão de uma nova variável não contribua significativamente com o modelo (LOCKHART et al., 2014). No entanto, uma vez que a variável entra no modelo não sai mais, mesmo que sua contribuição não seja mais significante após a entrada de outra. Sendo assim, a seleção ocorre de forma automática, não permitindo a limitação dos números de variáveis a serem selecionadas (WILKINSON e DALLAL, 1981; LOFTUS E TAYLOR, 2014).

A análise de trilha (WRIGHT, 1921) permite, através do desdobramento dos coeficientes de correlação de Pearson existente entre as variáveis, o estudo dos efeitos diretos e indiretos de

variáveis independentes sobre uma variável de interesse (CORREIA et al., 1996; SILVA PINHEIRO et al., 2021). Isso possibilita uma análise mais detalhada da associação entre duas variáveis por meio da predição de coeficientes que caracterizam a relação de causa e efeito entre estas (AZEVEDO et al., 2016; TRAUTENMÜLLER et al., 2019).

Já o algoritmo Random Forest cria várias "árvores de decisão", que estabelecem regras para tomada de decisão, criando assim uma estrutura com "pontos" onde uma condição é verificada, se atendida essa condição, o fluxo segue, caso contrário, muda de direção "ramo", sempre levando ao próximo ponto, até a finalização da "árvore". O modelo RF fornece estimativas confiáveis dos erros, utilizando dados conhecidos como "out-of-bag" (OOB), que é um subconjunto aleatório dos dados não utilizado pelo algoritmo para construção das árvores (LIAW e WIENER, 2015; AKPA et al., 2016; ZERAATPISHEH et al., 2019). Uma das vantagens deste método é a capacidade de apontar quais as variáveis mais importantes para o modelo, dando peso para cada uma delas (STROB et al., 2007).

De modo geral, esses métodos buscam identificar as variáveis explicativas mais associadas com a variável de interesse, permitindo selecionar aquelas com maior capacidade preditiva e que mantenha a parcimônia dos modelos. Embora os métodos mencionados sejam utilizados em diversas áreas do conhecimento, a maioria dos estudos que envolvem a predição de variáveis na área florestal, a exemplo dos modelos hipsométricos, não mencionam sobre os métodos utilizados para a seleção das variáveis explicativas que compõem os modelos, e tão pouco é realizada uma comparação da precisão das estimativas em modelos que tiveram as variáveis selecionadas por diferentes métodos. Diante do exposto, este estudo teve como objetivo avaliar a precisão das estimativas da altura de árvores de eucalipto a partir de modelos hipsométricos que tiveram as variáveis explicativas selecionadas por diferentes métodos, com e sem a inclusão da variável Hd. Foram avaliadas as seguintes hipóteses: $H_0(1)$: Os métodos de seleção das variáveis explicativas proporcionam a seleção de variáveis com maior capacidade preditiva, que resultam na melhoria da precisão da estimativa da altura de árvores de eucalipto; e $H_0(2)$: A variável Hd pode ser substituída por outras de mais fácil obtenção, sem a perda da capacidade preditiva dos modelos.

2. MATERIAL E MÉTODOS

2.1 Área de Estudo

Os dados foram obtidos em plantios comerciais de quatro clones de eucalipto (AEC 0144 – *Eucalyptus urophylla* (C1); AEC 0224 – *E. urophylla* (C2); VM01 – *E. urophylla* x *E.*

camaldulensis (C3); H77 – *E. urophylla x E. grandis* (C4) (Tabela 1), cultivados em primeira e segunda rotação (total de 8 talhões), com espaçamento médio de plantio de 3 x 3 m, localizados no município de Ribas do Rio Pardo, no Estado de Mato Grosso do Sul, Brasil, nas coordenadas 20◦19′14″S e 53◦17′28″W. A altitude média é de 380 m acima do nível do mar, a média anual precipitação é de 1.252 mm ano^{-1} e a temperatura média anual é de 24,9 ◦C. O solo dominante foi um Latossolo (13,75% argila, 2,5% silte e 83,75% areia) com pH 5,1 e 4,9 g kg^{-1} de matéria orgânica. A idade dos plantios atualmente, varia de 5,9 a 12,9 anos.

Tabela 1. Características silviculturais e dendrométricas das plantações avaliadas.

CLONE	C1	C1	C2	C2	C3	C3	C4	C4
Rotação	1ª	2ª	1ª	2ª	1ª	2ª	1ª	2ª
Idade (anos)	6,8	13,6	6,6	9,5	6,3	12,7	14,5	13,5
Condução da rebrota (anos)	-	6,0	-	3,1	-	5,7	-	5,9
N° de fustes por hectare	987	1703	933	1088	1049	1635	976	1067
Área basal (m² ha^{-1})	25,25	30,16	26,58	16,55	23,67	26,76	34,80	24,32
Volume (m³ ha^{-1})	279,85	369,02	285,63	144,59	210,27	235,31	387,98	208,64

Em que: C1 = clone AEC 0144, C2 = clone AEC 0224, C3 = clone VM01, C4 = clone H77

2.2 Obtenção dos dados

Em cada um dos talhões foram delimitadas cinco parcelas retangulares, com aproximadamente 540 m² de área cada (6 linhas de plantio x 10 plantas na linha), distribuídas aleatoriamente na área. No campo, foram mensuradas nas árvores vivas a circunferência à altura do peito (CAP – em centímetros), com casca, e a altura total das árvores (H – em metros). A CAP foi obtida na altura de 1,3 metros do nível do solo, com auxílio de uma fita métrica, em todas as árvores de cada parcela. A altura foi obtida em cerca de 33% das árvores das parcelas (duas primeiras linhas de cada parcela – cerca de 20 árvores), com o auxílio de um clinômetro Haglof. Além disso, para identificação das árvores dominantes, conforme o conceito de Assmann (1970), ainda no campo, foram identificadas as cinco árvores de maior CAP em cada parcela (árvores dominantes) e obtida a sua altura caso estas ainda não tivessem sido mensuradas. Nos plantios de segunda rotação, para a identificação das árvores dominantes, foram considerados apenas os fustes de maior diâmetro em cada cepa.

Posteriormente, os dados de CAP foram transformados em diâmetro a altura do peito (D) e foram obtidas algumas variáveis derivadas em unidade de área para cada uma das parcelas.

Foram calculados o diâmetro médio quadrático (Dg) e a área basal (G) (RETSLAFF, 2014). O diâmetro dominante (Dd) e a altura dominante (Hd) em cada uma das parcelas foram obtidos pela média aritmética de D e H das árvores dominantes, respectivamente. O número de fustes por cepa (F) foi obtido pela contagem do número de brotos conduzidos em cada cepa. O número de árvores (N) e número de fustes (Nf) por hectare foram obtidos, respectivamente, pela contagem do número de cepas e do número de fustes em cada parcela, e extrapolados para a unidade de hectare.

2.3 Análise dos dados

Inicialmente o conjunto de dados obtidos foi dividido de forma aleatória em dois subconjuntos: ajuste (75%) e validação (25%), os quais apresentam distribuição semelhante (Figura 1). Os dados de ajuste foram utilizados na aplicação de diferentes estratégias para a seleção de variáveis explicativas candidatas a serem incluídas nos modelos hipsométricos, e posteriormente, nos ajustes dos modelos selecionados. Os dados de validação foram utilizados como um banco de dados independentes para a avaliação da qualidade das estimativas a partir das equações hipsométricas obtidas em cada uma das estratégias.

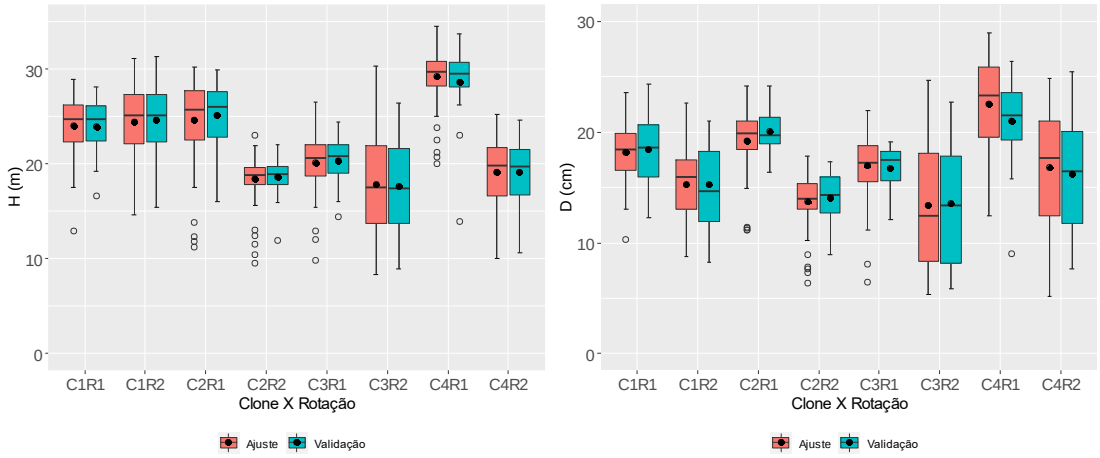

Figura 1. Boxplot para as variáveis altura (H) e diâmetro a altura do peito (D) de árvores de quatro clones de eucalipto (prefixos C1 a C4), cultivadas em primeira e segunda rotação (sufixos R1 e R2), considerando os dados de ajuste e validação.

Foram utilizados diferentes procedimentos estatísticos para a seleção das variáveis explicativas: correlação de Pearson (STANTON, 2001), regressão Stepwise (HOCKING, 1976), análise de trilha (WRIGHT, 1921) e algoritmo Random Forest (BREIMAN, 2001). Esses procedimentos foram executados com e sem a variável Hd no conjunto de variáveis explicativas candidatas ao modelo. Para se ter uma referência nos ganhos de precisão das estimativas com a utilização das variáveis selecionadas a partir de cada um dos procedimentos,

foram utilizados como referência modelos de regressão linear simples e múltipla. Para os modelos sem inclusão de Hd foi utilizada apenas a variável independente D (Modelo 1) e para os modelos com inclusão de Hd foram utilizadas as variáveis independentes D e Hd (Modelo 2). Assim, no total foram adotadas 10 estratégias para a seleção das variáveis a serem inclusas nos modelos (Tabela 2).

$$H = \beta_0 + \beta_1.D + \varepsilon \quad (1)$$

$$H = \beta_0 + \beta_1.D + \beta_2.Hd + \varepsilon \quad (2)$$

Tabela 2. Estratégias adotadas para a seleção de variáveis explicativas a serem utilizadas no ajuste de modelos hipsométricos para árvores de eucalipto.

Estratégia	Procedimento para a seleção das variáveis explicativas	Inclusão de HD
RE	Regressão Linear Simples	Não
REHD	Regressão Linear Múltipla	Sim
CR	Correlação de Pearson	Não
CRHD	Correlação de Pearson	Sim
FO	Regressão stepwise pelo método forward	Não
FOHD	Regressão stepwise pelo método forward	Sim
PA	Análise de trilha	Não
PAHD	Análise de trilha	Sim
RF	Randon Forest	Não
RFHD	Randon Forest	Sim

Nas estratégias CR e CRHD foram incluídas nos modelos as três variáveis explicativas com maior correlação com H. Essa análise foi realizada no software R, com auxílio do pacote "metan" (OLIVOTO e LÚCIO, 2020). Nas estratégias FO e FOHD foram utilizadas nos modelos todas as variáveis selecionadas no procedimento Stepwise pelo método Forward. Essa análise foi realizada no software R com auxílio do pacote "olsrr" (HEBBALI e HEBBALI, 2018).

Nas estratégias PA e PAHD foram selecionadas para os modelos as três variávieis independentes com maior efeito direto sobre a variável dependente H. Para controle do grau de multicolinearidade entre as variáveis, a análise de trilha foi realizada em crista, com a

introdução de uma constante K igual a 0,05 na diagonal da matriz de correlação (BARBOSA et al., 2017; 435 MOREIRA et al., 2013; OLIVOTO et al., 2017). A análise de trilha foi realizada no sofware R, com auxílio do pacote "metan" (OLIVOTO e LÚCIO, 2020). Nas estratégias RF e RFHD foram seleciondas as três variáveis de maior importância para a variável dependente H. Essa análise foi realizada no sofware R com auxílio do pacote "randomForest" (LIAW e WIENER, 2002).

Após a seleção das variáveis independentes a partir de cada uma das estratégias de seleção adotadas (Tabela 1) foram ajustados os modelos hipsométricos. Foram utilizados modelos de regressão linear, sem a transformação das variáveis, e as equações obtidas foram empregadas para estimar a variável dependente H no banco de dados de validação. O ajuste dos modelos foi realizado com auxílio da função "lm" do pacote "stats", disponível no software R (R CORE TEAM, 2021).

Para realizar a validação, as equações obtidas foram utilizadas para estimar a altura das árvores no banco de dados destinado à validação. A precisão das estimativas nas etapas de ajuste e validação dos modelos foi avaliada com base nos seguintes critérios: coeficiente de determinação ajustado (R_{aj}^2); média absoluta dos erros (MAE) e; análise gráfica dos erros em porcentagem (Erro (%). De forma complementar, foram obtidos também o fator de inflação da variância (VIF), cujo valores inferiores a 10 indicam a ausência de multicololinearidade (BERK, 1977; ALIN, 2010; CRUZ et al., 2019), com auxílio do pacote "car" (FOX e WEISBERG, 2019). Essa análise permite avaliar se há problemas de multicolinearidade entre as variáveis que compõem o modelo.

Como o banco de dados utilizado no presente estudo é bastante heterogêneo, proveniente de talhões com quatro clones em primeira e segunda rotação, adicionalmente, verificou-se o modelo de melhor desempenho geral também é adequado para estimar a altura das árvores em condições específicas. Para tanto, o modelo de melhor desempenho e os modelos de referência (modelos 1 e 2) foram reajustados e revalidados separadamente para cada condição de clone e rotação. A precisão das estimativas foi realizada utilizando os mesmos critérios utilizados na etapa inicial de ajuste dos modelos, exceto para o MAE, que foi calculado em porcentagem, facilitando a comparação da magnitude do erro nas diferentes combinações de clones e rotações.

3. RESULTADOS

A variável dependente H apresentou correlação significativa com todas as variáveis explicativas a um nível de 0,01% de probabilidade (Figura 2). Em ordem decrescente, as maiores correlações foram com D, Hd, Dg, G, Dd, N, F e Nf. O diagnóstico de colinearidade da matriz de correlação do conjunto total de variáveis explicativas indica que há problemas de multicolinearidade com os dados (03 números de VIFs ≥ 10) (MONTGOMERY; PECK, 1981).

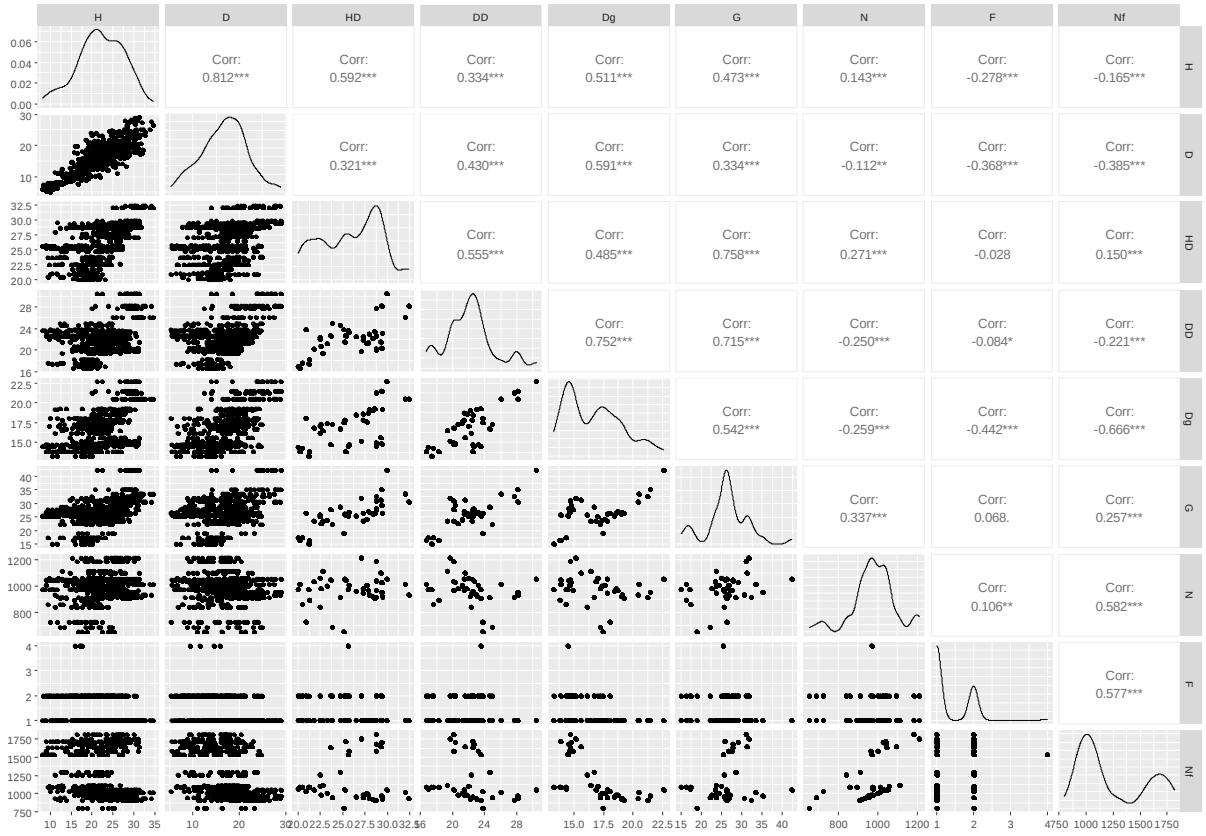

Figura 2. Correlação entre variáveis dendrométricas de árvores clonais de eucalipto cultivadas em primeira e segunda rotação. Em que: H = altura total; D = diâmetro; HD = altura dominante; DD = diâmetro dominante; Dg = diâmetro médio quadrático; G = área basal; N = número de árvores por hectare; F = número de fustes por cepa; Nf = número de fustes por hectare.

Dessa forma, com base nas três maiores correlações entre H e as variáveis explicativas, para as estratégias CR e CRHD, foram selecionados os modelos 3 e 4, respectivamente:

$$H = \beta_0 + \beta_1.D + \beta_2.Dg + \beta_3.G + \varepsilon \quad (3)$$

$$H = \beta_0 + \beta_1.D + \beta_2.Hd + \beta_3.Dg + \varepsilon \quad (4)$$

O procedimento Stepwise gerou modelos com seis variáveis explicativas para as estratégias FO e FOHD. Em ambas as estratégias, a variável D foi a primeira a ser inserida nos

modelos (R^2_{aj} = 0,6588 e MAE = 2,44). A inclusão das demais variáveis promoveu melhorias nas estatísticas de ajuste dos modelos (FO: R^2_{aj} = 0,7643 e MAE = 2,02; FOHD: R^2_{aj} = 0,8349 e MAE = 1,68) (Tabela 3).

Tabela 3. Sumário de seleção das variáveis do modelo pela análise de regressão stepwise pelo método de Forward.

Estratégia	Passo	Variável de Entrada	R^2_{aj}	MAE
FO	1	D	0,6588	2,44
	2	N	0,7139	2,28
	3	G	0,7304	2,23
	4	Dd	0,7499	2,07
	5	Dg	0,7553	2,05
	6	Nf	0,7643	2,02
FOHD	1	D	0,6588	2,44
	2	Hd	0,7813	1,96
	3	Dd	0,8273	1,73
	4	G	0,8308	1,72
	5	Nf	0,8328	1,70
	6	Dg	0,8349	1,68

Em que: R^2_{aj} = coeficiente de determinação ajustado; MAE = média absoluta dos erros; FO = regressão stepwise pelo método forward sem a altura dominante; FOHD = regressão stepwise pelo método forward com a altura dominante; D = diâmetro; Hd = altura dominante; Dd = diâmetro dominante; Dg = diâmetro médio quadrático; G = área basal; N = número de árvores por hectare; F = número de fustes por cepa; Nf = número de fustes por hectare.

Assim, com base nas estratégias FO e FOHD foram selecionados os modelos 5 e 6, respectivamente:

$$H = \beta_0 + \beta_1.D + \beta_2.N + \beta_3.G + \beta_4.Dd + \beta5.Dg + \beta_6.Nf + \varepsilon \quad (5)$$

$$H = \beta_0 + \beta_1.D + \beta_2.Hd + \beta_3.Dd + \beta_4.G + \beta5.Nf + \beta_6.Dg + \varepsilon \quad (6)$$

Na análise de trilha, as variáveis com maior efeito direto sobre H, em ordem decrescente, foram D, Hd, Dd, G, N, Nf, Dg e F (Tabela 4). Observa-se que o método foi eficiente para determinar a relação causa-efeito entre as variáveis, com R^2 de 0,7931. Além disso, o Número de Condição (CN) menor que 100, o Fator de Inflação (VIF) menor que 10 e a Determinante da Matriz (D) maior que zero, indicam que a análise não é enviesada por problemas de multicolinearidade.

Assim, com base nas estratégias PA e PAHD, foram selecionados os modelos 7 e 8, respectivamente:

$$H = \beta_0 + \beta_1.D + \beta_2.Dd + \beta_3.G + \varepsilon \quad (7)$$

$$H = \beta_0 + \beta_1.D + \beta_2.Hd + \beta_3.Dd + \varepsilon \quad (8)$$

Tabela 4. Resultado da análise de trilha para avaliação dos efeitos diretos (diagonal principal) e indiretos (fora da diagonal) das variáveis independentes sobre a variável dependente altura.

Variáveis	D	Hd	Dd	Dg	G	N	F	Nf
D	0,7337	0,1277	-0,1184	0,0204	0,0359	-0,0043	0,0024	0,0147
Hd	0,2241	0,4179	-0,1529	0,0167	0,0815	0,0104	0,0002	-0,0057
Dd	0,3002	0,2208	-0,2894	0,0259	0,0768	-0,0096	0,0005	0,0084
Dg	0,4127	0,1931	-0,2074	0,0362	0,0583	-0,0100	0,0028	0,0254
G	0,2336	0,3017	-0,1970	0,0187	0,1129	0,0129	-0,0004	-0,0098
N	-0,0785	0,1077	0,0689	-0,0089	0,0362	0,0403	-0,0007	-0,0222
F	-0,2570	-0,0112	0,0230	-0,0152	0,0073	0,0041	-0,0068	-0,0220
Nf	-0,2689	0,0598	0,0610	-0,0229	0,0276	0,0224	-0,0037	-0,0401
Máximo VIF:						9,5800		
Determinante:						0,0049		
R^2:						0,7931		
Resíduo:						0,4549		

Em que: D = diâmetro; HD = altura dominante; DD = diâmetro dominante; Dg = diâmetro médio quadrático; G = área basal; N = número de árvores por hectare; F = número de fustes por cepa; Nf = número de fustes por hectare; VIF = fator de inflação da variância; R^2 = coeficiente de determinação.

A análise pelo procedimento Random Forest demonstrou boa capacidade assertiva dos modelos propostos na predição dos dados out-of-bag (Tabela 5), com erro quadrático médio inferior a 3,5 e percentagem da variância explicada superior a 87,45%. As variáveis mais importantes para estimar H foram D, G e Nf para a estratégia RF e D, Hd e G para a estratégia RFHD.

Tabela 5. Importância das variáveis explicativas, com e sem inclusão da altura dominante (HD), no modelo de Randon Forest.

Variáveis	Sem Hd		Com Hd	
	% IncMSE	Inc Pureza Nós	% IncMSE	Inc Pureza Nós
D	110	9017,8	103,6	7891,6
HD	-	-	39,9	3388,4
DD	22,3	1323,9	19,2	924,3
Dg	25,3	2323,2	20	1544,9
G	35,8	2031,2	21	1264,2
N	24,1	480,2	17,2	268,3
F	13,2	163,5	13,4	174,1
Nf	28,8	784,4	17	426,4
Quadrado médio dos resíduos:	3,495		3,500	
% variação explicada:	87,430		87,430	

Em que: % IncMSE = incremento no erro médio quadrático; Inc Pureza Nós = Incremento na pureza dos nós. D = diâmetro; HD = altura dominante; DD = diâmetro dominante; Dg = diâmetro médio quadrático; G = área basal; N = número de árvores por hectare; F = número de fustes por cepa; Nf = número de fustes por hectare.

Assim, com base nas estratégias RF e RFHD, foram selecionados os modelos 9 e 10, respectivamente:

$$H = \beta_0 + \beta_1.D + \beta_2.G + \beta_3.Nf + \varepsilon \quad (9)$$

$$H = \beta_0 + \beta_1.D + \beta_2.Hd + \beta_3.G + \varepsilon \quad (10)$$

O resultado do ajuste dos 10 modelos é apresentado na Tabela 6. Na maioria das estratégias utilizadas os coeficientes associados às variáveis selecionadas para o modelo foram significativos, exceto para as estratégias FO (coeficientes associados a G e N foram não significativos) e para RF (intercepto não significativo), sem a inclusão da variável Hd. Os valores de VIF indicam que não há problemas de multicolinearidade entre as variáveis selecionadas nos modelos, exceto para o procedimento Stepwise (FO e FOHD), em que VIF foi maior do que 10 para as variáveis Dg, G e Nf.

Tabela 6. Resultados do ajuste dos coeficientes hipsométricos, com base na seleção de variáveis explicativas por diferentes estratégias (Estr.)

Estr.	Par.	Intercepto	D	Hd	Dd	Dg	G	N	Nf
RE	Coef.	6,9335*	0,9052*						
RE	VIF								
REHD	Coef.	-6,1366*	0,7729*	0,5906*					
REHD	VIF		1,11	1,11					
CR	Coef.	3,8784*	0,8668*			-0,1993*	0,2669*		
CR	VIF		1,54			1,93	1,42		
CRHD	Coef.	-4,1077*	0,8488*	0,6646*		-0,3119*			
CRHD	VIF		1,54	1,31		1,81			
FO	Coef.	-25,3011*	0,8560*		-0,7194*	2,3132*	-0,2847^{ns}	0,0017^{ns}	0,0131*
FO	VIF		1,54		5,44	91,59	53,34	2,63	66,29
FOHD	Coef.	18,4547*	0,8312*	0,7700*	-0,6146*	-1,2027*	0,5396*		-0,0087*
FOHD	VIF		1,55	3,30	3,54	127,12	59,48		90,77
PA	Coef.	8,2969*	0,8978*		-0,5829*		0,4429*		
PA	VIF		1,23		2,23		2,04		
PAHD	Coef.	-1,8867*	0,8581*	0,7929*	-0,49*				
PAHD	VIF		1,24	1,46	1,61				
RF	Coef.	0,6054^{ns}	0,8728*				0,1923*		0,0015*
RF	VIF		1,54				1,40		1,46
RFHD	Coef.	-6,6101*	0,7837*	0,7006*			-0,0970*		
RFHD	VIF		1,14	2,38			2,40		

Em que: Coef. = coeficientes estimados; Sig. = significância dos coeficientes; VIF = fator de inflação da variância; * e ns = significativo a 5% e não significativo a 5% de probabilidade respectivamente. Significado das siglas das estratégias disponíveis na Tabela 1.

A inclusão da variável Hd nos modelos contribuiu para a melhoria da precisão das estimativas de H em todos os procedimentos de seleção das variáveis (Tabela 7). O modelo 2 (estratégia REHD), que teve as variáveis selecionadas de forma empírica, incluindo a variável Hd, proporcionou estimativas com maior precisão do que todos os modelos selecionados pelas estratégias que não incluíram Hd entre as variáveis explicativas (RE, CR, FO, PA e RF). Quando analisada a precisão separadamente para as estratégias que não incluíram e que incluíram a variável Hd, todos os procedimentos adotados para a seleção das variáveis contribuíram para a melhoria da precisão das estimativas de H em relação à obtida nos modelos de referência (estratégias RE e REHD), o que torna a hipótese $H_0(1)$ verdadeira. Nos modelos que não incluíram Hd, da maior para a menor precisão, as estratégias se comportaram na seguinte ordem: FO > PA > RF > CR > RE; enquanto nas estratégias que incluíram HD foi observada a seguinte ordem: FOHD > PAHD > CRHD > RFHD > REHD.

Tabela 7. Estatísticas de ajuste e validação dos modelos obtidas para as diferentes estratégias de seleção das variáveis explicativas, para o ajuste de modelos hipsométricos, para plantios clonais de eucalipto em primeira e segunda rotação.

Estratégia	Ajuste		Validação	
	R^2aj	MAE	R^2aj	MAE
RE	0,658	2,439	0631	2,550
REHD	0,781	1,958	0,766	1,967
CR	0,707	2,318	0,693	2,312
CRHD	0,791	1,931	0,777	1,911
FO	0,764	2,015	0,722	2,118
FOHD	0,834	1,685	0,817	1,684
PA	0,749	2,069	0,721	2,145
PAHD	0,827	1,728	0,810	1,711
RF	0,708	2,315	0,693	2,311
RFHD	0,784	1,914	0,766	1,907

Em que: R^2aj. = coeficiente de determinação ajustado; mae = média absoluta dos erros. Significado das siglas das estratégias disponíveis na Tabela 1.

Observando o maior valor de R^2_{aj} resultante da validação dos modelos sem a inclusão de Hd, obtido pela regressão Stepwise, é possível verificar que para o modelo com a inclusão da variável houve um acréscimo de 13,16% no valor desta estatística. Considerando o menor valor de MAE para modelos sem Hd, também obtido pela regressão stepwise, a inclusão da altura dominante diminuiu o erro em 20,49%, sendo essa, a maior diferença para todas as estatísticas de validação comparando os modelos com e sem Hd.

Os modelos ajustados apresentaram leve tendência em superestimar (valores negativos) os menores valores de H (Figura 3).

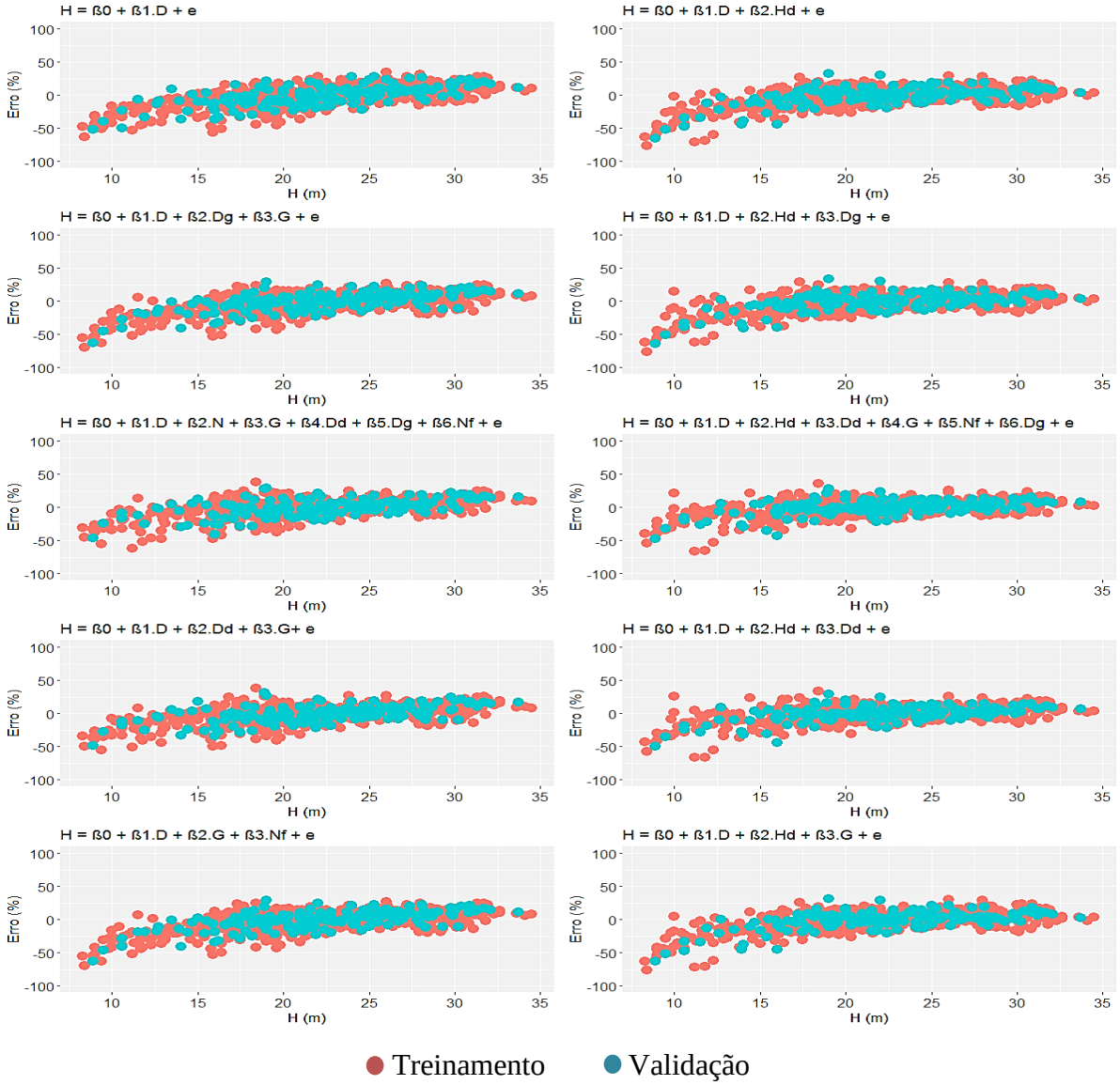

Figura 3. Distribuição dos erros em porcentagem na estimativa da altura de árvores de clones de eucalipto, cultivados em primeira e segunda rotação, com base nos modelos ajustados com as variáveis explicativas selecionadas por diferentes métodos.

Nota-se também, que os modelos contendo a variável Hd apresentaram distribuição gráfica levemente mais favoráveis, com maior uniformidade e menor tendenciosidade dos valores de erros. Já em relação aos métodos de seleção das variáveis, os modelos obtidos pelos procedimentos stepwise e análise de trilha mostraram-se ligeiramente mais acertivos. Assim, apesar da distribuição de erros seguir um padrão semelhante entre os modelos, os modelos 6 e 8 proporcionaram distribuição ligeiramente mais favoráveis, o que corrobora com as estatísticas obtidas nas etapas de ajuste e validação dos modelos.

Portanto, como nos modelos ajustados pelo procedimento stepwise houve a inclusão de variáveis com efeito não significativo e valores de VIF maiores do que 10, considera-se o modelo 8 como o de melhor desempenho para a estimativa da altura para as condições em que foram realizados os ajustes. O ajuste desse modelo, juntamente com os modelos 1 e 2, para cada condição específica de clone e rotação demonstrou que o modelo 8 também apresentou bom desempenho em cinco das oito condições (Tabela 8).

Tabela 8. Estatísticas de ajuste e validação obtidas para as diferentes combinações de clone x rotação, considerando os modelos de controle e o modelo válido com melhor desempenho para o povoamento (Modelo 8).

Variáveis Explicativas	C x R	b0	D	Hd	Dd	Treinamento		Validação	
						R^2aj	MAE %	R^2aj	MAE %
D, Hd, Dd	C1R1	4,615690	0,994968*	0,342492	-0,363009	0,756	4,75	0,733	4,71
D, Hd	C1R1	-2,579963	0,987509*	0,320356		0,756	4,78	0,734	4,78
D	C1R1	5,963956*	0,990373*			0,750	4,81	0,769	4,49
D, Hd, Dd	C1R2	18,585996*	0,982707*	-0,192367	-0,181899	0,675	7,46	0,619	8,12
D, Hd	C1R2	17,246986*	0,982909*	-0,274920		0,678	749	0,630	8,17
D	C1R2	9,332508*	0,986737*			0,677	7,49	0,659	7,95
D, Hd, Dd	C2R1	-8,680886	1,204915*	1,650041*	-1,62344*	0,797	5,82	0,603	7,13
D, Hd	C2R1	-18,598796	1,258714*	0,668538		0,741	7,05	0,397	8,37
D	C2R1	0,517067	1,254093*			0,737	7,21	0,392	8,58
D, Hd, Dd	C2R2	8,311698	0,81686*	0,510099	-0,672901	0,612	6,65	0,141	7,98
D, Hd	C2R2	8,638661	0,797273*	-0,058844		0,610	6,71	0,201	7,82
D	C2R2	7,389074*	0,799833*			0,615	6,72	0,254	7,79
D, Hd, Dd	C3R1	13,726311	0,835844*	0,288384	-0,69095*	0,659	6,72	0,335	8,13
D, Hd	C3R1	2,881191	0,818526*	0,147534		0,613	7,21	0,245	9,21
D	C3R1	6,119099*	0,823843*			0,619	7,23	0,301	9,21
D, Hd, Dd	C3R2	22,448098*	0,957116*	0,708853*	-1,55487*	0,913	7,20	0,885	7,65
D, Hd	C3R2	-12,601519*	0,953756*	0,686441*		0,900	7,79	0,893	7,83
D	C3R2	5,008074*	0,957402*			0,871	8,78	0,854	9,45
D, Hd, Dd	C4R1	20,887124*	0,452803*	0,545383*	-0,66069*	0,583	5,19	0,720	5,64
D, Hd	C4R1	-4,309032	0,398503*	0,805342*		0,503	5,54	0,603	6,46
D	C4R1	20,284652*	0,394387*			0,345	6,22	0,545	8,49
D, Hd, Dd	C4R2	7,452651	0,602358*	0,866089*	-0,77269*	0,747	7,74	0,745	7,83
D, Hd	C4R2	-6,034939	0,588754*	0,663974*		0,721	8,37	0,722	7,91
D	C4R2	9,083079*	0,59889*			0,647	9,27	0,711	8,39

Em que: Variáveis Explicativas = variáveis integrantes dos modelos; CxR = Clone x Rotação; b0 = intercepto; D, Hd e Dd = coeficientes da regressão associados as variáveis diâmetro à altura do peito, altura dominante e diâmetro dominante, respectivamente; R^2aj = coeficiente de correlação ajustado; MAE = média dos erros absolutos.

Assim, é possível verificar que a seleção de variáveis através da análise de trilha, pôde gerar um modelo com boa capacidade preditiva também em situações mais homogêneas, além disso, a inclusão do diâmetro dominante nos modelos tradicionais, presente no modelo 8, pode contribuir com a melhoria da capacidade preditiva dos modelos em diferentes situações.

4. DISCUSSÃO

A utilização de modelos hipsométricos baseados apenas no diâmetro para estimativa indireta da altura total de árvores de eucalipto é grandemente difundida na literatura (MORAES NETO et al., 2010; CAMPOS e LEITE, 2013; VENDRUSCOLO et al., 2015a). De modo geral, é possível verificar que em modelos dessa natureza os valores estimados costumam ser confiáveis (SANQUETTA et al., 2014 e VENDRUSCOLO et al., 2015b), o que corrobora com o desempenho satisfatório dos modelos de referência adotados (RE e REHD). No entanto, comparado aos demais modelos, nota-se que os modelos de referência apresentaram desempenho inferior aos demais em todos os procedimentos de seleção das variáveis, tanto no ajuste quanto na validação (Tabela 7). A utilização desses modelos pode não ser a mais indicada em situações onde haja heterogeneidade das parcelas em relação a: posição sociológica, região, idade, densidade de plantio, silvicultura, altura dominante, capacidade produtiva do sítio entre outros, tendo em vista a influência desses fatores sob sua qualidade preditiva (THIERSCH et al., 2013; VENDRUSCOLO et al., 2017a; ACOSTA et al., 2020).

De modo geral, a correlação procura entender o comportamento conjunto de duas variáveis em um cenário dinâmico, buscando identificar e quantificar através dos coeficientes de correlação alguma relação entre ambas (FIGUEIREDO FILHO et al., 2014). Desta forma, a utilização da correlação na seleção de variáveis para formulação de um modelo hipsométrico se mostra coerente, com a tendência de apresentar estimativas consideravelmente precisas como as obtidas neste estudo (Tabela 7). No entanto, a correlação pode não representar integralmente a relação de causa-efeito entre duas variáveis, pois, uma terceira variável pode se relacionar de forma isolada com as duas que estão sendo analisadas, influenciando diretamente a correlação observada, podendo sub ou superestimar o resultado, levando a conclusões precipitadas sobre a variável de interesse (MARTINS, 2014).

Nos modelos com maior número de variáveis (modelos 5 e 6), resultantes da seleção pelo método Stepwise, as estimativas da altura foram mais precisas em comparação com aquelas obtidas nos demais, compostos por apenas três variáveis cada (Tabela 7). Diferente dos demais métodos, na regressão Stepwise as variáveis são selecionadas de forma automática, não sendo

possível determinar de forma discriminada o número de variáveis a serem selecionadas. O método, que começa com o modelo vazio, adiciona variáveis de forma automática até que a última a ser adicionada não contribua significativamente com a seleção (LOFTUS e TAYLOR, 2014).

Embora um banco de dados que contenha muitas variáveis possa influir na praticidade e confiabilidade do modelo, pela possível associação deste a um alto nível de ruído, colinearidade e variáveis redundantes (TANG et al., 2014), cada variável independente adicionada, contribui de acordo com sua relação de causa-efeito com a variável dependente, para um melhor ajuste do mesmo (ANDERSEN e BRO 2010; BURGEL e ANZANELLO 2018). Portanto, é razoável afirmar que os modelos 5 e 6 foram, provavelmente, beneficiados pelo maior número de variáveis em sua composição. Contudo, a qualidade de um modelo hipsométrico passa também indiscriminadamente pelo nível de relação entre as variáveis independentes que o compõe, característica conhecida como multicolinearidade (GREENE, 2012; TOEBE E CARGNELUTTI FILHO, 2013a). A multicolinearidade indica que existe uma relação de linearidade entre as variáveis independentes, prejudicando assim as estimativas dos coeficientes de regressão (DORMANN et al., 2013; THOMPSON, 2017; SARI et al., 2018).

A literatura descreve diversos métodos para detecção da multicolinearidade. Altos valores de coeficiente de correlação (r) entre as variáveis, normalmente acima de 0.8, são um forte indicativo quanto a ocorrência de colinearidade (GUJARATI e PORTER, 2011). No entanto, a multicolinearidade não corresponde apenas ao nível de correlação entre as variáveis, mas sim ao quanto uma variável pode explicar outra, e o quanto estas compartilham de informações entre si. Logo, é possível a ocorrência de multicolinearidade mesmo em cenários com baixos valores de coeficiente de correlação entre as variáveis independentes (MONTGOMERY et al., 2012).

Nesse contexto, a utilização de outras técnicas para o diagnóstico de multicolinearidade se mostra prudente, podendo ser realizada por exemplo através de métodos que se baseiam em indicadores como: Número de Condição (NC), determinante da matriz de correlação (DET) e Fator de Inflação da Variância (VIF) (GUJARATI e PORTER, 2011; MONTGOMERY et al., 2012). Dentre estes, sendo um dos mais utilizados, o cálculo do fator de inflação da variância (VIF) (BERK, 1977) mede o quanto cada variável explicativa do modelo é explicada pelas demais, através da variação de sua variância (ALIN, 2010; CRUZ et al., 2019). Comparado aos demais indicadores, o VIF se destaca por demonstrar o efeito da inflação da variância para cada variável, diferente de NC e DET que observam as variáveis explicativas em conjunto. Dessa forma, através da interpretação dos valores de VIF, é possível identificar qual ou quais as

variáveis apresentam problemas de multicolinearidade e eliminá-las do modelo (AZEVEDO et al., 2016; ALVES et al., 2017a).

Na interpretação das grandezas de VIF, assume-se que valores iguais a 1 (VIF = 1) representam a ausência de multicolinearidade, enquanto valores iguais ou maiores que 10 (VIF =>10) indicam que há um forte efeito de relação entre as variáveis (GUJARATI e PORTER, 2011; ALVES et al., 2017b), como observado nos modelos 5 e 6, onde três das seis variáveis presentes nos modelos, apresentaram altos valores de VIF (Tabela 6), o que torna inviável a sua utilização.

Apesar da melhor precisão nas estimativas, pode se considerar que no presente estudo, o procedimento Stepwise não foi eficiente para a seleção das variáveis explicativas a serem incluídas nos modelos. Esse comportamento também foi observado por Eisfeld et al. (2018), ao utilizar o método Forward na seleção de variáveis para estimativa da biomassa foliar e potencial extrativista da pimenta *Pseudocaryophyllus* (CATAIA), verificando que o mesmo não foi capaz de detectar e retirar do modelo variáveis multicolineares. Desta forma, para a regressão Forward, seria prudente o diagnóstico prévio da ocorrência e o uso de alguma técnica para redução do grau de multicolinearidade, como a retirada das variáveis enviesadas por uma alta relação (MOREIRA et al., 2013; TOEBE e CARGNELUTTI FILHO, 2013a; SALLA et al., 2015).

Com a precisão das estimativas ligeiramente inferior as do procedimento Stepwise para as duas estratégias abordadas (PA e PAHD), os modelos determinados pela análise de trilha através da regressão em crista não apresentaram enviesamento por multicolinearidade, com valores de VIF inferiores a dez (Tabela 6). Proposta por Wright (1921), a análise de trilha desdobra a correlação em uma relação de efeitos diretos e indiretos de um grupo de variáveis explicativas sobre uma variável de interesse (SALLA et al., 2015), sendo amplamente utilizada em estudos referentes ao melhoramento genético de plantas para a seleção de caracteres de forma indireta (TEIXEIRA et al. 2012; LÚCIO et al. 2013; TOEBE e CARGNELUTTI FILHO, 2013b).

No entanto, para que os resultados da análise de trilha sejam confiáveis, é extremamente importante controlar o grau de multicolinearidade entre as variáveis. Para tanto, uma das possibilidades se baseia na utilização da análise com regressão em crista. A regressão em crista consiste basicamente na introdução de uma constante K à diagonal da matriz de correlação, com a utilização do menor valor a partir do qual os coeficientes se estabilizem, mantendo assim os indicadores de multicolinearidade dentro do aceitável (Tabela 3) (RIOS et al., 2012; TOEBE

e CARGNELUTTI FILHO, 2013a; AZEVEDO et al., 2016; BARBOSA et al., 2017; MOREIRA et al., 2013; OLIVOTO et al., 2017).

Ao analisar o desempenho dos modelos compostos por apenas três variáveis (modelos 3, 4, 7, 8, 9 e 10), percebe-se que a inclusão da variável Hd contribuiu para a melhoria da precisão das estimativas de H, como pode ser observado para as estatísticas de validação (Tabela 7). A inclusão de características do povoamento nos modelos hipsométricos pode resultar em vantagens na obtenção de estimativas mais precisas (BINOTI et al., 2013). Embora os modelos clássicos estimem em sua maioria a altura apenas com base na relação hipsométrica da altura com o diâmetro, resultados mais precisos podem ser obtidos com a inclusão de demais variáveis, como a capacidade de sítio, idade e altura dominante (MACHADO e FIGUEIREDO FILHO, 2006; SANQUETTA et al., 2009; MORAES NETO et al., 2010; RIBEIRO et al., 2010; SCOLFORO et al., 2015).

A altura dominante já é comumente utilizada em métodos de classificação da capacidade produtiva de sítios florestais para o ajuste de modelos de crescimento e produção (SELLE et al., 2008, LEITE et al., 2011), principalmente por ser uma variável pouco influenciada por fatores extrínsecos como densidade e tratamentos silviculturais (CAMPOS e LEITE, 2013). Portanto, quando utilizada, é possível que esta variável condicione o modelo a diferenciar a estimativa da altura total de árvores que possuam o mesmo diâmetro, mas que estejam em parcelas com diferentes capacidades produtivas (LEITE et al., 2003), o que pode explicar o melhor desempenho dos modelos obtidos com a inclusão de Hd.

A semelhança no padrão de distribuição dos resíduos para todos os modelos obtidos (Figura 3), superestimando a altura dos indivíduos mais baixos, sugere a influência de um mesmo fator ou grupo de fatores, sobre a capacidade preditiva de cada. Considerando as circunstâncias dos plantios estudados quanto a amplitude das idades, diferentes clones/espécies, classe de sítio e manejo, é provável que a heterogeneidade destes fatores possa ter influenciado diretamente na capacidade preditiva das árvores mais baixas uma vez que, diferentes clones/espécies apresentam ritmos de crescimento distintos (AZEVEDO et al., 2011; CURTO et al., 2014).

Já em relação as idades, em plantios mais novos, a variabilidade da altura total nas classes inferiores desta variável é maior, como observado por Miranda et al. (2014), ao estudar a estratificação hipsométrica em plantios clonais de eucaliptos. Utilizando Stepwise e Random forest na predição da altura total de *Eucalyptus spp.*, Lopes et al. (2021), também verificaram perda na qualidade preditiva dos modelos ajustados para povoamentos com diferentes idades (2 e 7 anos), obtendo valores superestimados para as menores alturas.

Considerando seu efeito sobre a qualidade dos modelos, a idade se mostra uma variável fundamental para as modelagens quando se tem interesse em ajustar uma equação geral para todo povoamento, a fim de reduzir a complexidade, o tempo e os custos atrelados ao processo quando realizado de forma estratificada (SANTAMARÍA et al., 2013; RIVAS et al., 2014). Quando inclusa, a idade pode ser capaz de identificar variações na altura das árvores em diferentes estratos, aumentando a qualidade preditiva dos modelos principalmente em povoamentos inequiâneos (RETSLAFF et al., ZANG, 2016).

O bom desempenho do Modelo 8, quando ajustado para todo o povoamento ou para cada combinação de clone x rotação separadamente, em comparação aos demais, pode estar relacionado ao peso das variáveis que o compõe, sugerindo que estas, apesar de serem suscetíveis, são menos influenciadas por fatores genéticos, ambientais e de manejo. (LEITE et al., 2011; CAMPOS e LEITE, 2013).

A variável Hd, já comumente utilizada em equações hipsométricas, tem a capacidade de identificar diferentes capacidades produtivas existentes nas áreas amostradas, permitindo com que haja a diferenciação no valor da altura de árvores que possuam mesmo diâmetro, localizadas em diferentes locais (LEITE et al., 2003; LEITE et al., 2011; MELO et al., 2017).

Já a variável Dd, embora ainda pouco utilizada em modelagens de crescimento, pode também contribuir para a diferenciação da altura de árvores individuais em diferentes locais, conforme descrito por Leite et al. (2011), cuja utilização desta variável no lugar de Hd, não apresentou perda na qualidade das estimativas. A possibilidade de substituir o uso de Hd por Dd em modelos hipsométricos implica em consideráveis vantagens para o processo de inventário florestal, principalmente no que diz respeito à dificuldade de obtenção destas. Para se obter a variável Hd, é necessário na maioria das vezes medir a altura de um maior número de árvores por parcela, sendo ainda passível à erros, uma vez que é medida de forma indireta, através do uso de hipsômetros por exemplo (SANQUETTA et al., 2014; FERRAZ FILHO et al., 2018). Já o diâmetro dominante é obtido de forma facilitada, por ser derivado da variável D, que por sua vez já é obtida para todas as árvores de cada parcela (LEITE et al., 2011; ARAÚJO JÚNIOR et al., 2016. Assim, o uso da variável Dd implicaria na redução de possíveis erros, do tempo e dos custos atrelados ao processo de coleta dos dados em campo.

No entanto, com base no melhor modelo válido selecionado para todo o povoamento, no qual a variável Hd fez parte (Modelo 8), e nos resultados de ajuste e validação obtidos para cada combinação de clone x rotação (Tabela 8), os modelos cuja altura dominante foi retirada obtiveram desempenho inferior a aqueles cuja Hd fez parte, em pelo menos metade das

observações, reforçando a importância desta variável na predição da altura de árvores de eucalipto, sendo, portanto, a hipótese $H_0(2)$, falsa.

Conforme amplamente descrito pela literatura, a formulação de modelos hipsométricos deve considerar os diversos fatores que influenciam essa relação, sejam genéticos, ambientais ou silviculturais (THIERSCH et al., 2013; VENDRUSCOLO et al., 2017; ACOSTA et al., 2020), levando assim a necessidade de ajustar equações específicas para cada unidade amostral, determinadas com base em alguma característica homogênea presente nestas (RIBEIRO et al., 2010; VENDRUSCOLO et al., 2015). Contudo, o ajuste de modelos hipsométricos para cada situação leva a um número elevado de equações, o que aumenta o tempo e, consequentemente, os custos do inventário florestal, evidenciando assim a necessidade de construir modelos cuja capacidade preditiva seja capaz de considerar os fatores heterogêneos intrínsecos ao povoamento (SANTAMARÍA et al., 2013; MENDONÇA et al., 2015; CERQUEIRA et al., 2019).

Nesse sentido, o bom desempenho do Modelo 8, quando ajustado para todo o povoamento em comparação aos demais modelos obtidos (Modelos 1, 2, 3, 4, 5, 6, 7, 9 e 10), e para cada combinação de clone x rotação, em comparação aos modelos de referência (Modelos 1 e 2), sugere que há potencial no uso de modelos generalizados para estimativa da altura total de árvores, corroborando com diversos autores que já abordaram esse tema (SANTAMARÍA et al., 2013; RIVAS et al., 2014; ZANG, 2016; XIE et al.,; ZHANG et al., 2020), e reafirmando a necessidade de dar continuidade a estes estudos.

Embora amplamente difundidas e consolidadas as equações hipsométricas, seja na literatura ou nas empresas florestais, a busca por novas técnicas para determinação da altura mostra-se promissora. Apesar de ainda pouco utilizados e carentes de mais estudos, com base nos resultados obtidos no presente trabalho, comprovou-se a aplicabilidade dos métodos estatísticos para seleção de variáveis na construção de modelos hipsométricos, incentivando seu uso pela maior praticidade, eficiência, capacidade de precisão e ineditismo, trazendo benefícios em relação principalmente a redução de custos e aumento na precisão das estimativas.

5. CONCLUSÃO

A análise de trilha é o procedimento estatístico mais eficiente para a seleção das variáveis explicativas na construção dos modelos hipsométricos, com desempenho superior aos modelos de referência comumente utilizados. Portanto, a hipótese $H_0(1)$ foi aceita.

Todos os modelos obtidos sem a inclusão da variável Hd tem perda na qualidade de precisão das estimativas, quando comparados aos modelos cuja altura dominante estava inclusa. Portanto, a hipótese $H_0(2)$ foi rejeitada.

6. REFERÊNCIAS

ACOSTA, H. A. B.; GARRETT, A. T. A.; LANSSANOVA, L. R., DIAS, A. N.; TAMBARUSSI, E. V., FILHO, A. F., GUIMARÃES, F. A. R.; CABRAL, O. M. V. Identidade de modelos hipsométricos para clones de eucalipto na região oriental do Paraguai. **Rev. Scientia Forestalis**, v. 48, n. 125, p. 1 - 12, 2020.

AKPA, S. I.; ODEH, I. O.; BISHOP, T. F.; HARTEMINK, A. E.; AMAPU, I. Y. Total oil organic carbon and carbon sequestration potential in Nigeria. **Geoderma**, v. 271, p. 202-215, 2016.

ALIN, A. Multicollinearity. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, n. 3, p. 370-374, 2010.

ALVES, B. M.; CARGNELUTTI FILHO, A. Genotypic correlation and path analysis in early and super-early maize genotypes. **Genetics and Molecular Research**, Ribeirão Preto. v. 16, n. 2, p. 1–12, 2017b.

ALVES, B. M.; CARGNELUTTI FILHO, A.; BURIN, C. Multicollinearity in canonical correlation analysis in maize. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 1, p. 1–14, 2017a.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 1, n. 1, 2013.

ANDERSEN, C. M.; BRO, R. Variable selection in regression - a tutorial. **Journal of Chemometrics,** v. 24, n. 11-12, p. 728-737, 2010.

ANZANELLO, M. J. Seleção de variáveis para classificação de bateladas produtivas com base em múltiplos critérios. **Production Journal**, v. 23 n. 4, 858-865, 2013.

ARAÚJO JÚNIOR, C. A.; SOARES, C. P. B.; LEITE, H. G. Site index curves in eucalyptus plantation obtained by quantile regression. **Pesquisa Agropecuária Brasileira**, v. 51, p. 720-727, 2016.

ASSMANN, E. **The principles of forest yield study: studies in the organic production, structure, increment and yield of forest stands**. Elsevier, 2013.

AZEVEDO, A. M.; SEUS, R.; GOMES, C. L.; FREITAS, E. M.; CANDIDO, D. M.; SILVA, D. J. H.; CARNEIRO, P. C. S. Correlações genotípicas e análise de trilha em famílias de meios-irmãos de couve de folhas. **Pesquisa Agropecuária Brasileira**, v.51, n.1, p.35-44, 2016.

AZEVEDO, G. B., REZENDE, A. V., AZEVEDO, G. T. D. O. S., MIGUEL, E. P., AQUINO, F. D. G., TEODORO, L. P. R., TEODORO, P. E. Prognosis of aboveground woody biomass in a central Brazilian Cerrado monitored for 27 years after the implementation of management systems. **European Journal of Forest Research**, v. 141, n. 1, p. 1-15, 2022.

AZEVEDO, G. B.; SOUSA, G. T. O.; SILVA, H. F.; BARRETO, P. A. B.; NOVAES, A. B. Seleção de modelos hipsométricos para quatro espécies florestais nativas em plantio misto no planalto da conquista na Bahia. **Enciclopédia Biosfera**, v. 7, n. 12, p. 1-12, 2011.

BARBOSA, R. P. et al. Early selection of sugarcane using path analysis. **Genetics and Molecular Research,** Ribeirão Preto, v. 16, n. 1, p. 1–8, 2017.

BHERING, S. B.; CHAGAS, C. D. S.; CARVALHO JUNIOR, W. D.; PEREIRA, N. R.; CALDERANO FILHO, B.; PINHEIRO, H. S. K. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. **Pesquisa Agropecuária Brasileira**, v. 51, p. 1359-1370, 2016.

BINOTI, M. L. M. S.; BINOTI, D. H. B.; LEITE, H. G. Aplicação de redes neurais artificiais para estimação da altura de povoamentos equiâneos de eucalipto. **Revista Árvore**, Viçosa, v.37, n.4, p.639-645, 2013.

BURGEL, E.; ANZANELLO, M. J. Abordagem para Seleção de Variáveis Preditivas no Contexto de Controle de Inventários. **R. Gest. Industr.**, Ponta Grossa, v. 14, n. 4, p. 154-195, 2018.

CAMPOS, B. P. F.; SILVA, G. F. D.; BINOTI, D. H. B.; MENDONÇA, A. R. D.; LEITE, H. G. Predição da altura total de árvores em plantios de diferentes espécies por meio de redes neurais artificiais. **Pesq. flor. bras.**, Colombo, v. 36, n. 88, p. 375-385, 2016.

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

CARVALHO JUNIOR, W. D.; CALDERANO FILHO, B.; CHAGAS, C. D. S.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa Agropecuária Brasileira**, v. 51, p. 1428-1437, 2016.

CERQUEIRA, C. L.; MÔRA, R.; TONINI, H.; VENDRUSCOLO, D. G. S.; LANSSANOVA, L. R.; ARCE, J. E.; DINIZ, C. C. C. Efeito do espaçamento e arranjo de

plantio na relação hipsométrica de eucalipto em sistema consorciado de produção. **Nativa**, v. 7, n. 6, p. 763-770, 2019.

CERVO, V. L.; ANZANELLO, M. J. Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento kernel. **Production**, v. 25, n. 4, 2015.

CORREIA, J. R.; COSTA, L. M.; NEVES, J. C. L.; CRUZ, C. D. Estudo do relacionamento entre características físicas e químicas do solo e a produtividade do gênero Pinus. **Revista Árvore**, Viçosa, v. 20, n. 2, p. 161-169. 1996.

CRUZ, C. D.; CARNEIRO, P. C. S.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético vegetal.** 3. ed. Viçosa: Editora UFV, 2014. 668p.

CRUZ, G. S.; SILVA RIBEIRO, R. B. S.; GAMA, J. R. V.; DE ALMEIDA, B. R. S.; XIMENES, L. C.; GOMES, K. M. A.; BEZERRA, T. G. Ajuste e avaliação na estimativa volumétrica para Lecythis lurida (Miers) SA Mori em uma área de manejo florestal. **Advances in Forestry Science**, v. 6, n. 1, p. 549-554, 2019.

CURTO, R. D. A.; LOUREIRO, G. H.; MÔRA, R.; MIRANDA, R. O. V.; NETTO, S. P.; SILVA, G. F. Relações hipsométricas em floresta estacional semidecidual. **Revista de Ciências Agrárias Amazonian Journal of Agricultural and Environmental Sciences**, v. 57, n. 1, p. 57-66, 2014.

DORMANN, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, v. 36, n. 1, p. 27–46, 2013.

EISFELD, R. L.; ANSOLIN, R. D.; D'ANGELIS, A. S. R.; JÚNIOR, E. R. S.; COSTA FILHO, S. V. S. Estimativa da Biomassa Foliar e Potencial Extrativista da Pimenta *Pseudocaryophyllus* (CATAIA). **BIOFIX Scientific Journal**, v. 3, n. 1, p. 145-151, 2018.

FERRAZ FILHO, A. C.; Mola-Yudego, B.; Ribeiro, A.; Scolforo, J. R. S.; Loos, R. A.; Scolforo, H. F. Height-diameter models for Eucalyptus sp. plantations in Brazil. **Cerne**, v. 24, p. 9-17, 2018.

FIGUEIREDO FILHO, D. B.; ROCHA, E. C.; SILVA JR, J.; PARANHOS, R. A.; NEVES, J. A. B.; SILVA, M. B. "Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o Retorno". **Leviathan**, São Paulo, V. 8, p. 66-95, 2014.

FOX, J.; WEISBERG, S.; AN, R. Companion to Applied Regression, Third. https://socialsciences.mcmaster.ca/jfox/Books/Companion/, 2019.

FREITAS, A. P.; GUTERRES, M. X.; LAMPERT, V. N.; SILVA, A. H. S.; BARCELLOS, J. O. J.; MARQUES, P. R. Aplicação de método de seleção de variáveis em um modelo DEA na produção de bovinos de corte. **Engevista**, v. 19, n.4, p. 881-889, 2017.

GENUER, R.; POGGI, J.; TULEAU-MALOT, C. VSURF: an R package for variable selection using random forests. **The R Journal**, v. 7, n. 2, p. 19-33, 2015.

GREENE, W. H. Econometric analysis. **Seventh Edition**, Pearson, 2012. 1238 p.

GUJARATI, D. N.; Porter, D. **Econometria Básica.** 5. ED. Rio de Janeiro, RJ: Editora Mc Graw-Hill, 2011. 402 p.

HAMNER, B.; FRASCO, M. Metrics: Evaluation Metrics for Machine Learning. **R package version 0.1.4**. https://CRAN.R-project.org/package=Metrics, v. 4, p. 2018, 2018.

HEBBALI, A.; HEBBALI, M. A. Maintainer Aravind. Package 'olsrr'. **Version 0.5**, v. 3, 2017.

HOCKING, R. R. The Analysis and Selection of Variables in Linear Regression. **Biometrics**, Washington, v. 32, n. 1, p. 1-49, 1976.

IBÁ - INDUSTRIA BRASILEIRA DE ÁRVORES. **Relatório Anual**. São Paulo, Ano base 2020. 2021.

LEITE, H. G.; ANDRADE, V. C. L. Importância das variáveis altura dominante e altura total em equações hipsométricas e volumétricas. **Revista Árvore**, Viçosa, v. 27, n. 3, p.301-310, 2003.

LEITE, H. G.; CASTRO, R., SILVA, A., JÚNIOR, C.; BINOTI, D.; CASTRO, A. F.; BINOTI, M. Classificação da Capacidade Produtiva de Povoamentos de Eucalipto Utilizando Diâmetro Dominante. **Silva Lusitana**, Lisboa, v. 19, n. 2, p. 181-195, 2011.

LI, J.; MALLEY, J. D.; ANDREW, A. S.; KARAGAS, M. R.; JASON, H. M. Detecting gene-gene interactions using a permutation-based random forest method. **BioData mining**, v. 9, n. 1, p. 1-17, 2016.

LI, J.; SUN, L.; YAN, Q.; LI, Z.; SRISA-AN, W.; YE, H. Significant permission identification for machine-learning-based android malware detection. **IEEE Transactions on Industrial Informatics**, v. 14, n. 7, p. 3216-3225, 2018.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R News**, v. 2, n. 3, p. 18-22, 2002.

LIAW, A.; WIENER, M. Random Forest: Breiman and Cutler's random forests for classification and regression. **R package version**, v. 4, p. 14, 2015.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. AND TIBSHIRANI, R. A significance test for the lasso. **Annals of statistics**, v. 42, n. 2, p. 413, 2014.

LOFTUS, J. R.; TAYLOR, J. E. A significance test for forward stepwise model selection. **arXiv preprint arXiv**:1405.3920, 2014.

LOPES, I. L.; ARAÚJO, L. A.; MIRANDA, E. N.; ABREU, V. S.; GOMES, V. S.; ALMEIDA, B. C.; GONÇALVES, A. F. A.; BARBOSA, L. O.; GOMIDE, L. R. **Aplicação de técnicas de regressão linear e aprendizagem de máquinas na predição da altura total de árvores de eucalyptus spp. Silvicultura e manejo florestal: técnicas de utilização e conservação da natureza.** 1. ed. Lavras, Minas Gerais MG: Editora Científica Digital, 2021, 29 – 43, 440 p.

LÚCIO, A.D.C.; STORCK, L.; KRAUSE, W.; GONÇALVES, R.Q.; NIED, A.H. Relações entre os caracteres de maracujazeiro-azedo. **Ciência Rural**, v.43, p.225-232, 2013.

MA, LI; FAN, S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. **BMC bioinformatics**, v. 18, n. 1, p. 1-18, 2017.

MACHADO, S. A.; FIGUEIREDO FILHO, A. **Dendrometria.** 2. ed. Guarapuava, Paraná: Editora Unicentro, 2006, 316 p.

MARTINS, M. E. G. Coeficiente de correlação amostral. **Revista de Ciência Elementar**, v. 2, n. 2, p. 69, 2014.

MASIERO, M. S.; ANZANELLO, M. J. Seleção de variáveis para predição utilizando regressão linear em processos logísticos de distribuição. **Universidade Federal do Rio Grande do Sul**, 2011.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SAEBO, S. A review of variable selection methods in partial least squares regression. **Chemometrics and intelligent laboratory systems**, v. 118, p. 62-69, 2012.

MELO, E. D. A.; CALEGARIO, N.; MENDONÇA, A. R. D.; POSSATO, E. L.; ALVES, J. D. A.; ISAAC, M. A. Modelagem Não Linear Da Relação Hipsométrica e Do Crescimento Das Árvores Dominantes e Codominantes de Eucalyptus sp. **Ciência Florestal**, v. 27, p. 1325-1338, 2017.

MENDONÇA, A. R. D.; CARVALHO, S. D. P. C.; CALEGARIO, N. Modelos hipsométricos generalizados mistos na predição da altura de *eucalyptus sp*. **Cerne**, v. 21, p. 107-115, 2015.

MIRANDA, R. O. V.; DAVID, H. C.; EBLING, Â. A.; MÔRA, R.; FIORENTIN, L. D.; SOARES, I. D. Extratificação hipsométrica em classes de sítio e de altura total em plantios clonais de eucaliptos. **Advances in Forestry Science**, v. 1, n. 4, p. 113-119, 2014.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analyses**. (5thedition). Wiley, 2012, 661p.

MONTGOMERY, D.C.; PECK, E.A. Introduction to linear regression analysis. New York: John Wiley, 1981. 504p.

MORAES NETO, S. P.; Pulrolnik, K.; Vilela, L.; Munhoz, D. J. M.; Gumarães Junior, R.; Marchão, R. L. Modelos Hipsométricos para Eucalyptus cloeziana e Eucalyptus urophylla x Eucalyptus grandis em Sistema Agrossilvipastoril. **Embrapa Cerrados-Boletim de Pesquisa e Desenvolvimento (INFOTECA-E)**, v. 286, 2010.

MORAES NETO, S. P.; PULROLNIK, K.; VILELA, L.; OLIVEIRA, P. D.; GUIMARAES JUNIOR, R.; MACIEL, G. Verificação da identidade de modelos hipsométricos em diversos arranjos de sistema agrossilvipastoril. **Embrapa Cerrados-Boletim de Pesquisa e Desenvolvimento (INFOTECA-E)**, 2012.

MOREIRA, S. O. et al. Correlações e análise de trilha sob multicolinearidade em linhas recombinadas de pimenta (*Capsicum annuum L*.). **Revista Brasileira de Ciências Agrárias**, v. 8, n. 1, p. 15-20, 2013.

OLIVOTO, T. et al. Multicollinearity in path analysis: A simple method to reduce its effects. **Agronomy Journal**, v. 109, n. 1, p. 131–142, 2017.

OLIVOTO, T.; LÚCIO, A. D. Metan: An R package for multi-environment trial analysis. **Methods in Ecology and Evolution**, v. 11, n. 6, p. 783-789, 2020.

PEREIRA, L. E. C. Estratégias para seleção de variáveis em diferentes modelos preditivos para tuberculose animal. **UNESP – Tese de Doutorado,** 2022.

PINHEIRO, H. S. K.; OWENS, P. R.; ANJOS, L. H. C.; CARVALHO JÚNIOR, W.; CHAGAS, C. S. Treebased techniques to predict soil units. **Soil Research**, v. 55, n. 8, p. 788-798, 2017.

R CORE TEAM R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL https://www.R-project.org/, 2019.

RETSLAFF, F. A. D. S.; FIGUEIREDO FILHO, A.; DIAS, A. N.; BERNETT, L. G.; FIGURA, M. A. Curvas de sítio e relações hipsométricas para Eucalyptus grandis na região dos Campos Gerais, Paraná. **Cerne**, v. 21, p. 219-225, 2015.

RETSLAFF, F. A. Simulador para prognose da produção de *Pinus taeda* com diagrama de manejo da densidade associado à modelagem em classes de diâmetro. **Tese (Doutorado em Engenharia Florestal)**, Universidade Federal do Paraná - PR, 167 f, 2014.

RIBEIRO, A.; FERRAZ FILHO, A. C.; MELLO, J. M.; FERREIRA, M. Z.; LISBOA, P. M. M.; SCOLFORO, J. R. S. Estratégias e metodologias de ajuste de modelos hipsométricos em plantios de Eucalyptus sp. **Cerne**, v. 16, p. 22-31, 2010.

RIOS, S. de A.; BORÉM, A.; GUIMARÃES, P.E. de O.; PAES, M.C.D. Análise de trilha para carotenoides em milho. **Revista Ceres**, v. 59, n. 3, p. 368-373, 2012.

RIVAS, S. C.; GONZÁLEZ, J. G. A.; CAMPO, F. C.; RIVAS, J. J. C. Local and generalized height-diameter models with random parameters for mixed, uneven-aged forests in northwestern Durango, Mexico. **Forest Ecosystems**, v. 1, n. 1, p. 1–9, 2014.

RODRÍGUEZ-LADO, L.; RIAL, M.; TABOADA, T.; CORTIZAS, A.M. A pedotransfer function to map soil bulk density from limited data. **Procedia Environmental Sciences**, v. 27, p. 45-48, 2015.

SALLA, V. P. et al. Análise de trilha em caracteres de frutos de jabuticabeira. **Pesquisa Agropecuária Brasileira**, v. 50, n. 3, p. 218–223, 2015.

SANQUETTA, C. R.; BEHLING, A.; CORTE, A.P.; RUZA, M. S.; SIMON. A.; JOSÉ, J. F. B. S. Relação hipsométrica em inventários pré-corte em povoamentos de Acacia mearnsii De Wild. **Científica**, Jaboticabal, v. 42, n. 1, p. 80-90, 2014.

SANQUETTA, C. R.; WATZLAWICK, L. F.; CÔRTE, A. P. D.; FERNANDES, L. A. V.; SIQUEIRA, J. D. P. Inventários florestais: planejamento e execução. **Multi-Graphic**, Curitiba, v. 2, 2009.

SANTAMARÍA, J. C.; Campo, F. C.; Martínez, J. L. F.; Anta, M. B.; Obeso, J. R. Tree height prediction approaches for uneven-aged beech forests in northwestern Spain. **Forest Ecology and Management**, v. 307, p. 63–73, 2013.

SARI B. G.; LÚCIO, A. D.; OLIVOTO, T.; KRYSCZUN, D. K.; TISCHLER, A. L.; DREBES, L. Interferência do tamanho de amostra no diagnóstico de multicolinearidade em análise de trilha. **Pesq. agropec. bras.**, Brasília, v. 53, n. 6, p. 769-773, 2018.

SCOLFORO, H. F.; RAIMUNDO, M. R.; SCOLFORO, J. R. S.; MELLO, J. M.; BATISTA, A. P. B.; BULLOCK, B. Hypsometric approaches to Eucalyptus experiments in Brazil. **African Journal of Agricultural Research**, v. 10, n. 45, p. 4176-4184, 2015.

SELLE, G. L.; PAULESKI, D. T.; BRAZ, E. M. Como classificar sítios florestais através da altura dominante do povoamento. Colombo: **Embrapa Florestas**, 46 p, 2008.

SILVA PINHEIRO, L. et al. Analise de trilha da massa da espiga de milho e seus atributos físicos. **Research, Society and Development**, v. 10, n. 1, p. e41510111912, 2021.

SOUZA, H. S.; TSUKAMOTO FILHO, A. A.; VENDRUSCOLO, D. G. S.; CHAVES, A. G. S.; MOTTA, A. S. Modelos hipsométricos para eucalipto em sistema de integração lavoura-pecuária-floresta. **Revista Nativa**, Sinop, v. 4, n. 1, p. 11-14, 2016.

STANTON, J. M. Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. **Journal of Statistics Education**, v. 9, n. 3, 2001.

STEIN, M.; ANZANELLO, M. J.; KAHMANN, A. Sistemática para identificação das variáveis preditivas mais relevantes em um processo do setor metal-mecânico. **Revista Gestão Industrial**, v. 10, n. 1, 2014.

STROB, C.; BOULEXTEIX, A. L.; ZEILEIS, A.; HOTHORN, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. **BMC bioinformatics**, v. 8, n. 1, p. 25, 2007.

TABAKHI, S.; MORADI, P.; AKHLAGHIAN, F. An unsupervised feature selection algorithm based on ant colony optimization. **Engineering Applications of Artificial Intelligence**, v. 32, p. 112-123, 2014.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. **Data classification: Algorithms and applications**, p. 37, 2014.

TEAM, R. Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2021.

TEIXEIRA, D. H. L.; OLIVEIRA, M. do S. P.; GONÇALVES, F. M. A.; NUNES, J. A. R. Correlações genéticas e análise de trilha para componentes da produção de frutos de açaizeiro. **Revista Brasileira de Fruticultura**, v. 34, p. 1135-1142, 2012.

THIERSCH, C. R.; ANDRADE, M. G. D.; MOREIRA, M. F. B.; LOIBEL, S. Estimativa da relação hipsométrica em clones de Eucalyptus sp. com o modelo de curtis ajustado por métodos bayesianos empíricos. **Revista Árvore**, Viçosa, v. 37, n. 1, p. 01-08, 2013.

THOMPSON, C. G.; KIM, R. S.; ALOE, A. M.; BECKER, B. J. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. **Basic and Applied Social Psychology**, v. 39, n. 2, p. 81-90, 2017.

TOEBE, M.; CARGNELUTTI FILHO, A. Multicollinearity in path analysis of maize (*Zea mays* L.). **Journal of Cereal Science**, v. 57, n. 3, p. 453–462, 2013a.

TOEBE, M.; CARGNELUTTI FILHO, A. Não normalidade multivariada e multicolinearidade na análise de trilha em milho. **Pesquisa Agropecuária Brasileira**, v. 48, p. 466-477, 2013b.

TRAUTENMÜLLER, J. W.; NETTO, S. P.; BALBINOT, R.; DALLA CORTE, A. P.; BORELLA, J. Path analysis applied to evaluation of biomass estimates in subtropical forests at Brazil. **Floresta**, v. 49, n. 3, p. 587-596, 2019.

VENDRUSCOLO, D. G. S.; CHAVES, A. G. S.; MEDEIROS, R. A.; DA SILVA, R. S.; SOUZA, H. S.; DRESCHER, R.; LEITE, H. G. Estimativa da altura de árvores de *Tectona grandis L.f.* utilizando regressão e redes neurais artificiais. **Revista Nativa**, v. 5, n. 1, p. 52-58, 2017.

VENDRUSCOLO, D. G. S.; CHAVES, A. G. S.; MEDEIROS, R. A.; SILVA, R. S.; SOUZA, H. S.; DRESCHER, R.; LEITE, H. G. Estimativa da altura de árvores de (*Tectona grandis L.f.*) utilizando regressão e redes neurais artificiais. **Revista Nativa**, Sinop, v. 5, n. 1, p. 52-58, 2017a.

VENDRUSCOLO, D. G. S.; DRESCHER, R.; SOUZA, H. S.; Moura, J. P. V. M.; MAMORÉ, F. M. D.; SIQUEIRA, T. D. S. Estimativa da altura de eucalipto por meio de regressão não linear e redes neurais artificiais. **Rev. Bras. Biom.**, São Paulo, v. 33, n. 4, p. 556-569, 2015.

VENDRUSCOLO, D.; DRESCHER, R.; SOUZA, H.; SILVA, R. Extratificação hipsométrica em plantios de eucaliptos na região sudeste de Mato Grosso. **Agrarian Academy**, Goiânia, v. 02, n. 03, p. 52-61, 2015a.

VIBRANS, ALEXANDER C. et al. Generic and specific stem volume models for three subtropical forest types in southern Brazil. **Annals of Forest Science**, v. 72, n. 6, p. 865-874, 2015.

WILKINSON, Leland; DALLAL, Gerard E. Tests of significance in forward selection regression with an F-to-enter stopping rule. **Technometrics**, v. 23, n. 4, p. 377-380, 1981.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, Washington v. 20, n.6, p. 557-585, 1921.

XIE, L.; WIDAGDO, FRA.; DONG, L.; LI, F. Modeling height–diameter relationships for mixed-species plantations of Fraxinus mandshurica Rupr. and Larix olgensis Henry in northeastern China. **Forests**, v. 11, n. 6, p. 610, 2020.

ZANG, H.; LEI, X.; ZENG, W. Height-diameter equations for larch plantations in northern and northeastern China: A comparison of the mixed-effects, quantile regression and generalized additive models. **Forestry**, v. 89, n. 4, p. 434–445, 2016.

ZERAATPISHEH, M.; AYOUBI, S.; JAFARI, A.; TAJIK, S.; FINKE, P. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. **Geoderma**, v. 338, p. 445-452, 2019.

ZHANG, B.; SAJJAD, S.; CHEN, K.; ZHOU, L.; ZHANG, Y.; YONG, K.; SUN, Y. Previsão da relação altura-diâmetro da árvore a partir dos níveis de competição relativa usando modelos de regressão quantílica para o abeto chinês (Cunninghamia lanceolata) na província de Fujian, China. **Florestas**, v. 11, n. 2, pág. 183, 2020.

ZHU, H.; YOU, X.; LIU, S. Multiple ant colony optimization based on pearson correlation coefficient. **IEEE Access**, v. 7, p. 61628-61638, 2019.

ZUFFO, A. M.; RIBEIRO, A. B. M.; BRUZI, A. T.; ZAMBIAZZI, E. V.; FONSECA, W. L. Correlações e análise de trilha em cultivares de soja cultivadas em diferentes densidades de plantas. **Revista Cultura Agronômica**, v. 27, n. 1, p. 78-90, 2018.