UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

ICA
INSTITUTO DE CIÊNCIAS AGRÁRIAS

MESTRADO EM CIÊNCIAS FLORESTAIS

Mylla Vyctória Coutinho Sousa

Modelagem hipsométrica utilizando regressão simbólica e variável ambiental

Montes Claros
2021

Mylla Vyctória Coutinho Sousa

Modelagem hipsométrica utilizando regressão simbólica e variável ambiental

Dissertação apresentada ao Mestrado em Ciências Florestais da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Ciências Florestais.

Orientador: Carlos Alberto Araújo Júnior

Coorientadores: Christian Dias Cabacinha; Renato Dourado Maia

Montes Claros
Março de 2021

Sousa, Mylla Vyctória Coutinho.

C871m 2021 Modelagem hipsométrica utilizando regressão simbólica e variável ambiental / Mylla Vyctória Coutinho Sousa. Montes Claros, 2021.
72 f.: il.

Dissertação (mestrado) - Área de concentração em Ciências Florestais. Universidade Federal de Minas Gerais / Instituto de Ciências Agrárias.

Orientador: Carlos Alberto Araújo Júnior.
Banca examinadora: Renato Vinícius Oliveira Castro, Adriana Leandra Assis.

Inclui referências: f. 22-26; 37-38; 47-49; 59-60; 68-69.

1. Florestas. 2. Inteligência artificial. 2. Levantamentos florestais. I. Araújo Júnior, Carlos Alberto. II. Universidade Federal de Minas Gerais. Instituto de Ciências Agrárias. III. Título.

CDU: 630

ELABORADA PELA BIBLIOTECA UNIVERSITARIA DO ICA/UFMG
Edélzia Cristina Sousa Versiani - Bibliotecária CRB-6 1349

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS AGRÁRIAS
MESTRADO EM CIÊNCIAS FLORESTAIS

ATA DE DEFESA DE DISSERTAÇÃO

Aos 29 dias do mês de março do ano de dois mil e vinte e um, às 08:00 horas, sob a Presidência do Professor Carlos Alberto Araújo Júnior, D. Sc. (Orientador - ICA-UFMG) e com a participação dos Professores Christian Dias Cabacinha, D. Sc. (Coorientador - ICA-UFMG), Adriana Leandra de Assis, D. Sc. (ICA/UFMG) e Renato Vinícius Oliveira Castro, D. Sc. (UFSJ), reuniu-se, por videoconferência, a Banca de Defesa de Dissertação de **MYLLA VYCTÓRIA COUTINHO SOUSA**, aluna do Curso de Mestrado em Ciências Florestais. Após a avaliação da referida aluna, a Banca Examinadora procedeu à publicação do resultado da defesa da Dissertação intitulada "**Modelagem hipsométrica utilizando regressão simbólica e variável ambiental**", sendo a aluna considerada **APROVADA**. E, para constar, eu, Professor Carlos Alberto Araújo Júnior, Presidente da Banca, lavrei a presente ata que depois de lida e aprovada, será assinada por mim e pelos demais membros da Banca examinadora.

OBS.: A aluna somente receberá o título após cumprir as exigências do ARTIGO 74 do regulamento do Curso de Mestrado em Ciências Florestais, conforme apresentado a seguir:

> **Art. 74 – Para dar andamento ao processo de efetivação do grau obtido, o candidato deverá, após a aprovação de sua Dissertação e da realização das modificações propostas pela banca examinadora, se houver, encaminhar à secretaria do colegiado do Curso, com a anuência do orientador, no mínimo 3 (três) exemplares impressos e 1 (um) exemplar eletrônico da dissertação, no prazo de 60 (sessenta) dias.**

Montes Claros, 29 de março de 2021.

Documento assinado eletronicamente por **Carlos Alberto Araujo Junior**, **Professor do Magistério Superior**, em 31/03/2021, às 20:02, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriana Leandra de Assis**, **Professora do Magistério Superior**, em 31/03/2021, às 20:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Christian Dias Cabacinha**, **Professor do Magistério Superior**, em 02/04/2021, às 12:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Renato Vinícius Oliveira Castro**, **Usuário Externo**, em 07/04/2021, às 10:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0654505** e o código CRC **3E8AEE65**.

DEDICATÓRIA

Aos meus pais, fonte de inspiração e ensinamentos e razão do meu viver.

Dedico

AGRADECIMENTO

Primeiramente a Deus, por me dar sabedoria e iluminar meus caminhos permitindo alcançar essa grande conquista.

Aos meus pais Nancy e Otávio, que sempre estiveram ao meu lado me apoiando e me ajudando.

Ao meu orientador Carlos por todos os ensinamentos, confiança, apoio e paciência.

Ao meus coorientadores (Christian e Renato) e aos membros da banca (Adriana e Renato) por todas as contribuições neste trabalho.

Aos meus colegas da liga da justiça (Patrícia, Roberta e Moreno) não somente pela amizade e momentos de descontração, como pelo apoio e aprendizado nesses dois anos.

Aos colegas do laboratório, por todos os momentos e ajuda durante o trabalho.

A Universidade Federal de Minas Gerais, pela oportunidade, infraestrutura e conhecimento que me permitiram chegar até aqui.

A CAPES pela concessão da bolsa de pesquisa, no qual gerou este trabalho.

A todos que de forma direta ou indireta contribuíram para a realização deste trabalho e que compartilharam momentos de alegria durante esse período, no qual ficará marcado por toda minha vida.

Muito obrigada!

MODELAGEM HIPSOMÉTRICA UTILIZANDO REGRESSÃO SIMBÓLICA E VARIÁVEL AMBIENTAL

RESUMO

Em florestas plantadas costuma-se medir a altura de apenas algumas árvores da parcela e utilizar da relação hipsométrica para se estimar as restantes, reduzindo dessa forma os custos do inventário florestal. Estas estimativas costumam ser feitas através de modelos de regressão e técnicas de inteligência artificial (IA). Contudo os modelos de regressão estão sujeitos a dependência de suposições estatísticas e às vezes um elevado número de equações, já as técnicas de IA estão concentradas apenas nas RNA. Outras técnicas têm surgido na comunidade cientifica, mas ainda são pouco estudadas, como é o caso da regressão simbólica (RS). Diante disso o presente trabalho objetivou verificar a viabilidade da utilização da regressão simbólica no processo de modelagem hipsométrica. Os dados para realização do estudo são provenientes de um plantio clonal de *Eucalyptus spp.,* localizado na região norte do estado de Minas Gerais. A base de dados é composta por 57 materiais genéticos, implantados em seis espaçamentos com idades variando entre 2 e 14 anos. A base de dados foi particionada em 70% para treinamento e 30% para validação. Para comparação foram ajustados 5 modelos tradicionais (Curtis, Trorey, Linear simples, Stoffels e Henricksen) e RNA, em seguida para alcançar as melhores estimativas dos dados com RS foram testadas 5 diferentes estratégias como variáveis de entrada, sendo elas E1: Dap; E2: Dap e Idade; E3: Dap, projeto, espécie e espaçamento; E4: Dap, projeto, clone e espaçamento; E5: Dap, idade, projeto e clone. Para o teste de inclusão de variáveis ambientais foi selecionado um clone amplamente distribuído em toda a área e obtido as variáveis precipitação pluviométrica e temperatura média provenientes de estações meteorológicas. Para avaliar a qualidade das estimativas, foram calculadas a correlação (r), média do erro absoluto (MAE) e Raiz quadrada do erro médio em porcentagem (RMSE%). Os principais resultados foram que o modelo gerado pela regressão simbólica, com r de 0,7861, e RMSE% de 11,72% se mostrou mais eficiente que os demais modelos e levemente inferior à RNA. A E5 com erro médio absoluto de 1,44 m apresentou os melhores valores para todas as estatísticas apresentadas. Com as variáveis qualitativas a RS apresentou r de 0,8338, MAE de 1,53 m e RMSE de 9,96%. Com as variáveis ambientais a RS apresentou r de 0,91 e RMSE de 5,49%, não apresentando ganho em precisão em relação ao modelo sem as variáveis. A regressão simbólica se mostrou um método viável e eficiente para estimativas hipsométricas, apresentando superioridade aos modelos hipsométricos tradicionais, porém quando comparada à RNA atingiu resultados semelhantes, mas não superiores. A adição de variáveis dap, idade, projeto e clone quando utilizadas em conjunto no modelo de regressão simbólica apresentaram os melhores resultados. Por se tratar de um tema inédito dentro dessa abordagem do manejo florestal, recomenda-se ainda que novos estudos sejam realizados, para que a técnica possa ser aprimorada e consolidada.

Palavras-chave: Inteligência artificial. Mensuração florestal. Programação genética.

HYPSOMETRIC MODELING USING SYMBOLIC REGRESSION AND ENVIRONMENTAL VARIABLE

ABSTRACT

In planted forests, it is customary to measure the height of only a few trees in the plot and use the hypsometric relationship to estimate the remaining ones, thus reducing forest inventory costs. These estimates are usually made using regression models and artificial intelligence (AI) techniques. However, regression models are subject to dependence on statistical assumptions and sometimes a high number of equations, while AI techniques are concentrated only on ANN. Other techniques have emerged in the scientific community, but are still poorly studied, as is the case of symbolic regression (SR). In view of this, the present study aimed to verify the feasibility of using symbolic regression in the hypsometric modeling process. The data for perform the study come from a clonal plantation of *Eucalyptus spp*., located in the northern region of the state of Minas Gerais. The database is composed of 57 genetic materials, implanted in six spacings with ages ranging among 2 and 14 years. The database was partitioned into 70% for training and 30% for validation. For comparison, 5 traditional models (Curtis, Trorey, Simple Linear, Stoffels and Henricksen) and ANN were adjusted, then, to reach the best estimates of the data with SR, 5 different strategies were tested as input variables, being them E1: Dap; E2: Dap and Age; E3: Dap, project, species and spacing; E4: Dap, project, clone and spacing; E5: Dap, age, project and clone. For the inclusion test of environmental variables, a clone widely distributed throughout the area was selected and the variables pluviometric precipitation and average temperature obtained from meteorological stations were obtained. To assess the quality of the estimates were calculated correlation (r), mean absolute error (MAE) and square root of mean error in percentage (RMSE%). The main results were that the model generated by symbolic regression, with r of 0.7861, and RMSE% of 11.72%, proved to be more efficient than the other models and slightly inferior to the ANN. The E5 with mean absolute error of 1.44 m presented the best values for all presented statistics. With the qualitative variables the SR presented r of 0.8338, MAE of 1.53 m and RMSE of 9.96%. With the environmental variables, the SR presented r of 0.91 and RMSE of 5.49%, showing no gain in precision compared to the model without the variables. Symbolic regression proved to be a viable and efficient method for hypsometric estimates, presenting superiority to traditional hypsometric models, but when compared to ANN it achieved similar results, but not superior. The addition of dap, age, project and clone variables when used together in the symbolic regression model presented the best results. As this is an unprecedented topic within this approach to forest management, further studies are recommended so that the technique can be improved and consolidated.

Key word: Artificial intelligence. Forest management. Genetic programming.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmo Genético
AM	Aprendizagem de Máquina
DAP	Diâmetro à 1,30m do solo
IA	Inteligência Artificial
k-NN	k-vizinhos mais próximos
ML	Machine Learning
PG	Programação Genética
RF	*Randon Forest*
RNA	Rede Neural Artificial
RS	Regressão Simbólica
RT	*Modified Regression Trees*
SVM	*Support Vector Machines*
SVR	*Support Vetor Regression*

SUMÁRIO

1 INTRODUÇÃO

No setor florestal, a análise de regressão é uma técnica empregada para a definição de equações que tem como objetivo estimar os valores de altura de árvores e volume de madeira (BARROS et al., 2002), dentre outras finalidades. As relações hipsométricas possibilitam obter valores da variável de interesse através de equações de regressão, o que pode reduzir o tempo e o custo na coleta de dados (SCHNEIDER; SCHNEIDER; SOUZA, 2009).

Essa técnica utiliza, geralmente, modelos de regressão linear clássicos amplamente difundidos na literatura devido a sua facilidade de ajuste. Contudo, ela está sujeita à maiores erros de estimativas quando as relações hipsométricas apresentam comportamento do tipo não linear (CALEGARIO et al., 2005; OLIVEIRA et al., 2015). Além disso, questões importantes são a geração de parâmetros com valores não significativos e a incapacidade de lidar com outliers, o que gera incertezas em relação aos valores estimados, implicando na subestimação ou superestimação da altura total das árvores (HESS et al., 2014).

Apesar de haver diversas equações com o propósito de estimar a altura das árvores em um povoamento florestal, ainda se faz necessário o desenvolvimento de novos métodos ou modelos que visem, por exemplo, a diminuição do número de equações geradas para toda a floresta, redução na quantidade de árvores medidas em campo e aumento na precisão das estimativas.

Com o avanço tecnológico, pesquisas sobre a aplicabilidade de novos modelos para estimativas da altura de árvores intensificaram-se grandemente. Vários estudos têm sido realizados nas mais variadas situações de povoamentos florestais, sempre com o intuito de melhorar as estimativas da variável de interesse. Porém, na grande maioria destes estudos, os modelos de regressão considerados são aqueles já consagrados na literatura, a exemplo do modelo de Curtis (1967). Ainda que haja um consenso de que nem sempre os mesmos modelos vão servir para qualquer espécie, região ou regime de manejo (MIGUEL et al., 2014).

Assim, nos últimos anos, ferramentas de inteligência artificial tem se destacado, a exemplo das redes neurais artificiais (MIGUEL et al., 2015; VENDRUSCOLO et al.,2017; SILVA et al., 2018). Outras técnicas de *machine learning* (ML) também tem despertado interesse da comunidade científica, como é o caso da regressão simbólica (RS) via programação genética (BOURQUE; BAYAT: ZHANG, 2019; CHAABENE; NEHDI, 2021; VALSARAJ et. al, 2020,).

Enquanto a regressão tradicional determina coeficientes de equações conhecidas, procurando-se apenas estimar os valores ótimos dos parâmetros das equações, na regressão simbólica nenhuma expressão matemática específica é necessária como ponto de partida. Nesse caso, as expressões primárias são formadas pela combinação aleatória de funções básicas (lineares ou não lineares) de variáveis de entrada com operadores numéricos. A equação definida pela abordagem será, então, aquela que melhor estimar os valores de saída em relação aos valores observados (BOURQUE; BAYAT: ZHANG, 2019).

A RS busca descobrir as estruturas do modelo e os parâmetros correspondentes utilizando as características da programação genética (CHEN; XUE; ZHANG; 2015). O método começa com um conjunto cuidadosamente escolhido de operadores, funções, variáveis e constantes matemáticas, como blocos de construção que são combinados aleatoriamente e recombinados várias vezes sucessivamente,

usando princípios do processo evolutivo (AUGUSTO; BARBOSA, 2000). Tudo isso para chegar a um conjunto de equações que provavelmente representam a dinâmica do fenômeno modelado.

O algoritmo é conduzido por uma função de adequação que, em cada estágio, mantém as melhores soluções para “reprodução e mutação” e abandona as que não são apropriadas. Tal processo continua até que um nível desejado de precisão seja alcançado (VALSARAJ et. al, 2020).

A grande vantagem desse trabalho é a apresentação de uma ferramenta de análise de dados que tem aplicação vasta, visto que vários dos problemas florestais se assemelham a esse (aproximação de função). Diante disso, vê-se como a regressão simbólica se mostra como um método promissor para estudos dentro das ciências florestais, podendo-se propor novas ferramentas para a área de mensuração e manejo florestal utilizando-se de técnicas mais avançadas, tais como a regressão simbólica.

2 OBJETIVOS

2.1 Objetivo geral

Obter equações hipsométricas a partir da utilização do aprendizado de máquina via regressão simbólica de modo que as mesmas gerem estimativas mais precisas que os modelos hipsométricos tradicionais.

2. 2 Objetivos específicos

- Apresentar metodologia para utilização da regressão simbólica;
- Obter equações via regressão simbólica;
- Comparar os resultados da regressão simbólica com os modelos tradicionais e com redes neurais artificiais;
- Avaliar o comportamento da regressão simbólica para bancos de dados com diferentes atributos**;**
- Avaliar a inclusão de variáveis ambientais na modelagem via regressão simbólica.

3 REVISÃO DE LITERATURA

3.1 Importância da variável altura

A altura de uma árvore é a distância linear ao longo do eixo principal, partindo do solo até o topo, ou até outro ponto de referência a depender do tipo de altura que se quer medir, já que na mensuração florestal ela divide-se em altura total, comercial, do fuste e da copa (MACHADO; FIGUEIREDO FILHO, 2014). Ela constitui uma importante característica da árvore e pode ser medida ou estimada. Sua medição ou estimação é muito importante para o cálculo do volume, de incrementos em altura e, em determinadas situações, pode servir como indicadora da qualidade produtiva de um local (SILVA et al., 2012).

Nos métodos estimativos, a altura também entra como uma segunda variável independente nas tabelas de volume, funções de afilamento e em algumas relações dendrométricas. Quando relacionada à idade em povoamentos equiâneos, expressa o índice de sítio, que é usado como variável independente na construção de tabelas de produção ou em funções de crescimento e produção (MACHADO; FIGUEIREDO FILHO, 2014).

Em florestas nativas, a altura total das árvores pode ter importante significado ecológico e para fins de manejo, além de ajudar a compreender a estrutura vertical da comunidade (SILVA et al., 2012). Por meio dessa variável é possível estimar as espécies que apresentam maior importância ecológica, considerando esse tipo de estrutura (SOUZA E SOUZA, 2004).

Souza e Leite (1993) apresentam metodologia para cálculo da posição sociológica tomando-se em conta as alturas das árvores amostradas na comunidade. Esses autores ainda destacam a importância de se medir a altura das árvores em estudos fitossociológicos.

3.2 Variáveis que influenciam a estimativa da altura

A medição da altura de todas as árvores na parcela, em um inventário florestal, é uma atividade que requer um elevado tempo de dedicação, pois é dificultada pela ação de alguns fatores como: dificuldade de observar o ápice da árvore na presença de um sub-bosque denso ou em espaçamentos pequenos; inclinação acentuada do terreno, e da operação inadequada dos equipamentos para medição de alturas (SANQUETTA et al., 2014).

Neste sentido, tem-se como prática medir o diâmetro de toda as árvores da parcela e a altura apenas de algumas e posteriormente estabelecer uma relação matemática que possibilite estimar as alturas das demais árvores contidas na parcela, resultando em uma sensível redução nos custos do inventário florestal e em uma operacionalização mais eficaz (RIBEIRO et al., 2010).

Muitas variáveis influenciam a estimativa de altura das árvores, dentre elas a de maior destaque é o DAP, como mostra o estudo de Silva, Xavier e Rodrigues (2007) que em seu estudo com *Eucalyptus grandis* aos 5, 6 e 7 anos de idade, concluíram que a variável DAP foi mais importante para estimar a altura do que a variável idade, sendo a influência da variável idade na estimação da altura maior nas árvores de maior DAP.

Mendonça et al. (2015) observaram que as variáveis utilizadas em seu estudo influenciaram a altura, dentre elas estão DAP, idade, área basal, índice de sítio, número de indivíduos por hectare e diâmetro

quadrático. Já a variável espaçamento exerce pouco efeito sobre a altura total, principalmente se considerando idades menores (LEITE; NOGUEIRA; MOREIRA, 2006).

A relação entre diâmetro e altura das árvores é chamada de relação hipsométrica. Essa relação é de fundamental importância nos procedimentos de inventário florestal (SOARES; PAULA NETO; SOUZA, 2011), pois permite que apenas algumas árvores tenham sua altura medida no campo, aumentando a velocidade dos levantamentos e reduzindo seus custos. Por representar um padrão biológico, diversos fatores e características do povoamento influenciam a relação hipsométrica como estrutura da floresta, idade da floresta, espécie/material genético, qualidade do sítio, variações ambientais e características qualitativas (BATISTA, 1998; FANG; BAILEY, 1998).

3.3 Modelos hipsométricos

A relação hipsométrica, se expressa corretamente por meio dos modelos de regressão, pode estimar a altura de povoamentos florestais medindo apenas o seu diâmetro, reduzindo dessa forma os custos do inventário (SOARES; PAULA NETO; SOUZA, 2011). Para a construção adequada de modelos hipsométricos devem ser observados os diversos fatores que influenciam essa relação, tais como: posição sociológica, região, idade, densidade do plantio e práticas silviculturais em geral (RIBEIRO et al., 2010).

Além disso, muitas vezes, a relação hipsométrica não expressa uma relação dendrométrica muito forte, devido à grande variabilidade das alturas em uma mesma classe de diâmetro. Nesses casos, a inclusão de variáveis qualitativas nos modelos hipsométricos é de grande importância para a obtenção de estimativas que se aproximem ao máximo da realidade. Entretanto, adicionar variáveis categóricas em modelos de regressão nem sempre resulta em ganho de exatidão, uma vez que a inclusão dessas, requer representatividade para todos os níveis das variáveis qualitativas na amostra, o que nem sempre é possível (MARTINS et al., 2016).

A precisão e a capacidade de modelar relações complexas entre variáveis são as principais características de um modelo bom e adequado. Qualquer tipo de modelo de diâmetro-altura pode nem sempre ser adequado para todos os tipos de condições onde uma espécie de árvore em particular pode ser encontrada, porque as condições do local podem afetar a relação diâmetro-altura (BAYAT et al. 2020).

A inclusão de características do povoamento nos modelos hipsométricos, como índice de local e idade, resulta em vantagens, como a obtenção de estimativas mais precisas e o maior realismo biológico, tornando a equação aplicável em diferentes locais. As principais dificuldades para a inclusão de muitas variáveis em um modelo hipsométrico são a dificuldade de modelagem e quantificação das influências sobre a variável a ser estimada, pois, essas relações apresentam características não lineares ou valores categóricos, a exemplo do tipo de solo, podendo ser incluídas em regressões somente como variáveis binárias e ocasionar aumento na complexidade de modelagem. (BINOTI; BINOTI; LEITE, 2013).

As relações hipsométricas podem ser classificadas em locais e gerais. As locais são funções apenas do DAP e podem ser aplicadas apenas para o povoamento em que os dados foram medidos ou em povoamento com características homogêneas. As relações hipsométricas gerais, além de recorrerem ao DAP, também exprimem a altura da árvore em função da altura dominante, diâmetro dominante, densidade, idade e outras características (TOMÉ; RIBEIRO; FARIAS, 2007).

Os modelos hipsométricos se dividem em lineares e não lineares. Os modelos lineares descrevem a variável Y como a soma de uma quantidade determinística (X) e uma quantidade aleatória (ε) de inúmeros fatores que podem, conjuntamente, ter influência sobre Y e podem assumir as seguintes formas: lineares simples e lineares múltiplos (CHARNET et al., 2008). Os modelos não lineares são aqueles que possuem seus parâmetros agregados na forma não aditiva e podem ser classificados em linearizáveis e não linearizáveis (DRAPER; SMITH, 1981).

Modelos lineares têm sido muito utilizados para a relação hipsométrica, principalmente devido à facilidade de ajuste (FERREIRA, 2009), pois os modelos não lineares apresentam dificuldade de ajuste e a necessidade de muitos recursos computacionais para tal (SENA et al. 2015). Porém, nem sempre são os mais adequados para expressar esta relação, já que, muitas vezes, esta apresenta-se como não linear (CALEGARIO et al., 2005).

A relação hipsométrica pode ser afetada por uma série de fatores ambientais, os quais tem uma relação não linear no seu comportamento, e por isso equações tradicionais podem não simular com precisão a relação entre estas variáveis (SHEN et al., 2020)

Diversos modelos hipsométricos são encontrados na literatura, dentre esses merece destaque o modelo de Trorey, modelo da linha reta, modelo de Stoffels, modelo de Curtis, modelo de Henriksen, modelo de Gompertz, modelo de Chapman e Richards, modelo de Prodan, modelo de Naslund e modelo de Meyer.

A grande dificuldade da escolha do melhor modelo para representar essas relações hipsométricas se deve à não linearidade da relação entre as variáveis envolvidas e as restrições impostas aos parâmetros dos modelos, por razões biológicas (BARTOSZECK et al., 2004; SOARES et al., 2004).

3.4 Modelagem hipsométrica com inteligência artificial

A partir da década de 1970, técnicas baseadas em inteligência artificial começaram a ser usadas de forma mais ampla na solução de problemas reais (RUSSELL e NORVIG, 2010).

Quando se trata de povoamentos em diferentes arranjos de plantio é necessário desenvolver e aplicar equações específicas, o que acaba se tornando insustentável quando a quantidade de amostragem é muito grande indicando-se assim optar por técnicas de inteligência artificial para otimizar o gerenciamento dos povoamentos florestais (CERQUEIRA et al., 2018).

Recentemente, estudos na área de Inteligência Artificial (IA), mostraram que essas técnicas se adequam ao tipo de predição feito na mensuração florestal. Trabalhos como o de Martins et al. (2016), Campos et al. (2016) e Vieira et al. (2018) mostram estimativas de altura feitas com o uso de inteligência artificial.

3.5 *Machine Learning*

O aprendizado de máquina (ou do inglês, *Machine Learning*) é um ramo da inteligência artificial que induz hipóteses a partir de um conjunto de dados, que é a experiência passada de algum problema (FACELI et al., 2011). Aprendizado de máquina (AM), portanto, é responsável por desenvolver técnicas computacionais sobre o aprendizado e construir sistemas capazes de adquirir conhecimento de forma autônoma (REZENDE, 2003).

O AM visa permitir que as máquinas realizem seus trabalhos com habilidade usando softwares inteligentes. Os métodos estatísticos de aprendizado constituem a espinha dorsal do software inteligente usado para desenvolver a inteligência da máquina. Como os algoritmos de aprendizado de máquina exigem que os dados sejam aprendidos, a disciplina deve ter conexão com a disciplina do banco de dados (MOHAMMED; KHAN; BASHIER, 2017).

No AM geralmente o conjunto de dados é dividido em dois sub conjuntos: treinamento e validação. Enquanto, o desempenho do conjunto de treinamento fornece uma ideia da capacidade de generalização do modelo, os conjuntos de validação formam um conjunto independente de dados para determinar a previsibilidade (KUMAR et al., 2014),

Os algoritmos de AM possuem parâmetros específicos que devem ser configurados para a realização da etapa de treinamento. De modo geral, os modelos de AM não produzem resultados ótimos sem que estes parâmetros sejam apropriadamente configurados, necessitando de métodos de busca que minimizam o erro do modelo (BERGSTRA e BENGIO, 2012).

Os algoritmos de AM podem ser estruturados em paradigmas, tais como: simbólico, estatístico, baseado em exemplos, conexionista e evolutivo (REZENDE, 2003). Exemplos desses algoritmos são C4.5, *Support Vector Machines* (SVM), k-vizinhos mais próximos (k-NN), redes neurais artificiais (RNA), algoritmos genéticos e *Random Forest*.

Estudos na área de mensuração florestal têm utilizado métodos de aprendizagem de máquina (AM) principalmente em tarefas de aproximação de funções, sendo as redes neurais artificiais os mais estudados, apresentando em diversos casos performance superior aos modelos de regressão tradicionais na modelagem de variáveis como volume (BINOTI; BINOTI; LEITE, 2014; GORGENS et al., 2014) e altura total (VENDRUSCOLO et al., 2016; COSTA FILHO et al., 2019).

Outros modelos de AM promissores ainda são pouco investigados quanto às aplicações na modelagem de atributos florestais, no entanto, têm sido amplamente estudados em outras áreas como a de sensoriamento remoto (ABDOLLAHNEJAD et al., 2017; BLANCHETTE et al., 2015; GORGENS; MONTAGHI; RODRIGUEZ, 2015).

3. 5. 1. Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs) são sistemas paralelos compostos por elementos de processamento simples (neurônios) que calculam funções matemáticas. Esses elementos são dispostos em uma ou mais camadas e interligados entre si por meio de conexões associadas a pesos. O procedimento inicial é uma fase chamada de aprendizagem, que consiste em extrair características para representar a informação e armazenar o conhecimento. Depois do aprendizado, é feita a generalização, que nada mais é do que aplicar a RNA em dados não conhecidos (BRAGA; CARVALHO; LUDEMIR, 2007).

Atualmente RNA vêm sendo empregadas para estimar altura de árvores em povoamentos florestais. No Brasil, as RNA têm sido utilizadas por empresas florestais, com as vantagens de simplificação nas rotinas de coleta e processamento de dados, diminuição no número de medições de alturas e consequente redução do custo do inventário (CAMPOS; LEITE, 2017).

Quanto às desvantagens de uso da técnica, tem-se que a configuração ótima associada às RNA não são simples de ser encontrados, além de, a configuração especificamente, ser um processo que

demanda muito tempo. O uso de redes neurais ainda é criticado também pela falta de interpretabilidade dos pesos obtidos durante o processo de construção do modelo. Comparativamente a outros modelos estatísticos que se destacam por permitir a interpretação dos coeficientes das variáveis individuais (PALIWAL; KUMAR, 2009).

Em alguns casos, as RNA têm apresentado desempenho superior aos modelos de regressão devido a diversos fatores, como: estrutura maciça e paralelamente distribuída (camadas); habilidade de aprender e generalizar, que as tornam capazes de resolver problemas complexos; são tolerantes a falhas e ruídos; podem modelar diversas variáveis e suas relações não lineares; possibilidade de modelagem com variáveis categóricas (qualitativas), além das numéricas (quantitativas); e analogia neurobiológica (HAYKIN, 2001).

Vários trabalhos demonstram os bons resultados alcançados a partir da utilização de RNA em estimativas florestais, tais como estimativas de altura total (DIAMANTOPOULOU, 2012; BINOTI; BINOTI; LEITE, 2013; BINOTI et al., 2013; ÖZÇELIK at al., 2013).

3.5.2 Algoritmos Genéticos

Algoritmos Genéticos (AG) são definidos como uma técnica de busca baseada no processo seleção natural, ou seja, são uma técnica heurística de otimização global. Nos AG populações de indivíduos são criados e submetidos aos operadores: seleção, recombinação e mutação (LINDEN, 2008).

Em vez de processar uma única solução por vez, os algoritmos genéticos trabalham com uma população de soluções experimentais. Em cada geração a população atual é formada pelo conjunto de soluções consideradas. Então alguns integrantes sobrevivem e se tornam pais que depois têm filhos (novas soluções) que compartilham características de ambos os pais. Finalmente os indivíduos mais adaptados sobrevivem, levando o AG a uma solução próxima da ótima (HILLIER e LIBERMAN, 2013).

Uma desvantagem dos algoritmos genéticos é que a divisão e a recombinação dos pais geralmente levam a um número elevado de soluções inviáveis quando há muitas restrições (BETTINGER et al., 2009).

No meio florestal os algoritmos genéticos são geralmente utilizados na área de planejamento e otimização florestal como é o caso dos trabalhos de Jin, Pukkala e Li (2016), Matos et al. (2019) e MirarabRazi et al. (2020).

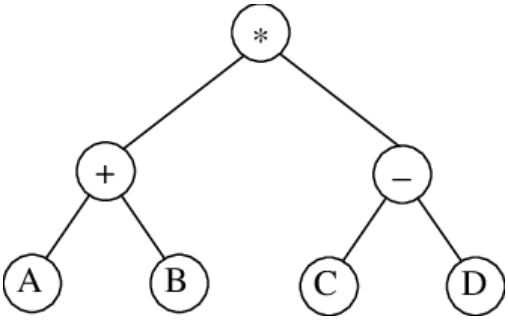
3.5.3 Regressão simbólica

A regressão simbólica (RS) difere da regressão tradicional, pois não depende de uma estrutura de modelo determinada previamente. A única suposição feita na RS é que a superfície de resposta pode ser descrita por uma expressão algébrica. Em vez da abordagem tradicional em que a estrutura do modelo é fixado e os demais parâmetros livres são otimizados, na RS o problema de regressão é reformulado como um problema de pesquisa para encontrar a estrutura ideal do modelo. Uma vez encontrada uma estrutura de modelo de qualidade suficiente, técnicas tradicionais podem ser usadas para encontrar os coeficientes ideais (MINEBO; STIJVEN, 2011).

Um algoritmo comum usado para gerar modelo de regressão simbólica é a programação genética que é um algoritmo de busca baseado em populações que são representadas por árvores de expressão, onde cada nó pode representar uma função, um preditor ou um valor constante (FIGURA 1). Com esta

configuração, pode-se representar qualquer função de todo o espaço de busca de expressões matemáticas (FRANÇA; LIMA, 2021).

Figura 1 - Ilustração indicando um exemplo de árvore de busca gerada durante a execução de um algoritmo de programação genética.



Fonte: MATTOS; CASTRO, 2015.

A Programação Genética (PG) é a técnica mais popular usada na regressão simbólica e foi inventada por Cramer (1985) e desenvolvida por Koza (1992). Esta pode ser considerada como uma extensão dos algoritmos genéticos (AG) em que a definição fixa do problema (principal limitação do AG) é evitada com a ajuda de árvores de comprimento variável em vez de indivíduos de tamanho fixo (KOZA, 1994).

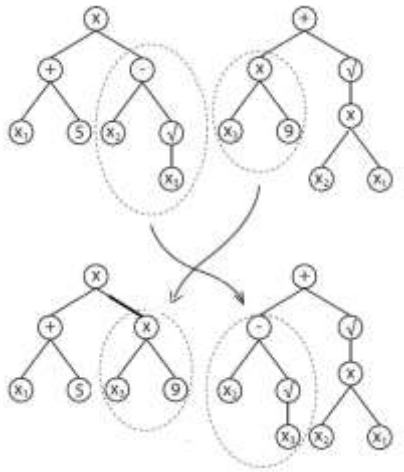
A PG utiliza o princípio Darwiniano de seleção natural juntamente com a recombinação genética, chamada de *crossover*, para criar populações geneticamente superiores. Essa operação envolve o cruzamento dos indivíduos selecionados. Então, a nova população será composta por partes da antiga, sendo obtidas após as operações de cruzamento e reprodução (KOZA, 1992). Isso geralmente é feito por meio do torneio, em que pares de soluções são amostradas aleatoriamente para disputar quem será escolhido para a próxima geração. Os novos descendentes passam a substituir a antiga geração e em seguida, faz-se a avaliação da nova população, repetindo-se o processo por várias gerações (FIGURA 2).

Figura 2 - Ilustração do processo de execução do algoritmo geral de programação genética

Fonte: Elaborado pela autora, 2021

Este algoritmo se baseia em duas operações principais: cruzamento e mutação. O primeiro combina duas árvores de expressão para formar uma nova expressão (FIGURA 3) e o último faz pequenas alterações em uma árvore de expressão (FIGURA 4). A ideia principal é que o operador de cruzamento tenha a chance de melhorar uma solução misturando partes de um conjunto de boas soluções e, o operador de mutação irá explorar a vizinhança de uma determinada expressão (FRANÇA; LIMA, 2021)

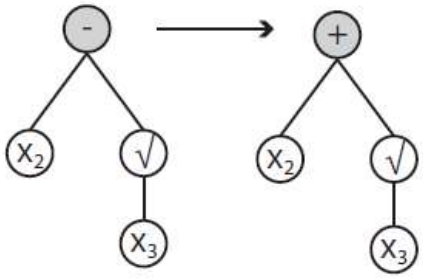
Figura 3 - Ilustração de uma operação de cruzamento durante o processo de execução do algoritmo de regressão simbólica



Fonte: Adaptado de MINEBO; STIJVEN, 2011.

Dessa forma, há uma maior flexibilidade para o ajuste dos dados, o que também é um desafio, já que o espaço de busca é enorme e com muitos locais ótimos, podendo às vezes exigir um esforço computacional maior para alcançar bons resultados (LA CAVA; MOORE 2019).

Figura 4 - Ilustração de uma operação de mutação durante o processo de execução do algoritmo de regressão simbólica



Fonte: MINEBO; STIJVEN, 2011.

Vários estudos mostram a sua utilização em diferentes áreas como o trabalho de Barmpalexis et al. (2011) que utilizaram a regressão simbólica via programação genética para otimização de um tablet com matriz de liberação zero de ordem farmacêutica e compararam seu desempenho preditivo aos dos modelos de redes neurais artificiais (RNA). Para isso utilizaram dois tipos de algoritmo de PG, como resultado verificou-se que a capacidade de predição da PG em um conjunto de validação externa foi maior em comparação com a das RNAs.

Já Karaboga et al. (2012) usaram um conjunto de problemas de referência de regressão simbólica com programação de colônias de abelhas artificiais e compararam com o método de programação genética. Os resultados da simulação indicam que o método proposto é muito viável e robusto nos problemas de teste considerados de regressão simbólica.

Ghosh, Behera e Paramanik (2020) estimaram a altura do dossel da floresta a partir de dados de SAR e sensoriamento remoto óptico usando dois modelos de aprendizado de máquina *Random Forest* e regressão simbólica. Os resultados mostraram que a coerência interferométrica e as variáveis biofísicas apresentam correlação razoável com a altura do dossel e que o modelo RS por programação genética apresentou os melhores resultados, mostrando-se como uma alternativa plausível para a estimativa da altura do dossel da floresta quando há ausência de fontes de dados comerciais.

Valsaraj et al. (2020) propuseram um novo método de aplicação de regressão simbólica aos dados de velocidade do vento para obter uma função simbólica capaz de estimar a velocidade do vento em grandes altitudes usando dados da velocidade do vento em altitudes mais baixas em diferentes locais. O novo método mostrou desempenho mais preciso em diferentes estações do ano, tanto em locais de referência quanto em locais distantes quando comparado com o método tradicional da lei de energia.

Ainda pode-se citar outros trabalhos como os de Kumar et al. (2014) e Chaabeni e Nedhi (2021) na construção civil; Bahrami et al., (2016) na engenharia de petróleo; Astarabadi; Ebadzadeh (2019) e França (2018) nas ciências da computação; Kodali et al., (2018) no meio ambiente e Bourque, Bayat e Zhang (2019) na gestão florestal.

3.5.3.1 Critérios e parâmetros de controle

A execução da técnica depende de um conjunto de parâmetros, cujos valores são normalmente escolhidos no início do processo. A seleção de diferentes valores para cada parâmetro afeta diretamente a evolução do sistema e os recursos consumidos (KOZA, 1992).

O tamanho da população e o número de gerações são os parâmetros mais utilizados para controlar o processo. O primeira influencia diretamente no rendimento e no esforço computacional necessário para produzir resultados, e o segundo serve como um critério de parada, caso nenhum outro critério seja estabelecido. Maiores populações também podem significar um maior espaço de busca e, portanto, maior probabilidade de encontrar boas soluções. O tamanho das escolhas depende da dificuldade do problema envolvido e por isso podem ser necessários valores maiores quando a estrutura da solução é considerada complexa (KOZA, 1992).

A recombinação das características para a próxima geração ocorre por meio da aplicação de um operador genético, sendo os principais o cruzamento (*crossover*) e a mutação. A probabilidade de ocorrência de cada um deles é chamado de “taxa”. A taxa de cruzamento é o percentual de reprodução no qual trechos dos indivíduos são intercambiados e a taxa de mutação é a probabilidade dos indivíduos terem partes substituídas. Ambas as operações afetam a composição da próxima geração e ajudam a manter a variabilidade das soluções da próxima geração (KOZA, 1992).

Outro parâmetro que afeta a variação no tamanho da solução é a profundidade máxima permitida nas gerações inicial e subsequentes. Restringir a profundidade significa que a árvore representativa da solução não pode conter mais do que um certo número de nós definidos como limite em sua ramificação

mais profunda. Árvores com maior profundidade significam soluções com mais pontos. Em muitas aplicações, aumentar o número de pontos é importante para resolver o problema, porém o aumento excessivo consumirá recursos ilimitados e prejudicará a convergência (KOZA, 1992).

O método de inicialização especifica de que forma é criada a primeira geração, para que ela tenha uma boa diversidade no que diz respeito ao tamanho dos indivíduos. Os métodos mais utilizados são o *full, o grow e o ramped-half-and-half*. A estratégia *full* cria soluções nas quais no mínimo um dos nodos atinge a profundidade máxima. O *grow* produz indivíduos com tamanhos irregulares. No *ramped-half-and-half*, são criados indivíduos com profundidades distribuídas entre tamanhos mínimos e máximos. Como exemplo podemos supor que a profundidade máxima é 6, neste caso ocorre uma distribuição igual de indivíduos com profundidades 2, 3, 4, 5 e 6. Esta última alternativa permite uma boa variabilidade na população sem soluções de tamanhos exagerados e é, portanto, a escolha mais comum (KOZA, 1992).

Outra escolha a ser feita no início do processo envolve o método de seleção de indivíduos para operações genética. Existem várias variações possíveis, entretanto as mais comuns são: Roleta ou proporcional ao *fitness* e torneio. Na primeira, todos recebem um valor proporcional ao seu *fitness*. Quanto melhor for o *fitness*, maior será a probabilidade de escolhê-lo. A seleção por torneio envolve a seleção aleatória de um certo número de indivíduos (tamanho do torneio) para formar subgrupos temporários. O indivíduo selecionado desse grupo é o indivíduo com maior valor de *fitness*. É um dos métodos mais utilizados por oferecer a vantagem de não exigir a comparação entre todos os indivíduos (BANZHAF et al., 1998).

3.6 Referências

ABDOLLAHNEJAD, A. PANAGIOTIDIS, D.; JOYBARI, S. S.; SUROVY, P. Prediction of Dominant Forest Tree Species Using QuickBird and Environmental Data. **Forests,** v. 8, n. 2, p.42-60, 2017.

ASTARABADI, S. S. M.; EBADZADEH, M. M. Genetic programming performance prediction and its application for symbolic regression problems. **Information Sciences.** v. 502, 2019, p. 346–362, 2019.

AUGUSTO, D. A.; BARBOSA H. J. C. Symbolic regression via genetic programming. In: **Proceedings**. Vol. 1. Sixth Brazilian symposium on neural networks, IEEE, 2000. p. 173–8.

BAHRAMI, P.; KAZEMI, P.; MAHDAVI, S.; GHOBADI, H. A novel approach for modeling and optimization of surfactant/Polymer flooding based on Genetic Programming evolutionary algorithm. **Fuel**, v.179, p.289-298, 2016.

BANZHAF, W.; NORDIN, P.; KELLER R.; FRANCONE, F.D. **Genetic Programming An Introduction**: On the Automatic Evolution of Computer Programs and Its Applications. San Francisco:Morgan Kaufmann, 1998. 470p.

BARMPALEXIS, P; KACHRIMANIS, K.; TSAKONAS, A.; GEORGARAKIS, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. **Information Sciences**, v. 107, n. 1, p. 75-82. 2011.

BARROS, D. A.; MACHADO, S. A.; ARCEBI JÚNIOR, F. W.; SCOLFORO, J. R. S. Comportamentos de modelos hipsométricos tradicionais e genéricos para plantações de *Pinus oocarpa* em diferentes tratamentos**. Boletim de Pesquisa Florestal**, v. 45, n. 1, p. 3-28, 2002.

BARTOSZECK, A. C. P. S.; MACHADO, S. A.; FIGUEIREDO FILHO, A.; OLIVEIRA, E. B. Dinâmica da relação hipsométrica em função da idade, do sítio e da densidade inicial de povoamentos de Bracatinga da Região Metropolitana de Curitiba, PR. **Revista Árvore**, v.28, n.4, p.517-533, 2004.

BATISTA, J. L. F. Mensuração de árvores: Uma introdução à Dendrometria. Departamento de Ciências Florestais, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 45p, 1998.

BAYAT, M.; GHORBANPOUR, M.; ZARE, R.; JAAFARI, A.; PHAM, B.T. Application of artificial neural networks for predicting tree survival and mortality in the Hyrcanian forest of Iran. **Computers and Electronics in Agriculture**, v.164, 2019.

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

BETTINGER P.; BOSTON K.; SIRY L.P; GREBNER R.L. **Forest management and planning**. 1. ed. Londres: Academic. 2009. 331 p.

BINOTI, D. H. B.; BINOTI, M. L. M. DA S.; LEITE, H. G.; SILVA, A. Redução dos custos em inventário de povoamentos equiâneos. **Revista Brasileira de Ciências Agrárias**, v.8, n.1, p.125-129, 2013.

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

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

BLANCHETTE, D. FOURNIER, R. LUTHER, J. E.; COTE, J. F. Predicting wood fiber attributes using local-scale metrics from terrestrial LiDAR data: A case study of Newfoundland conifer species. **Forest Ecology and Management**, v. 347, p.116-129, 2015.

BOURQUE, C. P. A.; BAYAT, M.; ZHANG, C. An assessment of height–diameter growth variation in an unmanaged Fagus orientalis-dominated forest. **European Journal of Forest Research**, v.138, p. 607–621, 2019.

BRAGA, A. P.; CARVALHO, A. P. L. F; LUDEMIR. T. B. **Redes Neurais Artificiais**: teoria e aplicações. 2 ed. Rio de Janeiro: LTC, 2007.

CALEGARIO, N.; CALEGARIO, C. L. L.; MAESTRI, R.; DANIELS, R.; Melhoria da qualidade de ajuste de modelos biométricos florestais pelo emprego da teoria dos modelos não-lineares generalizados. **Scientia Forestalis**, v. 69, p. 38-50, 2005.

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

CAMPOS, J. C. C.; LEITE, H. G. **Mensuração florestal**: perguntas e respostas. 5.ed. Viçosa: UFV, 2017. 636p.

CERQUEIRA, C. L.; MORA, 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.

CHAABENI, W. B.; NEDHI, M. L. Genetic programming based symbolic regression for shear capacity prediction of SFRC beams. **Construction and Building Materials,** v. 280, p. 1-14, 2021.

CHARNET, R, FREIRE, C. A.; CHARNET E. M. R.; BONVINO H. **Análise de modelos de regressão linear com aplicações**. 2. ed. Campinas: Unicamp, 2008, 356p.

CHEN QI.; XUE BING.; ZHANG M. Generalisation and domain adaptation in gp with gradient descent for symbolic regression. In: **Evolutionary computation** (CEC), 2015 IEEE congress on, IEEE, 2015. p. 1137–44.

COSTA FILHO, S. V. S.; ARCE, J. E.; MONTAÑO, R. N. R.; PELISSARI, A. L. 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, 2019.

CRAMER, N. L. A representation for the adaptive generation of simple sequential programs, in: **Proceedings** of the 1st International Conference on Genetic Algorithms and their Applications, Erlbaum, 1985.

CURTIS, R. O. Height diameter and height diameter age equations for second growthDouglas-fir. **Forest Science**, v.13, n.4, p.365-375, 1967.

DIAMANTOPOULOU, M. J. Assessing a reliable modeling approach of features of trees through neural network models for sustainable forests. **Sustainable Computing: Informatics and Systems**, v.2, n.4, p.190-197, 2012

DRAPER, N. M.; SMITH, H. **Applied Regression Analysis**. 2.ed., New York: Wiley. 1981, 709p.

FACELI, R.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC. 2011.394p.

FERREIRA, D. F. **Estatística básica**. Lavras: Editora UFLA, 2ª Ed. 2009. 664p

FRANÇA, F. O A greedy search tree heuristic for symbolic regression. **Information sciences**, v. 442, n. 1, p. 18-32. 2018.

FRANÇA, F. O; LIMA, M. Z. Interaction-transformation symbolic regression with extreme learning machine. **Neurocomputing**, v.423, p. 609-619, 2021.

GORGENS, E. B.; LEITE, H. G.*;* GLERIANI. J. M.; SOARES, C. P. B.; CEOLIN, A. Influência da arquitetura na estimativa de volume de árvores individuais por meio de redes neurais artificiais. **Revista Árvore,** v. 38, n. 2, p.289-295, 2014.

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

HAYKIN, S. **Redes neurais:** princípios e prática. Porto Alegre: 2001. 900p

HESS, A. F.; MUÑOZ, E. B.; THAINES. F.; MATTOS, P. P. Adjustment of the hypsometric relationship for species of Amazon Forest. **Ambiência**, v. 10, n. 1, p. 21–29, 2014.

HILLIER F.S.; LIEBERMAN G.J. **Introdução à pesquisa operacional**. 9. ed. São Paulo: MacGraw-Hill. 2013. 1006 p.

JIN X.; PUKKALA T.; LI, F. Fine-tuning heuristic methods for combinatorial optimization in forest planning. **European Journal of Forest Research**, v. 135, n.1, p. 765-779, 2016.

KARABOGA, D.; OZTURK, C.; KARABOGA, N.; GORKEMLI, B. Artificial bee colony programming for symbolic regression. **Information Sciences**, v. 209, n.1, p. 1–15, 2012.

KODALI, A.; SZUBERT, M.; DAS, K.; GANGULY, S.; BONGARD, J. Understanding climate-vegetation interactions in global rainforests through a GP-tree analysis. **Parallel Problem Solving from Nature (PPSN),** p. 525-536, 2018.

KOZA, J. R**.** Genetic progamming as means for programming computers by natural selection. **Statistics and Computing**, v. 4 p.87–112, 1994.

KOZA, J. R. **Genetic programming**: on the programming of computers by means of natural selection. Cambridge: The MIT Press, 1992. 819 p.

KUMAR, B.; JHA, A.; DESHPANDE, V. SREENIASULU, G. Regression model for sediment transport problems using multi-gene symbolic genetic programming. **Computers and Electronics in Agriculture.** v. 103, p. 82–90, 2014.

LA CAVA, W.; MOORE, J. H. Semantic variation operators for multidimensional genetic programming, in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '19, ACM, New York, NY, USA, 2019, p. 1056–1064. doi:10.1145/3321707.3321776.

LEITE, H.G.; NOGUEIRA, G. S; MOREIRA, A. M. Efeito do espaçamento e da idade sobre variáveis de Povoamentos de *Pinus taeda* L. **Revista Árvore,** Viçosa, v.30, n.4, p.603-612, 2006.

LINDEN, R. **Algoritmos genéticos**: uma importante ferramenta de inteligência computacional. 2. ed. Rio de Janeiro: Brasport, 2008. 402p..

MACHADO, S. A.; FIGUEIREDO FILHO, A. **Dendrometria**. 2 ed. Guarapuava: UNICENTRO, 2014. 316p.

MARTINS; E. R.; BINOTI, M. L. M.; LEITE, H. G.; BINOTI, D. H. B. DUTRA, G. C. Configuração de redes neurais artificiais para estimação da altura total de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias (Agrária)**, Recife, v.11, n.2, p.117-123, 2016.

MATOS, L. M. A.; ARAÚJO JÚNIOR, C. A.; ASSIS, A. L.; CABACINHA, C. D.; FERREIRA, P. H. B.; MAGALHAES, E. C. Influência dos parâmetros da metaheurística algoritmo genético em um problema de planejamento florestal. **Advances in forestry science**, v. 6, p. 767-774, 2019.

MENDONÇA, A. R.; CORADIN, C. M.; PACHECO, G. R. VIEIRA, G. C.; ARAÚJO, M. S. INTERAMNENSE, M. T. Modelos hipsométricos tradicionais e genéricos para *Pinus caribaea* var. *hondurensis.* **Pesquisa florestal brasileira**, v. 35, n. 81, p. 47-54, 2015.

MIGUEL, E. P.; LEAL, F.A.; ONO, H. A.; LEAL, U.A.S. Modelagem na predição do volume individual em plantio de *Eucalyptus urograndis.* **Revista Brasileira de Biometria**, v.32, n.4, p. 584-598, 2014.

MIGUEL, E. P.; REZENDE, A. V.; LEAL, F. A.; MATRICARDI, E. A. T.; VALE, A. T.; PEREIRA, R. S. Redes neurais artificiais para a modelagem do volume de madeira e biomassa do cerradão com dados de satélite. **Pesquisa Agropecuária Brasileira**, v. 50, n. 9, p.829-839, 2015.

MINEBO, W.; STIJVEN, S**. Empowering Knowledge Computing with Variable Selection**. 2011. 135 f. Dissertação (Mestrado) – Universidade de Antwerpen. 2011.

MIRARABRAZI, j.; NAVRODI, i. h.; GHAJAR, I.; SALAHI, M. Identifying optimal location of ecotourism sites by analytic network process and genetic algorithm (GA): (Kheyroud Forest). **International Journal of Environmental Science and Technology**, v. 17, p.2583–2592, 2020

MOHAMMED, M.; KHAN, M. B.; BASHIER, E. B. M. **Machine learning**: algorithms and applications. Boca Raton: CRC Press, 2017.

OLIVEIRA, G. M. V.; MELLO, J.M. M.; ALTOÉ. T. F.; SCALON. J. D.; SCOLFORO, J. R. S.; PIRES. J. V. Equações hipsométricas para *Eucalyptus* spp. não manejado em idade avançada com técnicas de inclusão de covariantes. **Cerne**, v. 21, n. 3, p. 483–492, 2015.

ÖZÇELIK, R.; DIAMANTOPOULOU, M. J.; CRECENTE-CAMPO, F.; ELER, U. Estimating *Crimean juniper* tree height using nonlinear regression and artificial neural network models. **Forest Ecology and Management**, v.306, p.52–60, 2013.

PALIWAL, M.; KUMAR, U. A. Neural networks and statistical techniques: A review of applications. **Expert Systems with Applications**, v.36, n.1 p.2-17, 2009.

REZENDE, S. O. **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri: Manole. 2003.

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**, Lavras, v. 16, n. 1, p. 22-31, 2010.

RUSSELL, S.; NORVIG, P. **The Artificial Intelligence**. Prentice Hall Press, Upper Saddle River, NJ, USA, 3 ed. 2010.

SANQUETTA, C. R.; WATZLAWICK, L. F.; DALLA CORTE, A. P.; FERNAND, L. V. **Inventários florestais:** planejamento e execução. 3 ed. Curitiba: Multi-Graphic Gráfica e Editora, 2014. 406p.
SENA, A. L. M.; SILVA NETO, A. J.; OLIVEIRA, G. M. V.; CALEGARIO, N. Modelos lineares e não lineares com uso de covariantes para relação hipsométrica de duas espécies de pinus tropicais. **Ciência Florestal**, v.25, n.4, p. 969-980, 2015.

SCHNEIDER, P. R.; SCHNEIDER, P. S. P.; SOUZA, C. A. M. **Análise de regressão aplicada a engenharia florestal**. 2. ed. Santa Maria: FACOS-UFSM, 2009. 294 p.

SHEN, J.; HU, Z.; SHARMA, R. P.; WANG, G.; MENG, X.; WANG, M.; WANG, Q.; FU, L. Modeling height–diameter relationship for poplar plantations using combined-optimization multiple hidden layer back propagation neural network. **Forest**, v.11, n.4, p. 1-19, 2020.

SILVA, J. P. M.; CABACINHA, C. D.; ASSIS, A. L.; MONTEIRO, T. C.; ARAÚJO JÚNIOR, C. A.; MAIA, R. D. Redes neurais artificiais para estimar a densidade básica de madeiras do cerrado. **Pesquisa Florestal Brasileira**, *[S. l.]*, v. 38, 2018. DOI: 10.4336/2018.pfb.38e201801656.

SILVA, G. F.; CURTO, R. A.; SOARES, C. P.B.; PIASSI, L. C.; Avaliação de métodos de medição de altura em florestas naturais. **Revista Árvore**, v.36, n.2, p.341-348, 2012.

SILVA, G. F.; XAVIER, A. C.; RODRIGUES, F. L. Análise da influência de diferentes tamanhos e composições de amostras no ajuste de uma relação hipsométrica para *Eucalyptus grandis*. **Revista Árvore**, v.31, n.4, p.685-694, 2007.

SOARES, C. P. B.; PAULA NETO, F.; SOUZA, A. L. **Dendrometria e inventário florestal**. Viçosa: Editora UFV, 2011. 272p.

SOARES, T. S. SCOLFORO, J. R. FERREIRA, S. O.; MELLO, J. M. Uso de diferentes alternativas para viabilizar a relação hipsométrica no povoamento florestal. **Revista Árvore**, v.28, n.6, p.845-854, 2004.

SOUZA A. L.; SOUZA, D. R. Estratificação vertical em floresta ombrófila densa de terra firme não explorada, Amazônia Oriental. **Revista Árvore**, v.28, n.5, p.691-698, 2004.

SOUZA, A. L.; LEITE, H. G. **Regulação da produção em florestas inequiâneas.** Viçosa, MG: Universidade Federal de Viçosa, 1993. 147p.

STOLLE, L.; VELOZO, D. R.; DALLA CORTE, A. P.; SANQUETA, R. C.; BEUTLING, A. Modelos hipsométricos para um povoamento jovem de *Khaya ivorensis* A.Chev. **Biofix ScientificJournal** v. 3 n. 2 p. 231-236 2018.

TOMÉ, M.; RIBEIRO, F.; FAIAS, S. Relação Hipsométrica geral para *Eucalyptus globulus* Labill. em Portugal. **Silva Lusitana**, v.15, n.1, p. 41-55, 2007.

VALSARAJ, P.; THUMBA, D. A.; ASOKAN, K. KUMAR, K. S.; Symbolic regression-based improved method for wind speed extrapolation from lower to higher altitudes for wind energy applications. **Applied Energy,** v. 260, n.1, p. 1- 10, 2020.

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

VENDRUSCOLO, D. G. S.; DRESCHER, R.; CARVALHO, S. P. C.; MEDEIROS, R. A.; SOUZA, H. S.; CERQUEIRA, C. L.; MOURA, J. P. V.; LEITE, H.G. Height prediction of Tectona grandis trees by mixed effects modelling and artificial neural networks**. International Journal of Current Research**, v. 8, n. 12, p.43189-43195, 2016.

VIEIRA, G. C.; MENDONÇA, A. R.; SILVA, G. F.; ZANETTI, S. S.; SILVA, M. M.; SANTOS, A. R. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment,** v. 619, p. 1473–1481, 2018.

4 ARTIGOS

4.1 Artigo 1- Modelagem hipsométrica via regressão simbólica

RESUMO

A altura é uma variável importante para o manejo florestal, por isso estima-la de forma precisa é fundamental. Recentemente, técnicas de inteligência artificial têm sido usadas para isso, mesmo assim diversas técnicas ainda não foram testadas como é o caso da regressão simbólica. Nesse contexto, o objetivo deste trabalho é avaliar a qualidade dos resultados produzidos estimando-se a altura total das árvores usando o método de regressão simbólica via programação genética. Os dados para realização do estudo são provenientes de um plantio clonal de *Eucalyptus* spp*.,* localizado na região norte do estado de Minas Gerais. A base de dados é composta por 57 materiais genéticos, implantados em seis espaçamentos com idades variando entre 2 e 14 anos. A base de dados foi particionada em 70% para treinamento e 30% para validação. Com base em testes iniciais e problemas anteriores foram definidos os parâmetros da regressão simbólica. Para comparação foram ajustados 5 modelos tradicionais (Curtis, Trorey, Linear simples, Stoffels e Henricksen) e RNA. Para avaliar a qualidade das estimativas, foram calculadas a correlação, quadrado médio do erro, média do erro absoluto, raiz quadrada do erro médio em porcentagem e a análise gráfica dos resíduos. Para comparação dos modelos realizou-se o teste de identidade entre os métodos. O modelo gerado pela regressão simbólica, com uma correlação de 0,7861 e um RMSE de 11,72% se mostrou mais eficiente que os demais modelos e levemente inferior aos resultados da RNA. No teste de identidade entre os métodos verificou-se que as estimativas obtidas pela regressão simbólica foram estatisticamente diferentes das estimativas obtidas por redes neurais artificias e demais modelos. Por fim concluiu-se que a regressão simbólica se mostra promissora na estimativa de altura total das árvores e assim como a RNA, estas têm uma precisão maior que os modelos tradicionais de relação hipsométrica.

Palavras-chave: Aprendizado de máquinas. Inteligência artificial. Programação genética. Redes neurais artificiais.

ABSTRACT

Height is an important variable for forest management, so it is essential to estimate it accurately. Recently, artificial intelligence techniques have been used for this purpose, even though several techniques have not yet been tested, as is the case of symbolic regression. In this context, the objective of this work is to evaluate the quality of the results produced by estimating the total height of trees using the symbolic regression method with genetic programming. The data for the study come from a clonal plantation of *Eucalyptus* spp. located in the northern region of the state of Minas Gerais. The database is composed of 57 genetic materials, planted in six spacings with ages ranging from 2 to 14 years. The database was partitioned into 70% for training and 30% for validation. Based on initial tests and previous problems the parameters of the symbolic regression were defined. For comparison 5 traditional (Curtis, Trorey, Simple Linear, Stoffels and Henricksen) and ANN models were fitted. To evaluate the quality of the estimates, correlation, mean square of the error, mean absolute error, square root of the mean error in percent, and graphical analysis of the residuals were calculated. To compare the models, the identity test between the methods was performed.

The model generated by symbolic regression, with a correlation of 0.7861, and an RMSE, of 11.72% showed to be more efficient than the other models and the same results for the ANN in the training data. In the test data, the estimates obtained by symbolic regression were statistically different from the estimates obtained by artificial neural networks and other models. Finally, it was concluded that the symbolic regression is promising in estimating the total height of trees and, as with ANN, it has a higher precision than traditional models.
Keywords: Machine learning. Artificial intelligence. Genetic programming. Artificial neural networks.

INTRODUÇÃO

A altura total das árvores é um fator importante no manejo florestal, sendo necessária para determinar o volume de madeira, a biomassa acima do solo e o estoque de carbono (ZHOU et al., 2018). Embora sua medição possa ser feita diretamente com dispositivos analógicos ou lasers, existem vários métodos para estimar a altura total da árvore (XIONG et al., 2016).

Quando estimada, a variável altura é obtida a partir do diâmetro a 1,30 m do solo (dap), estabelecendo-se a chamada relação hipsométrica. O dap pode ser considerado uma variável de fácil obtenção, porém, o mesmo não ocorre com a altura, a qual é obtida por meio de uma operação onerosa e sujeita a erros de medição (HUSCH; BEERS; KERSHAW, 2003). Sendo assim, essa relação pode reduzir os custos do inventário, já que só o diâmetro será medido para a grande maioria das árvores (SOARES; PAULA NETO; SOUZA, 2011).

Sendo assim a precisão de tais modelos é muito importante para a preparação de tabelas de volumes precisas e para o desenvolvimento de modelos de previsão de crescimento (ÖZÇELIK et al., 2013). Dessa forma nos últimos anos se buscado desenvolver formas para aumentar a precisão dessas estimativas. Como é o caso do uso de ferramentas de inteligência artificial (IA) para estimar diferentes variáveis em manejo florestal (VIEIRA et al., 2018; BAYAT et al., 2020). Redes Neurais Artificiais (RNAs) é aquela de maior destaque, sendo usada como uma alternativa às técnicas tradicionais de modelagem (VAN DAO et al., 2020).

Diante desse contexto, é possível considerar que outras ferramentas de IA também possam ser empregadas no manejo florestal, podendo-se citar a Regressão Simbólica (RS). A RS é uma técnica de *Machine Learning* (ML) que utiliza a Programação Genética (PG) para encontrar modelos matemáticos que melhor descrevem as relações entre as variáveis analisadas. Ela difere da regressão tradicional por não depender de uma estrutura de modelo previamente especificada (MINEBO; STIJVEN, 2011);

Assim, este trabalho tem como objetivo avaliar a qualidade dos resultados produzidos com a utilização da Regressão Simbólica para estimativas de altura total e compará-los com os modelos tradicionais e RNA. Espera-se dessa forma encontrar um método preciso e apropriado para prever a altura das árvores em questão.

MATERIAIS E MÉTODOS

Caracterização da área de estudo e base de dados

Os dados são provenientes de 2.109 parcelas de inventário florestal, distribuídas em 862 talhões, com plantios clonais de *Eucalyptus spp.,* localizados na região norte do estado de Minas Gerais. Segundo a classificação de Köppen (ALVARES et al., 2013), a região do estudo possui clima Aw (tropical com inverno seco) e Cw (subtropical com inverno seco). A temperatura média anual varia entre 18 e 27ºC, a precipitação média anual é igual a 1100 mm e a altitude média é de 567 m.

A base de dados é composta por 3 projetos, com 57 materiais genéticos pertencentes a 10 espécies, implantados em seis espaçamentos: 3,00m x 1,00 m; 3,00m x 2,00 m; 3,00m x 3,00 m; 4,00m x 2,00 m; 6,00m x 1,00 m; e 7,00m x 1,00 m. O plantio tinha idades variando entre 2 e 14 anos, estando as parcelas em maior proporção nas idades entre 3 e 7 anos (TABELA 1).

Tabela 1 - Análise descritiva para as variáveis diâmetro (DAP) e altura total (H).

Dados	N	DAP				H			
		Mínimo (cm)	Média (cm)	Máximo (cm)	Desvio Padrão (cm)	Mínimo (m)	Média (m)	Máximo (m)	Desvio Padrão (m)
Treino	10.321	4,17	12,78	26,48	2,56	4,00	18,96	35,30	3,60
Validação	4.451	3,98	12,74	23,49	2,60	4,20	18,88	35,20	3,60
Total	14.772	3,98	12,77	26,48	2,57	4,00	18,93	35,30	3,60

Nota: N = Número de indivíduos

Fonte: Elaborada pela autora, 2021

Para a avaliação da capacidade de generalização dos modelos, realizou-se o particionamento dos dados. Na qual foi feita a divisão da base de dados em dois conjuntos independentes, sendo um para treinamento e outro para validação dos modelos. A porcentagem utilizada para a divisão foi de 30% dos dados para validação e 70% para treinamento dos algoritmos e ajuste dos modelos de regressão. Para que a separação resultasse em dois grupos de amplitude e variação similares, a divisão foi feita dentro de cada parcela.

Regressão Simbólica

A definição dos parâmetros do algoritmo de regressão simbólica foi realizada após execuções iniciais com foco em tempo de processamento e qualidade de resultados e após revisões de problemas citados por outros autores (BARMPALEXIS et al., 2011; CHAABENI; NEDHI, 2021).

A população inicial foi gerada pelo método de *Ramped-half-and-half*. Tal método foi escolhido para que as soluções iniciais apresentassem uma boa diversidade no que diz respeito ao tamanho dos indivíduos. Cada indivíduo foi gerado considerando o seguinte conjunto de funções: soma (+); subtração (-); multiplicação (*); divisão (÷); exponenciação (exp); logaritmo (log); seno (sen); cosseno (cos); e a variável elevada ao quadrado (x^2).

O tamanho da população é o número total de soluções testadas em cada geração, sendo uma solução representada por uma equação obtida aleatoriamente (no caso da população inicial) ou pelo

cruzamento de soluções anteriores. Tal parâmetro foi definido com valor igual a 1.000 e a quantidade de gerações (ou iterações do algoritmo) igual a 200.

O tipo de seleção é o método utilizado para selecionar as soluções da população atual que serão utilizadas para cruzamento. Nesse caso, considerou-se a seleção por torneio com avaliação simultânea de cinco indivíduos ou soluções.

A taxa de *crossover*, que é a probabilidade se realizar o cruzamento de duas soluções da população de pais, foi definida como sendo igual a 0,90. Já a taxa de mutação, que é a probabilidade de se alterar um ponto específico da solução após o *crossover*, foi de 0,10.

A seleção dos indivíduos mais aptos ocorreu pela avaliação da função de *fitness*, dada pela estimativa do valor do erro quadrático médio. Para a população seguinte, foi considerado o método de elitismo, no qual os melhores indivíduos da população anterior são mantidos na população seguinte.

E a profundidade e o comprimento máximo limitam o tamanho da árvore quanto à quantidade de operações matemáticas inseridas no modelo. Considerou-se uma profundidade de 12 e comprimento de 16.

A variável DAP foi considerada como variável de entrada e altura como variável de saída. O algoritmo de RS foi executado utilizando-se o software HeuristicLab versão 3.3.16, desenvolvido por membros do *Heuristic and Evolutionary Algorithms Laboratory* (WAGNER et al, 2014). A melhor equação obtida e seus parâmetros estimados foram utilizados para estimar a altura das árvores.

Regressão Linear

Foram testados cinco modelos hipsométricos de regressão linear (aritméticos e logarítmicos) obtidos na literatura (TABELA 2) e ajustados no software R.

Tabela 2 - Modelos hipsométricos tradicionais testados para o ajuste dos dados.

Nº	Modelo	Equação
1	Curtis	$Ln(H) = \beta_0 + \beta_1 * (1/DAP)$
2	Trorey	$H = \beta_0 + \beta_1 * DAP + \beta_2 * DAP^2$
3	Linear simples	$H = \beta_0 + \beta_1 * DAP$
4	Stoffels	$Ln(H) = \beta_0 + \beta_1 * Ln(DAP)$
5	Henricksen	$H = \beta_0 + \beta_1 * Ln(DAP)$

Em que: Ln= logaritmo neperiano; H = altura total estimada da árvore (m); DAP = diâmetro a 1,30 m de altura (cm); βs = parâmetros a serem estimados.

Fonte: Elaborada pela autora, 2021

Redes Neurais Artificiais

A metodologia empregada neste trabalho para a definição de configurações de RNA apropriada para estimação da altura de árvores foi baseada na metodologia utilizada no trabalho de Martins et al. (2016).

Com o software NeuroForest (BINOTI; BINOTI; LEITE, 2013), duzentas redes neurais artificiais do tipo Multilayer Perceptron foram treinadas, sendo escolhida a de maior correlação entre os valores estimatidos e observados para o conjunto de validação.

A camada de entrada das RNAs possuía um único neurônio, sendo este equivalente ao DAP. Considerou-se apenas uma única camada oculta contendo 11 neurônios com a função de ativação do tipo sigmóide. A camada de saída foi construída com apenas um neurônio, correspondendo à variável altura total.

O treinamento foi realizado com o algoritmo Resilient Propagation. A execução foi interrompida quando o processamento alcançou 3000 épocas.

Avaliação dos modelos

Para avaliar a qualidade das estimativas, foram calculadas as estatísticas de correlação (r) (Equação 1), média do erro quadrático (MSE) (Equação 2), média do erro absoluto (MAE) (Equação 3) e raiz quadrada do erro médio em porcentagem (RMSE%) (Equação 4). Além disso, foram realizadas análises gráficas dos resíduos.

$$\mathbf{r} = \frac{\mathbf{S_{y\hat{y}}}}{\mathbf{S_y S_{\hat{y}}}} \tag{1}$$

$$\mathbf{MSE} = \sum_{\mathbf{i}}^{\mathbf{n}} \frac{(\mathbf{y}-\hat{\mathbf{y}})^{\mathbf{2}}}{\mathbf{n}} \tag{2}$$

$$\mathbf{MAE} = \sum_{\mathbf{i}}^{\mathbf{n}} \frac{|\mathbf{y}-\hat{\mathbf{y}}|}{\mathbf{n}} \tag{3}$$

$$\mathbf{RMSE\%} = \frac{\mathbf{1}}{\bar{\mathbf{y}}} \sqrt{\sum_{\mathbf{i}}^{\mathbf{n}} \frac{(\mathbf{y}-\hat{\mathbf{y}})^{\mathbf{2}}}{\mathbf{n}}} * \mathbf{100} \tag{4}$$

Em que $S_{y\hat{y}}$ é a covariância entre os valores observados e estimados; Sy é o desvio padrão de y; Sŷ é o desvio padrão de ŷ; n é o número de amostras; y é a altura observada; ŷ é a altura estimada.

Após a avaliação individual dos ajustes dos modelos, realizou-se o teste de identidade entre os métodos com os dados de validação. O procedimento foi feito através da submissão dos dados ao teste estatístico proposto por Leite e Oliveira (2002). Sendo este derivado da metodologia descrita por Graybill (1976) para realizar o teste de F.

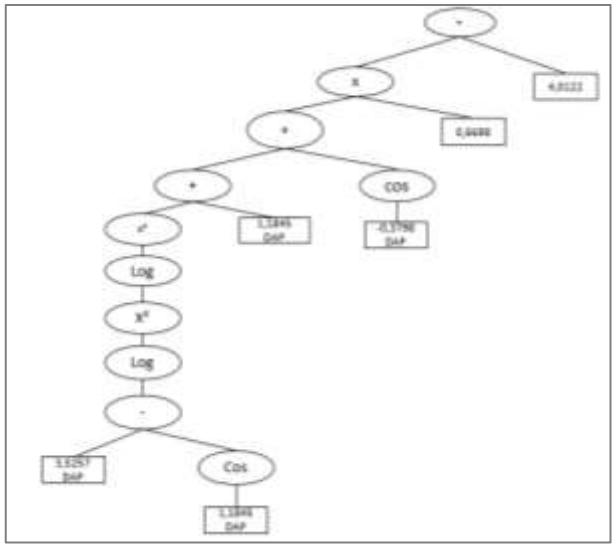
O teste considera Y_j com um método alternativo e Y_1 como um método padrão. Onde a relação entre os dois é expressa matricialmente como $Y_j = Y_1\beta + \varepsilon$. Para uma hipótese H_0: $\beta_0=0$ e $\beta_1=1$ foi aplicado o teste F, ao nível de 95% de probabilidade. Levando em consideração parâmetros como $F(H_0)$, erro médio ($t\bar{e}$) e $r_{Yj\hat{Y}1} \geq (1 - |\bar{e}|)$, foi estabelecida a identidade entre os métodos. Assim, quando $t\bar{e} > t\alpha^{(n-1)}$, a hipótese

H_0 é rejeitada. No entanto, se $t\bar{e} < t\alpha^{(n-1)}$, a hipótese H_0 é aceita, ou seja, indicando, juntamente com os outros critérios, que os métodos são estatisticamente idênticos.

RESULTADOS

A Regressão Simbólica apresenta como resultado uma árvore hierárquica de busca (FIGURA 1), a qual pode ser convertida facilmente em uma expressão matemática (equação 5).

Figura 1 - Representação da árvore hierárquica de busca



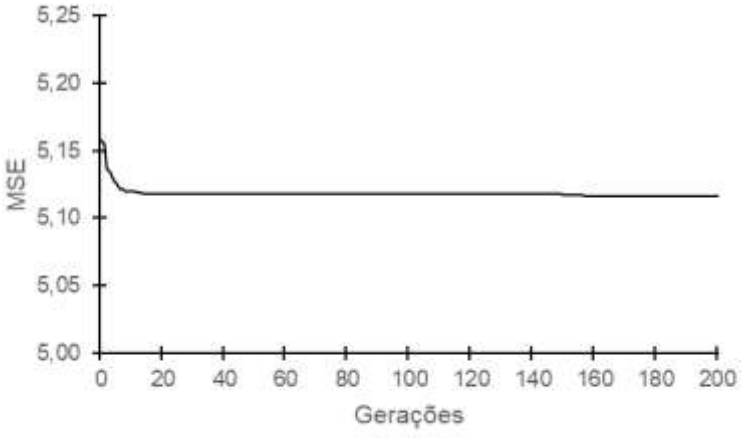
Fonte: Elaborada pela autora, 2021

$$H = \left(\left(\left(\log\left(\log\left((3{,}526 * DAP - \cos(1{,}185 * DAP)\right)\right)\right)^2\right)^2 + 1{,}185 * \ DAP\right) + \cos(-0{,}380 * DAP)) * 0{,}670 + 4{,}012\right) \quad (5)$$

Onde: H= altura total (m); C_s = parâmetros da equação; DAP= diâmetro a altura do peito (cm).

Em relação à evolução dos valores da função de *fitness* (FIGURA 2), nota-se que no início das gerações há uma queda brusca do valor, depois ele torna-se praticamente constanste, com suaves reduções ao longo das gerações.

Figura 2 - Evolução do *fitness* ao longo das gerações



Fonte: Elaborada pela autora, 2021

Os indicadores de ajuste sugerem que todos os modelos foram adequados para uso na estimativa da altura do povoamento (TABELA 3), sendo o modelo de Stoffels o que apresentou o menor erro padrão da estimativa (S_{yx}) e o maior coeficiente de determinação ajustado (R^2_{aj}).

Tabela 3 - Coeficientes de regressão e indicadores estatísticos dos modelos de hipsométricos ajustados.

Modelo	β_0	β_1	β_2	R^2_{aj}	S_{yx}
Curtis	3,5386*	-7,5081*		0,576	2,36
Trorey	2,7945*	1,4503*	-0,0140*	0,603	2,27
Linear Simples	5,0246*	1,0901*		0,601	2,27
Stoffels	1,0968*	0,7230*		0,611	2,27
Henricksen	-13,6807*	12,9191*		0,589	2,31

Nota: * = Significativo a 5% de probabilidade

Fonte: Elaborada pela autora, 2021

Excetuando-se o modelo de Curtis, o qual apresentou estatísticas inferiores aos demais, os modelos apresentaram indicadores favoráveis e similares entre si, com RMSE% inferiores a 12,5% (TABELA 4).

Tabela 4 - Valores das estatísticas para cada um dos modelos.

Modelo	Treino				Validação			
	r	MSE	MAE	RMSE%	r	MSE	MAE	RMSE%
RS	0,7780	5,12	1,74	11,93%	0,7861	4,94	1,73	11,72%
RNA	0,7780	5,11	1,74	11,93%	0,7866	4,93	1,73	11,71%
Curtis	0,7605	5,57	1,81	12,45%	0,7735	5,27	1,80	12,11%
Trorey	0,7765	5,15	1,75	11,97%	0,7851	4,96	1,73	11,75%
Linear Simples	0,7753	5,17	1,75	11,99%	0,7823	5,02	1,74	11,82%
Stoffels	0,7761	5,17	1,75	12,00%	0,7846	4,98	1,74	11,77%
Henricksen	0,7676	5,32	1,78	12,17%	0,7792	5,08	1,77	11,89%

Nota: r: correlação entre valores observados e valores estimados; MSE: média do erro quadrático; MAE: média do erro absoluto; RMSE%: raiz quadrada do erro médio em porcentagem; RS: regressão simbólica; RNA: rede neural artificial.

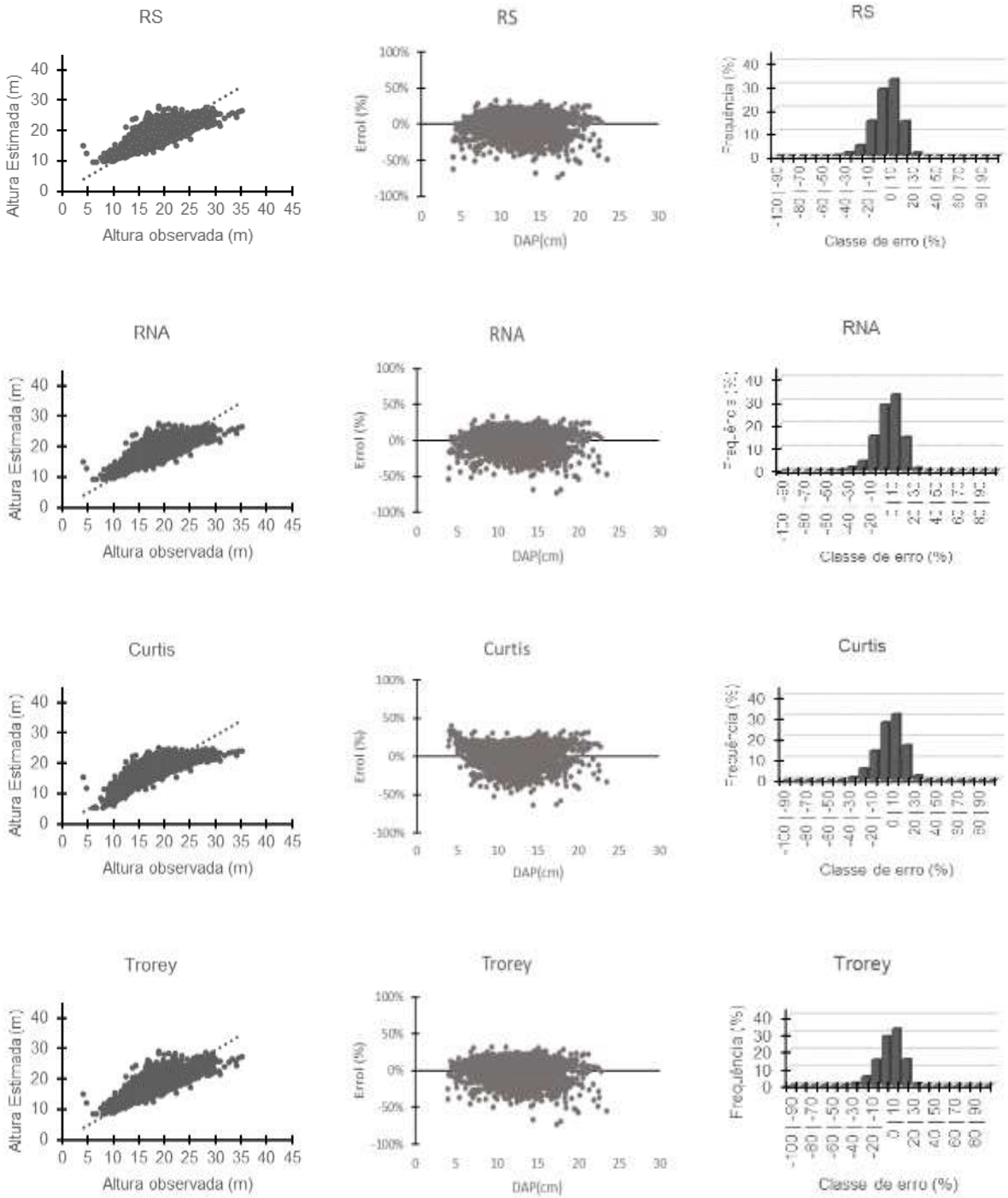
Fonte: Elaborada pela autora, 2021

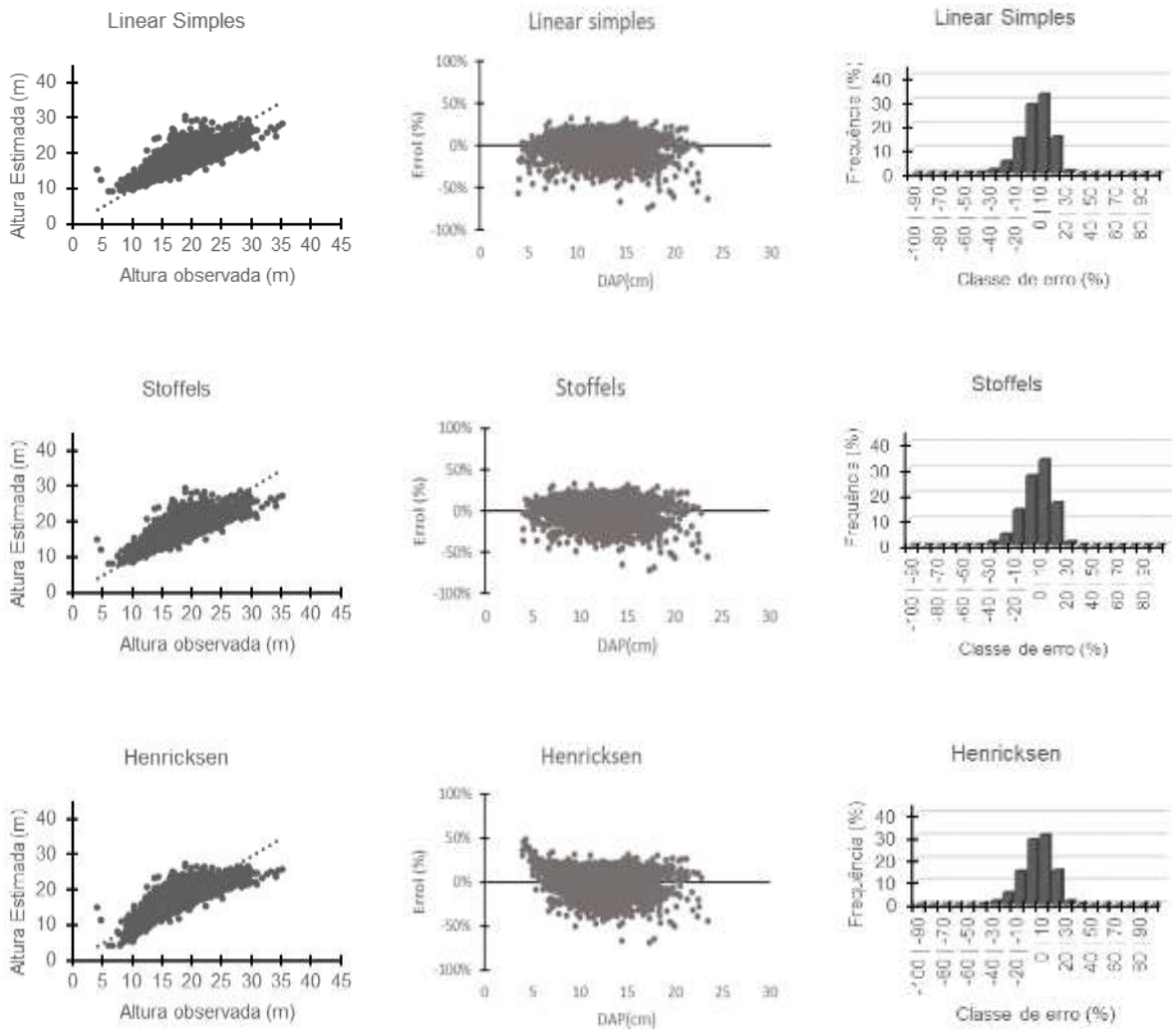
A RS e RNA apresentaram o menor RMSE% (11,93%) e a maior correlação entre as alturas estimadas e observadas (0,7780) no conjunto de treinamento. No entanto, na etapa de validação a RNA se mostrou levemente superior. A correlação foi maior e o MSE foi menor para os modelos de inteligência artificial em relação aos modelos clássicos de regressão.

Os gráficos de resíduos mostram uma tendência de subestimação para árvores mais altas, sendo menos acentuada nos modelos de AM (Figura 3). Os modelos de Curtis e Henricksen foram os menos

acurados. Ademais, todos os outros apresentaram dispersão de resíduos semelhantes. Em todos os modelos, aproximadamente 60% dos erros estão entre -10% e 10%.

Figura 3 - Alturas observada e estimada, distribuição gráfica dos resíduos e histogramas de frequência dos erros.





Fonte: Elaborada pela autora, 2021

Através do procedimento estatístico proposto por Leite; Oliveira (2002) foi rejeitada a hipótese H_0, pois verificou-se que as estimativas obtidas pela regressão simbólica foram estatisticamente diferentes das estimativas obtidas por redes neurais artificias e modelos tradicionais (TABELA 5).

Tabela 5 - Resultados de cada etapa do teste de proposto por Leite; Oliveira (2002) comparando os resultados obtidos por cada método com os resultados obtidos via regressão simbólica.

Modelo	r	$\bar{e}$	$F(H_0)$	$t\bar{e}$	$r_{YjY1} \geq (1 - \lvert \bar{e} \rvert)$	Conclusão
RNA	0,9998	-0,0001	$0,3621^{ns}$	$0,9940^{ns}$	Não	$Y_j \neq Y_1$
Curtis	0,9745	-0,0072	1146,50*	11,5355*	Não	$Y_j \neq Y_1$
Trorey	0,9979	-0,0003	$0,4515^{ns}$	$1,4639^{ns}$	Não	$Y_j \neq Y_1$
Linear simples	0,9967	0,0003	11,4644*	2,1544*	Não	$Y_j \neq Y_1$
Stoffels	0,9973	-0,0074	1124,16*	36,6953*	Não	$Y_j \neq Y_1$
Henricksen	0,9839	-0,0012	19,0208*	1,9634*	Não	$Y_j \neq Y_1$

Nota: RNA: rede neural artificial; r: correlação.

Fonte: Elaborada pela autora, 2021

DISCUSSÃO

Os resultados de ajuste do modelo mostram coeficientes de determinação abaixo de 0,7. Segundo Campos e Leite (2017) normalmente a relação altura diâmetro não é tão forte, sendo muito comum que o valor do coeficiente de determinação de equações hipsométricas seja menor que 0,80. Sousa et al. (2013) encontraram valores de R^2aj entre 0,40 à 0,43 para os mesmos modelos em estudos de relação hipsométrica para *Eucalyptus urophylla* em regime de alto fuste.

As técnicas de inteligência artificial apresentaram erros mais próximos de zero e estimativas mais precisas para a altura das árvores na área de estudo, o que está alinhado com os resultados de outras pesquisas (BAYAT et al., 2020; DIAMANTOPOULOU; OZÇELIL, 2012; VIEIRA et al., 2018).

A RNA foi levemente superior aos demais por apresentar menor valor de RMSE% e maior valor de correlação, sendo este modelo adequado para estimar a altura total do plantio. No entanto, a RS apresentou resultados semelhantes. Os modelos de aprendizado de máquina apresentaram distribuição de resíduos menos tendenciosa que os modelos de regressão, assim como no trabalho de Costa Filho et al., (2019).

O manejo florestal requer modelos precisos e capazes de prever características importantes da floresta, como diâmetros e alturas de árvores. Todos os métodos de modelagem que preveem o desempenho da floresta, como modelos de regressão e modelos de inteligência artificial, têm seus próprios pontos fortes e fracos. Embora os modelos de regressão tradicionais sejam capazes de fornecer fórmulas específicas, e isso possa facilitar a compreensão dos relacionamentos entre as variáveis nesses modelos, eles têm muitas limitações, incluindo a independência do intervalo de suposições estatísticas, como uma distribuição normal de dados, independência de variáveis e assim por diante (BAYAT et al. 2019; OU; LEI; SHEN, 2019; TAKAYAMA et al., 2019).

Uma das vantagens das técnicas de inteligência artificial na modelagem é que, em muitos casos, elas não têm as mesmas limitações dos modelos empíricos. Por exemplo, algumas pressuposições da regressão linear podem inviabilizar a utilização dos modelos clássicos. Outras vantagens das técnicas de inteligência artificial são a capacidade de trabalhar com variáveis qualitativas, além de apresentarem, geralmente, uma maior precisão (VIEIRA et al., 2018).

O teste de identidade de modelos aplicado nos dados de validação indicou que as alturas estimadas pela RS diferem estatisticamente das alturas estimadas pelos outros métodos. Lee et al. (2018) usaram três novas técnicas de aprendizado de máquina, *Support Vetor Regression* (SVR), *modified regression trees* (RT) e *Randon Forest* (RF) na Coréia do Sul para prever as alturas das áreas florestais e concluíram que esses três modelos eram capazes de estimar bem a altura média das árvores dos talhões, e as estimativas desses três modelos não foram estatisticamente diferentes.

Finalmente, Bourque, Bayat e Zhang (2019) usaram a regressão simbólica para examinar o controle de variáveis abióticas sobre o crescimento da altura e do diâmetro em uma floresta não manejada no norte do Irã. Nesse estudo, eles provaram a capacidade das abordagens de IA para previsão. Apesar das diferenças nos tipos de métodos e nas técnicas de inteligência artificial usadas em todos esses estudos, ou nos números e tipos de entradas de modelos, todos eles mostram a superioridade dos métodos de inteligência artificial sobre modelos de regressão. Sendo assim, as técnicas de IA podem substituir os modelos empíricos.

Embora o uso de equações hipsométricas seja consolidado, novas técnicas tornaram-se promissoras devido ao seu êxito na determinação da altura. Com os resultados encontrados no presente trabalho, comprovou-se a aplicabilidade da regressão simbólica. Desta forma, incentiva-se o seu uso pela precisão e ineditismo e a realização de novos trabalhos por ser uma técnica ainda carente de estudos na modelagem da altura das árvores.

CONCLUSÃO

A regressão simbólica mostrou resultados promissores na estimativa de altura total das árvores. Assim como a RNA, estas têm uma precisão maior do que os modelos tradicionais na estimativa da altura total das árvores.

A regressão simbólica apresentou resultados superiores aos dos modelos tradicionais, mas levemente inferiores aos da RNA.

A regressão simbólica e a RNA possuem maior flexibilidade, logo são mais capazes de modelar interfaces complexas e não lineares.

REFERÊNCIAS

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. M.; SPAROVEK, G.; Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728. 2013.

BARMPALEXIS, P; KACHRIMANIS, K.; TSAKONAS, A.; GEORGARAKIS, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. **Information Sciences**, v. 107, n. 1, p. 75-82. 2011.

BAYAT, M.; BETTINGER, P.; HEIDARI, S.; KHALYANI, A. H.; JOURGHOLAMI, M.; HADIMI, S. K. Estimation of Tree Heights in an Uneven-Aged, Mixed Forest in Northern Iran Using Artificial Intelligence and Empirical Models. **Forests**, v. 11, n. 324, 2020.

BAYAT, M.; GHORBANPOUR, M.; ZARE, R.; JAAFARI, A.; PHAM, B.T. Application of artificial neural networks for predicting tree survival and mortality in the Hyrcanian forest of Iran. **Computers and Electronics in Agriculture**, v.164, 2019.

BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G. **NeuroForest Star.** Depositante: INPI - Instituto Nacional da Propriedade Industrial. BR nº 13410-5. Depósito: 30 abr 2013.

BOURQUE, C. P. A.; BAYAT, M.; ZHANG, C. An assessment of height–diameter growth variation in an unmanaged Fagus orientalis-dominated forest. **European Journal of Forest Research**, v.138, p. 607–621, 2019.

CAMPOS, J. C. C.; LEITE, H. G. **Mensuração florestal**: perguntas e respostas. 5.ed. Viçosa: UFV, 2017. 636p.

CHAABENI, W. B.; NEDHI, M. L. Genetic programming based symbolic regression for shear capacity prediction of SFRC beams. **Construction and Building Materials,** v. 280, p. 1-14, 2021.

COSTA FILHO, S. V. S.; ARCE, J. E.; MONTAÑO, R. N. R.; PELISSARI, A. L. 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, 2019.

DIAMANTOPOULOU, M.J.; OZCELIK, R. Evaluation of different modeling approaches for total tree-height estimation in Mediterranean Region of Turkey. **Forest Systems**, v.21, n.3, p. 383-397, 2012.

GHOSH, S. M.; BEHERA, M. D.; PARAMANIK, S. Canopy height estimation using sentinel series images through machine learning models in a Mangrove Forest. **Remote sensing**, v.12. n.9, 2020.

GRAYBILL, F.A. **Theory and application of the linear model**. Belmonte: Wadsworth Publishing Company, 1976. 704 p.

HUSCH, B.; BEERS, T. W.; KERSHAW, J. A. 2003. Forest Mensuration, 4th Ed. John Wiley & Sons, Inc. Hoboken, New Jersey. 443 p.

LEE, J.; IM, J.; KIM, K.-M.; QUACKENBUSH, L.J. Machine Learning Approaches for Estimating Forest Stand Height Using Plot-Based Observations and Airborne LiDAR Data. **Forests**, v. 9, n.5, p.268-284, 2018.

LEITE, H.G.; OLIVEIRA, F.H.T. Statistical Procedure to Teste Identity Between Analytical Methods. **Communications in Soil Science and Plant Analysis,** v.33, n. 7/8, p.1105-1118, 2002.

MARTINS; E. R.; BINOTI, M. L. M.; LEITE, H. G.; BINOTI, D. H. B. DUTRA, G. C. Configuração de redes neurais artificiais para estimação da altura total de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias (Agrária)**, v.11, n.2, p.117-123, 2016.

MINEBO, W.; STIJVEN, S**. Empowering Knowledge Computing with Variable Selection**. 2011. 135 f. Dissertação (Mestrado) – Universidade de Antwerpen. 2011.

OU, Q.; LEI, X.; SHEN, C. Individual Tree Diameter Growth Models of Larch–Spruce–Fir Mixed Forests Based on Machine Learning Algorithms. **Forests**, v.10, n.2, p.187-207, 2019.

ÖZÇELIK, R.; DIAMANTOPOULOU, M.J.; CRECENTE-CAMPO, F.; ELER, U. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. **Forest Ecology Management**, v. 306, p. 52–60, 2013.

SOARES, C. P. B.; PAULA NETO, F.; SOUZA, A. L. **Dendrometria e inventário florestal**. Viçosa: Editora UFV, 2011. 272p.

SOUSA, G. T. O.; AZEVEDO, G. B.; BARRETO, P. A. B.; CONÇEIÇÃO JÚNIOR, V. Relações hipsométricas para *Eucalyptus urophylla* conduzidos sob regime de alto fuste e talhadia no Sudoeste da Bahia. **Scientia Plena**, v. 9, n. 4., p 1-7, 2013.

TAKAYAMA, T.; UMEZAWA, T.; KOMURO, N.; OSAWA, N. A regression model-based method for indoor positioning with compound location fingerprints. Geo-Spat. **Information Science**, v. 22, p.107–113, 2019.

VAN DAO, D.; JAAFARI, A.; BAYAT, M.; MAFI-GHOLAMI, D.; QI, C.; MOAYEDI, H.; LUU, C. A spatially explicit deep earning neural network model for the prediction of landslide susceptibility. **Catena,** v. 188, 2020.

VIEIRA, G. C.; MENDONÇA, A. R.; SILVA, G. F.; ZANETTI, S. S.; SILVA, M. M.; SANTOS, A. R. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment,** v. 619, p. 1473–1481, 2018.

WAGNER, S. et al. Architecture and Design of the HeuristicLab Optimization Environment. In **Advanced Methods and Applications in Computational Intelligence**, Topics in Intelligent Engineering and Informatics Series, Springer, pp. 197-261. 2014.

XIONG, B.M.; WANG, Z.X.; LI, Z.Q.; ZHANG, E.; TIAN, K.; LI, T.T.; LI, Z.; SONG, C.L. Study on the correlation among age, DBH and tree height of the Pseudotsuga sinensis in Qizimei Mountain Nature Reserve. **Forest Resources Management**., v. 4, p. 41–46, 2016.

ZHOU, R.; WU, D.; FANG, L.; XU, A.; LOU, X. A Levenberg–Marquardt Backpropagation Neural Network for Predicting Forest Growing Stock Based on the Least-Squares Equation Fitting Parameters. **Forests**, v.9, n. 12, p.757-773, 2018.

4.2 Artigo 2-Diferentes abordagens para modelagem de altura com regressão simbólica

RESUMO

A maioria dos estudos de aprendizado de máquina para modelagem de altura é limitado às redes neurais artificiais, diante disso vê-se a necessidade de estudar outras técnicas ainda pouco utilizadas para este fim, como a regressão simbólica. Diante disso o objetivo deste trabalho é estimar a altura total das árvores em plantio de eucalipto a partir de diferentes estratégias, usando o método de regressão simbólica. A área de estudo está localizada na região norte do estado de Minas Gerais, compreendendo uma área de plantio clonal de *Eucalytus* ssp. A base de dados foi composta por dados dendrométricos provenientes de inventário florestal, localizado na região norte do estado de Minas Gerais, sendo composta por 57 materiais genéticos, implantados em seis espaçamentos com idades variando entre 2 e 14 anos. A base de dados foi particionada em 70% para treinamento e 30% para validação. Para alcançar as melhores estimativas dos dados foram testadas 5 diferentes estratégias como variáveis de entrada, sendo elas E1: Dap; E2: Dap e Idade; E3: Dap, projeto, espécie e espaçamento; E4: Dap, projeto, clone e espaçamento; E5: Dap, idade, projeto e clone. Para avaliar a qualidade das estimativas, foram calculadas a correlação entre a variável observada e estimada, média do erro absoluto, raiz quadrada do erro médio em porcentagem e análise gráfica de resíduos. Para selecionar a melhor estratégia de ajuste foi realizado o teste de Scott & Knott a 95% de probabilidade. A estratégia 5 com média do erro absoluto de 1,44 apresentou os melhores valores para todas as estatísticas apresentadas tanto no treinamento como na validação, assim como foi diferente estatisticamente dos demais pelo teste de Scott & Knott. A adição da variável idade concomitantemente com as variáveis projeto e clone se mostrou como o melhor método para estimativa de altura.

Palavras- chave: Aprendizado de máquina. Programação genética. Manejo florestal.

ABSTRACT

Most studies of machine learning for height modeling are limited to artificial neural networks, thus the need to study other techniques still little used for this purpose is seen, such as symbolic regression. Therefore, the objective of this study is to estimate the total height of trees in eucalyptus plantations from different strategies, using the symbolic regression method. The study area is located in the northern region of Minas Gerais state, comprising an area of clonal plantation of *Eucalytus* ssp. The database was composed of dendrometric data from a forest inventory, located in the northern region of Minas Gerais state, consisting of 57 genetic materials, planted in six spacings with ages ranging from 2 to 14 years. The database was partitioned into 70% for training and 30% for validation. To achieve the best data estimates 5 different strategies were tested as input variables, being them E1: DBH; E2: DBH and Age; E3: DBH, project, species and spacing; E4: DBH, project, clone and spacing; E5: DBH, age, project and clone. To evaluate the quality of the estimates, the correlation between the observed and estimated variable, mean absolute error, square root of the mean error in percent, and graphical analysis of residuals were calculated. To select the best fitting strategy the Scott & Knott test was performed at 95% probability. Strategy 5, with a mean absolute

error of 1.44, showed the best values for all the statistics presented both in training and validation, and was statistically different from the others by the Scott & Knott test. The addition of the age variable concomitantly with the project and clone variables proved to be the best method for height estimation.
Key-words: Machine learning. Genetic programming. Forest management.

INTRODUÇÃO

A cultura do gênero *Eucalyptus* detém a maioria (77%) da área de florestas plantadas no Brasil, com 6,97 milhões de hectares. A indústria de florestas plantadas é responsável por 1,2% do PIB Nacional, com receita bruta total de 97,4 bilhões de reais e uma área total em 2019 com 9 milhões de hectares (IBÁ, 2020).

No manejo florestal, gerar equações e métodos confiáveis para estimar a altura do povoamento é uma tarefa muito importante. Sua medição ou estimativa é amplamente utilizada para calcular o volume, incremento em altura e em alguns casos, pode indicar a qualidade produtiva de um local (SILVA et al, 2012).

A medição da altura das árvores é considerada uma parte importante no custo do inventário florestal (BINOTI; BINOTI; LEITE, 2013). Logo, na prática é comum que haja a aferição da altura de apenas algumas árvores nas parcelas e a estimação das alturas das demais árvores em função de variáveis de rápida medição como o DAP (BINOTI et al., 2017).

A relação existente entre o diâmetro e a altura das árvores é denominada de relação hipsométrica, esta depende de fatores ambientais e de características do povoamento, como: capacidade produtiva, idade, genótipo e espaçamento do plantio (CURTIS, 1967).

Atualmente o interesse na busca por métodos mais eficientes de estimativa da altura das árvores tem aumentado, visto que encontrar relações de altura-diâmetro precisas permitem a economia de tempo e capital nos inventários florestais. Portanto, as técnicas de aprendizagem de máquina (AM) têm sido empregadas com sucesso na área de mensuração florestal (COSTA FILHO et al., 2019, BINOTI; BINOTI; LEITE 2013; MARTINS et al., 2016; VENDRUSCOLO et al., 2016).

Sabe-se que a maioria dos estudos com AM para modelagem de alturas de árvores são restritos a RNA. Logo outras técnicas como regressão simbólica (RS) são pouco estudadas nesse contexto. A RS é um método de encontrar a melhor equação de ajuste a partir da aplicação de operações genéticas, com base na teoria da seleção natural de Darwin em que indivíduos que se adaptam melhor ao seu ambiente têm uma maior possibilidade de sobreviver e de passar as suas características genéticas para os seus descendentes, gerando automaticamente soluções que naturalmente resolvem melhor o problema investigado (SPINOZA; POZO, 2003). A aplicação deste método tem apresentado êxito em áreas correlatas como hidráulica (KUMAR et al., 2014) sensoriamento remoto (GHOSH; BEHERA; PARAMANIK, 2020), meio ambiente (KODALI et al., 2018) e gestão florestal (BOURQUE; BAYAT; ZHANG, 2019).

Sendo assim, visando aumentar a precisão das estimativas de altura o objetivo deste artigo é estimar a altura total das árvores em plantios de eucalipto a partir de diferentes estratégias, usando o método de regressão simbólica.

MATERIAIS E MÉTODOS

Localização

A área de estudo está localizada na região norte do estado de Minas Gerais, compreendendo uma área de plantio clonal de *Eucalytus ssp*. A precipitação média anual na região é de 1100 mm. A temperatura média é de 22,5 ºC, com pequena amplitude. Segundo a classificação de Köppen (ALVARES et al., 2013) o clima da região é classificado em dois tipos:

Aw: tropical com inverno seco. Apresenta estação chuvosa no verão, de novembro a abril, e nítida estação seca no inverno, de maio a outubro (julho é o mês mais seco). A temperatura média do mês mais frio é superior a 18ºC. As precipitações são superiores a 750 mm anuais, atingindo 1800 mm.

Cwa: subtropical com inverno seco (com temperaturas inferiores a 18ºC) e verão quente (com temperaturas superiores a 22ºC).

Base de dados

A base de dados foi composta por dados dendrométricos provenientes de inventário florestal com 2.109 parcelas de inventário florestal, distribuídas em 862 talhões. Sendo composta por 3 projetos, com 57 materiais genéticos pertencentes a 10 espécies, implantados em seis espaçamentos: 3,00 x 1,00 m; 3,00 x 2,00 m; 3,00 x 3,00 m; 4,00 x 2,00 m; 6,00 x 1,00 m; e 7,00 x 1,00 m. O plantio tinha idades variando entre 2 e 14 anos, estando as parcelas em maior proporção nas idades entre 3 e 7 anos (TABELA 1).

Para a avaliação da capacidade de generalização dos modelos, realizou-se o particionamento dos dados. Na qual foi feita a divisão da base de dados em dois conjuntos independentes, sendo um para treinamento e outro para validação dos modelos. A porcentagem utilizada para a divisão foi de 30% dos dados para validação e 70% para treinamento dos algoritmos e ajuste dos modelos de regressão. Para que a separação resultasse em dois grupos de amplitude e variação similares, a divisão foi feita dentro de cada parcela.

Tabela 2 - Estatísticas descritivas das variáveis diâmetro (DAP) e altura total (H) do povoamento.

Dados	N	DAP				H			
		Mínimo (cm)	Média (cm)	Máximo (cm)	Desvio Padrão (cm)	Mínimo (m)	Média (m)	Máximo (m)	Desvio Padrão (m)
Treino	10.321	4,17	12,78	26,48	2,56	4,00	18,96	35,30	3,60
Validação	4.451	3,98	12,74	23,49	2,60	4,20	18,88	35,20	3,60
Todos	14.772	3,98	12,77	26,48	2,57	4,00	18,93	35,30	3,60

Nota: N: número de indivíduos; Min= mínimo; Med=média; Max= máximo e D.P=desvio padrão

Fonte: Elaborada pelo autor, 2021

Regressão simbólica

A definição dos parâmetros do algoritmo de regressão simbólica foi realizada após execuções iniciais com foco em tempo de processamento e qualidade de resultados e após revisões de problemas citados por outros autores (BARMPALEXIS et al., 2011; CHAABENI; NEDHI, 2021).

A população inicial foi gerada pelo método de *Ramped-half-and-half*. Tal método foi escolhido para que as soluções iniciais apresentassem uma boa diversidade no que diz respeito ao tamanho dos indivíduos. Cada indivíduo foi gerado considerando o seguinte conjunto de funções: soma (+); subtração (-); multiplicação (*); divisão (÷); exponenciação (exp); logaritmo (log); seno (sen); cosseno (cos); e a variável elevada ao quadrado (x^2).

O tamanho da população é o número total de soluções testadas em cada geração, sendo uma solução representada por uma equação obtida aleatoriamente (no caso da população inicial) ou pelo cruzamento de soluções anteriores. Tal parâmetro foi definido com valor igual a 1.000 e a quantidade de gerações (ou iterações do algoritmo) igual a 200.

O tipo de seleção é o método utilizado para selecionar as soluções da população atual que serão utilizadas para cruzamento. Nesse caso, considerou-se a seleção por torneio com avaliação simultânea de cinco indivíduos ou soluções.

A taxa de crossover, que é a probabilidade se realizar o cruzamento de duas soluções da população de pais, foi definida como sendo igual a 0,90. Já a taxa de mutação, que é a probabilidade de se alterar um ponto específico da solução após o *crossover*, foi de 0,10.

A seleção dos indivíduos mais aptos ocorreu pela avaliação da função de *fitness*, dada pela estimativa do valor do erro quadrático médio. Para a população seguinte, foi considerado o método de elitismo, no qual os melhores indivíduos da população anterior são mantidos na população seguinte.

E a profundidade e o comprimento máximo limitam o tamanho da árvore quanto à quantidade de operações matemáticas inseridas no modelo. Considerou-se uma profundidade de 12 e comprimento de 16.

Para o estudo se utilizou o software HeuristicLab versão 3.3.16, desenvolvido por membros do *Heuristic and Evolutionary Algorithms Laboratory* (WAGNER et al, 2014). A RS utiliza os conceitos da PG para o processamento dos dados e criação de equações que melhor descrevem as relações entre as variáveis dependente e independentes. Para alcançar as melhores estimativas dos dados foram testadas 5 diferentes abordagens como variáveis de entrada, sendo elas descritas a seguir:

- Estratégia 1: Dap;
- Estratégia 2: Dap e Idade;
- Estratégia 3: Dap e variáveis categóricas {projeto, espécie e espaçamento};
- Estratégia 4: Dap e variáveis categóricas {projeto, clone e espaçamento};
- Estratégia 5: Dap, idade e variáveis categóricas {projeto e clone};

Avaliação

Para avaliar a qualidade das estimativas, foram calculados a correlação entre a variável observada e estimada (r) (Equação 1), média do erro absoluto (MAE) (Equação 2), Raiz quadrada do erro médio em porcentagem (Equação 3) e análise gráfica de resíduos (Equação 4).

$$\mathbf{r} = \frac{\mathbf{Cov}(\mathbf{y}, \hat{\mathbf{y}})}{\mathbf{S_y S_{\hat{y}}}} \quad (1)$$

$$\mathbf{MAE} = \sum_{\mathbf{i}}^{\mathbf{n}} \frac{|\mathbf{y} - \hat{\mathbf{y}}|}{\mathbf{n}} \quad (2)$$

$$\mathbf{RMSE\%} = \frac{\mathbf{100}}{\bar{\mathbf{y}}} * \sqrt{\sum_{\mathbf{i}}^{\mathbf{n}} \frac{(\mathbf{y} - \hat{\mathbf{y}})^{\mathbf{2}}}{\mathbf{n}}} \quad (3)$$

$$\mathbf{Erro\%} = \frac{(\mathbf{y} - \hat{\mathbf{y}})}{\mathbf{y}} * \mathbf{100} \quad (4)$$

Onde: Sy= desvio padrão de y; Sŷ= desvio padrão de ŷ; n= número de amostras; ẏ= média da altura observada; y=altura observada; ŷ=altura estimada.

Para seleção adequada da melhor estratégia de ajuste, foi realizada uma análise de variância de um critério onde as 5 estratégias foram comparadas, cada uma contendo 14772 repetições (sendo cada árvore uma unidade amostral), essas consistiam no valor do erro absoluto em metros de cada árvore presente na base total dos dados, fazendo, dessa forma, uma validação preditiva. Sendo então os valores mais próximos de zero ideais, visto menor variação da altura estimada em relação à altura real da árvore. Havendo diferença entre as estratégias, aplicou-se o teste de Scott & Knott (1974) ao nível de 95% de probabilidade, gerado com os 14772 valores de erro absoluto em metros de todos os dados por estratégia, para identificação da melhor abordagem de ajuste.

O teste de comparações múltiplas de Scott-Knott foi escolhido por ser considerado robusto e por não apresentar resultados ambíguos, de acordo com Borges e Ferreira (2003). Ambos os testes foram realizados no software R (R CORE TEAM, 2020).

RESULTADOS

Os valores das estatísticas de precisão para seleção da melhor metodologia adotada são apresentados na Tabela 2. A E5 apresentou os melhores valores para todas as estatísticas apresentadas tanto no treinamento como na validação.

Tabela 2 - Estatísticas de precisão para os dados de treino e validação dos tratamentos.

Abordagem	Treino			Validação		
	r	MAE	RMSE%	r	MAE	RMSE%
E1	0,7780	1,74	11,93	0,7861	1,73	11,72
E2	0,8065	1,61	11,23	0,8134	1,60	11,08
E3	0,8009	1,67	11,37	0,8111	1,65	11,14
E4	0,8318	1,55	10,54	0,8418	1,52	10,28
E5	0,8480	1,45	10,07	0,8565	1,43	9,83

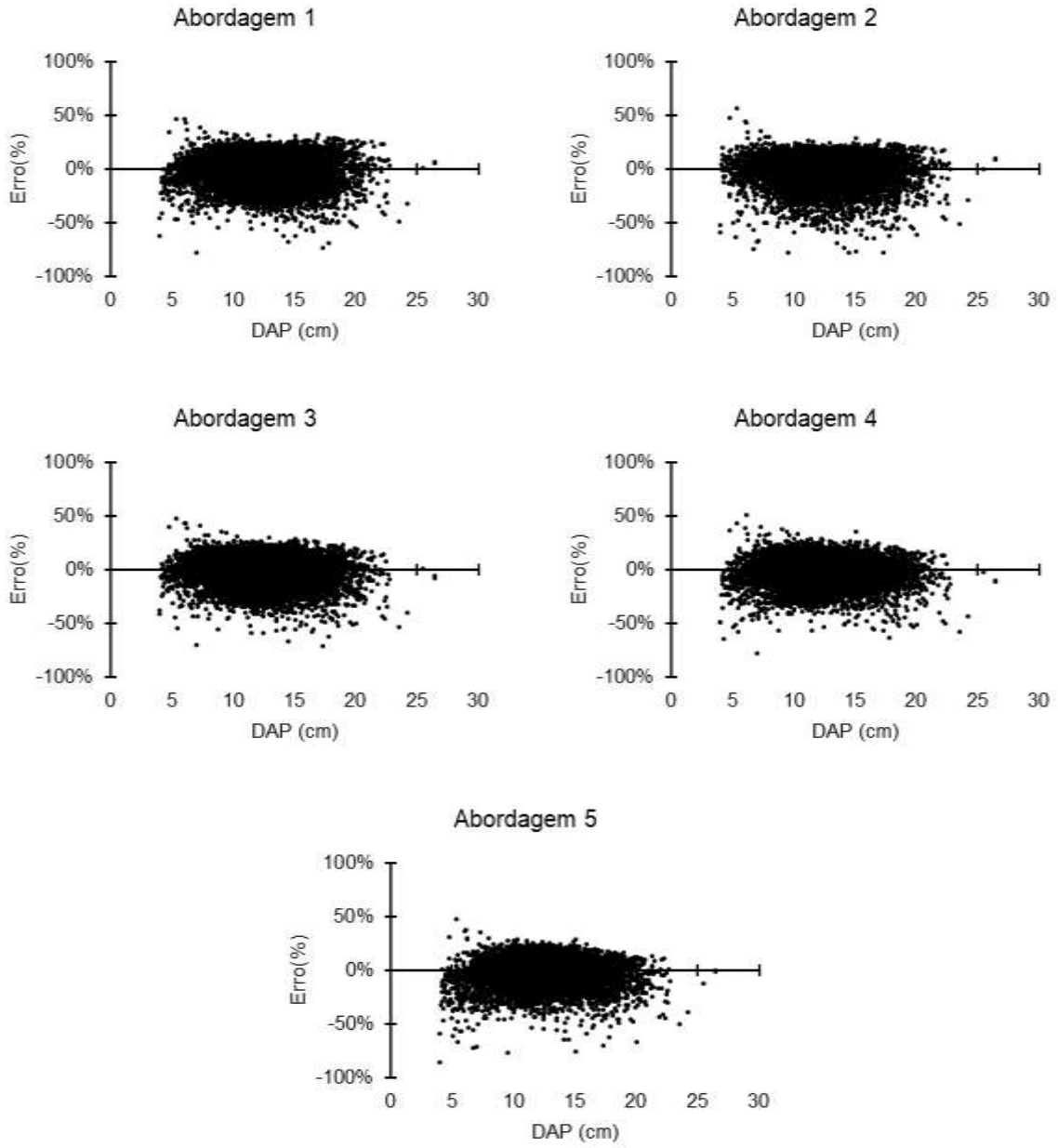
Nota: r: correlação entre valores observados e valores estimados; MAE: média do erro absoluto; RMSE%: raiz quadrada do erro médio em porcentagem.

Fonte: Elaborada pela autora, 2021

Para avaliação da qualidade dos ajustes, além do cálculo da raiz quadrada do erro médio e da correlação, utilizou-se também a análise gráfica de resíduos. Foram verificadas diferenças entres as abordagens, quando realizada a análise visual de gráficos (FIGURA 1).

De um modo geral, os gráficos não apresentaram grandes tendenciosidades para as abordagens adotadas. É importante ressaltar que os gráficos da Figura 1 foram elaborados com os resíduos de todas as árvores da base de dados, e não apenas com os dados de ajuste.

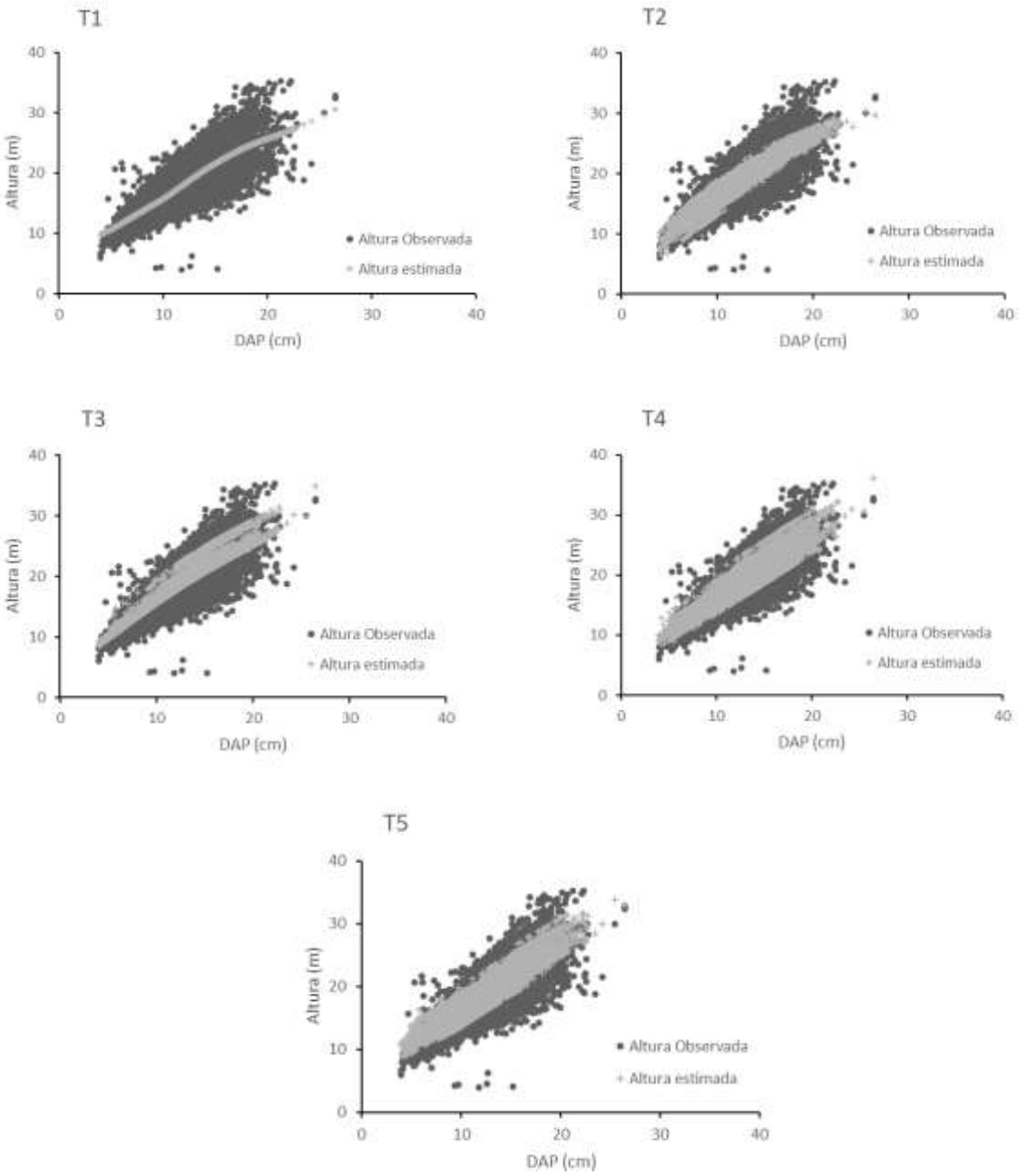
Figura 1 - Gráficos de resíduos para cada uma das abordagens avaliadas.



Fonte: Elaborada pela autora, 2021

As curvas das alturas observadas e estimadas em função do DAP mostram o comportamento das estimativas obtidas pelas estratégias (FIGURA 2). Nota-se que a E1 apresenta apenas uma linha de ajuste, enquanto os demais apresentam várias linhas em consequência da adição de mais variáveis.

Figura 2 - Curvas hipsométricas obtidas para os diferentes tratamentos.



Fonte: Elaborada pela autora, 2021

A seleção decisória da melhor estratégia de ajuste foi tomada com base na ANOVA. Os valores da análise de variância são apresentados na Tabela 3.

Tabela 3 - Análise de variância para diferença entre altura real e estimada pelos tratamentos.

FV	GL	SQ	QM	Fc	P-value
Estratégias	4	767,0026	191,7507	108,812	0,000
Erro	73855	13149,0237	1,7622		
Total corrigido	73859	130916,0264			

Fonte: Elaborada pela autora, 2021

Pela Tabela 4 foi detectado que houve diferença entre os tratamentos, implicando que pelo menos uma abordagem diferiu das demais. Para comparação das abordagens foi realizado o teste de média de Scott-Knott. Os resultados são apresentados na Tabela 4. A melhor estratégia adotada foi a inclusão das variáveis categóricas (clone e projeto) juntamente com o Dap e a idade e a pior estratégia de ajuste foi o ajuste somente com o Dap confirmando os resultados apresentados anteriormente.

Tabela 4 - Teste de média Scott-Knott a 95% de probabilidade dos valores médios de erro absoluto.

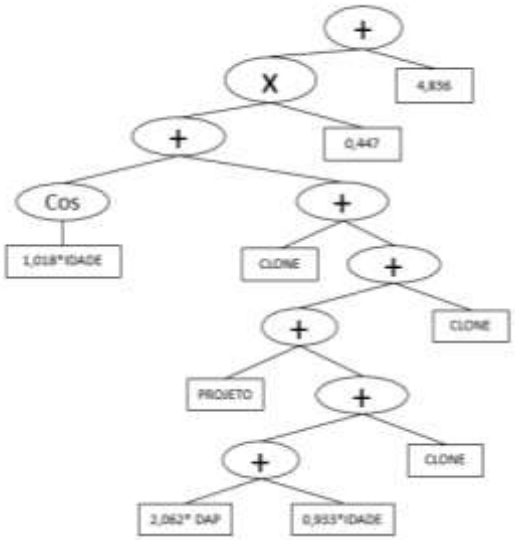
Estratégia	Média
E5	1,444145 a
E4	1,537445 b
E2	1,604993 c
E3	1,663541 d
E1	1,740268 e

Nota: letras diferentes apresentam diferença estatística a 95% de probabilidade

Fonte: Elaborada pela autora, 2021.

Como a regressão simbólica gera resultados passíveis de serem interpretados. A árvore de expressão da estratégia 5 (FIGURA 3) juntamente com a sua interpretação matemática são mostradas logo adiante.

Figura 3 - Árvore de expressão da E5



Fonte: Elaborada pela autora, 2021

$$h = \left(\left(\cos(c_0 . \text{Idade}) + \left(c_1 + \left(\left(c_2 + \left((c_3 . \text{DAP} + c_4 . \text{Idade}) + c_5\right)\right) + c_6\right)\right)\right) . c_7 + c_8\right)$$

DISCUSSÃO

Os valores de correlação de todas as estratégias foram acima de 0,75, mostrando o potencial destes para a estimativa da altura. Avaliando o RMSE% todos foram inferiores a 12%. A inclusão de outras

variáveis no modelo aumentou a precisão das estimativas, em relação à estratégia1 em 6%, 5%, 13% e 16% para as abordagens 2, 3, 4 e 5, respectivamente. A adição de mais variáveis aos ajustes pode trazer benefícios na redução dos custos e maior precisão das estimativas de altura.

A estratégia que utilizou a idade em conjunto com as variáveis categóricas (A5) foi a melhor opção para ajuste dos dados, levando em consideração as estatísticas encontradas. A possibilidade de inserção de variáveis categóricas (não numéricas) no ajuste geraram resultados precisos, e sem tendenciosidade.

Vendruscolo et al. (2015) inseriram uma variável categórica para modelagem de altura com RNA e observaram resultados mais precisos. Binoti et al. (2014) também apontaram vantagens com a inclusão de variáveis categóricas na modelagem volumétrica. Já SOUZA et al, (2018) não observaram diferenças perceptíveis na modelagem de volume com a inclusão da variável categórica (clone), isso pode ter se dado devido ao fato de que elas necessitam de representatividade para todos os níveis das da amostra para ser ter ganho em precisão, o que nem sempre é possível (MARTINS et al., 2016).

As estratégias de ajuste usadas para estimar as alturas diferiram entre si, segundo a análise de variância aplicada às alturas estimadas. Sendo assim pelo teste de Scott-Knott a estratégia 5 é o método recomendado ao conjunto de dados, considerando que não houve estratificação dos dados é notório o potencial da utilização da RS para estimativa da altura de árvores com inserção de variáveis categóricas. Além disso, do ponto de vista operacional, o método permite reduzir muito o tempo gasto com ajustes e avaliações de modelos de regressão principalmente em situações com muitos estratos.

CONCLUSÃO

A adição de variáveis categóricas melhora a precisão das estimativas de altura total com a técnica de regressão simbólica.

A adição da variável idade concomitantemente com as variáveis categóricas (projeto e clone) se mostrou como o melhor método para a modelagem de altura total com regressão simbólica dos dados em questão.

REFERÊNCIAS

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. M.; SPAROVEK, G.; Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728. 2013.

BARMPALEXIS, P; KACHRIMANIS, K.; TSAKONAS, A.; GEORGARAKIS, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. **Information Sciences**, v. 107, n. 1, p. 75-82. 2011.

BINOTI, D. H. B et al. Estimation of height of eucalyptus trees with Neuroevolution of augmenting topologies (NEAT). **Revista Árvore**, v. 41, n.3, p. e410314, 2017.

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

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.

BORGES, L. C.; FERREIRA, D. F. Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não normais dos resíduos. **Revista de Matemática e Estatística**, v. 21, n. 1, p. 67-83, 2003.

BOURQUE, C. P. A.; BAYAT, M.; ZHANG, C. An assessment of height–diameter growth variation in an unmanaged Fagus orientalis-dominated forest. **European Journal of Forest Research**, v.138, p. 607–621, 2019.

CHAABENI, W. B.; NEDHI, M. L. Genetic programming based symbolic regression for shear capacity prediction of SFRC beams. **Construction and Building Materials,** v. 280, p. 1-14, 2021.

COSTA FILHO, S. V. S.; ARCE, J. E.; MONTAÑO, R. N. R.; PELISSARI, A. L. 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, 2019.

CURTIS, R. O. Height diamenter and height diamenter age equations for second growth Douglas-fir. **Forest Science**, Washington, v. 13, n. 4, p. 356-375, 1967.

GHOSH, S. M.; BEHERA, M. D.; PARAMANIK, S. Canopy height estimation using sentinel series images through machine learning models in a Mangrove Forest. **Remote sensing**, v.12. n.9, 2020.

IBÁ - Indústria brasileira de árvores. **Anuário IBÁ 2020**: ano base 2019. Brasília: 2020. 122 p.

KODALI, A.; SZUBERT, M.; DAS, K.; GANGULY, S.; BONGARD, J. Understanding climate-vegetation interactions in global rainforests through a GP-tree analysis. **Parallel Problem Solving from Nature (PPSN),** p. 525-536, 2018.

KUMAR, B.; JHA, A.; DESHPANDE, V. SREENIASULU, G. Regression model for sediment transport problems using multi-gene symbolic genetic programming. **Computers and Electronics in Agriculture.** v. 103, p. 82–90, 2014.

MARTINS; E. R.; BINOTI, M. L. M.; LEITE, H. G.; BINOTI, D. H. B. DUTRA, G. C. Configuração de redes neurais artificiais para estimação da altura total de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias (Agrária)**, v.11, n.2, p.117-123, 2016

R CORE TEAM. **R: A language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing, 2020.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, v. 30, p. 507-512, 1974.

SILVA, G. F.; CURTO, R. de A.; SOARES, C. P. B.; PIASSI, L. de C. Avaliação de métodos de medição de altura em florestas naturais. **Revista Árvore**, Viçosa, v. 36, n. 2, p. 341-348, 2012.

SOUZA, S. R. R.; SILVA, J. A. A.; GUERA, O. G. M.; FERREIRA, T. A.E. Redes neurais para estimativa volumétrica de clones de *Eucalyptus* spp. no pólo gesseiro do Araripe. **Revista Brasileira de Biometria**, v.36, n.3, p.715-729, 2018.

SPINOZA, E.; POZO, A. Controlling the population size in genetic programming, in: E. CANTU-PAZ, J.A. FOSTER, K. DEB, L.D. DAVIS, R. ROY, U.M. O'REILLY, H.G. BEYER, R. STANDISH, G. KENDALL, S. WILSON, M. HARMAN, J. WEGENER, D. DASGUPTA, M.A. POTTER, A.C. SCHULTZ, K.A. DOWSLAND, N. JONOSKA, J. MILLER (Eds.), **Genetic and Evolutionary Computation-GECCO**, Springer, Berlin, Heidelberg, 2003, p. 1975–1985.

VENDRUSCOLO, D. G. S.; DRESCHER, R.; CARVALHO, S. P. C.; MEDEIROS, R. A.; SOUZA, H. S.; CERQUEIRA, C. L.; MOURA, J. P. V.; LEITE, H.G. Height prediction of Tectona grandis trees by mixed effects modelling and artificial neural networks**. International Journal of Current Research**, v. 8, n. 12, p.43189-43195, 2016.

VENDRUSCOLO, D. G. S.; DRESCHER, R.; SOUZA, H. S.; MOURA, J. P. V. M.; MAMORÉ, F. M. D.; SIQUEIRA, T. A. S. Estimativa da altura de eucalipto por meio de regressão não linear e redes neurais artificiais. **Revista Brasileira de Biometria**, v.33, n.4, p.556-569, 2015.

WAGNER, S. et al. Architecture and Design of the HeuristicLab Optimization Environment. In **Advanced Methods and Applications in Computational Intelligence**, Topics in Intelligent Engineering and Informatics Series, Springer, pp. 197-261. 2014.

4.3 Artigo 3-Modelagem hipsométrica via regressão simbólica com variáveis qualitativas

RESUMO

O objetivo deste trabalho é avaliar a qualidade dos resultados produzidos estimando-se a altura total das árvores usando o método de regressão simbólica via programação genética. Os dados para realização do estudo são provenientes de um plantio clonal de *Eucalyptus* spp. localizado na região norte do estado de Minas Gerais. A base de dados é composta por 57 materiais genéticos, implantados em seis espaçamentos com idades variando entre 2 e 14 anos. Os dados foram separados em 70% para treinamento e 30% para validação. Os parâmetros da regressão simbólica foram definidos com base em uma experimentação inicial. Para comparação foram ajustados 4 modelos tradicionais (Curtis, Trorey, Linha reta e Stoffels) e RNA. A regressão simbólica com uma correlação (r) de 0,8338, um erro médio absoluto de 1,53 m e RMSE de 9,96% se mostrou mais eficiente que os demais modelos tradicionais. Os dados de teste mantiveram o mesmo padrão de treinamento. A RNA com r de 0,8559 e RMSE de 9,85% apresentou resultados superiores aos demais métodos, mostrando que os modelos de inteligência artificial têm o potencial para complementar e substituir modelos empíricos na modelagem florestal. Por fim concluiu-se que a regressão simbólica, assim como a RNA possuem uma maior flexibilidade tendo uma precisão maior do que os modelos tradicionais na estimativa da altura total das árvores.

Palavras-chave: Modelos empíricos. Programação genética. Redes Neurais Artificiais.

ABSTRACT

The objective of this study is to evaluate the quality of the results produced by estimating the total height of trees using the symbolic regression method via genetic programming. The data for the study come from a clonal plantation of *Eucalyptus* spp. located in the northern region of the state of Minas Gerais. The database is composed of 57 genetic materials, planted in six spacings with ages ranging from 2 to 14 years. The data were separated into 70% for training and 30% for validation. The parameters of the symbolic regression were defined based on an initial experimentation. For comparison 4 traditional models (Curtis, Trorey, Straight Line and Stoffels) and ANN were fitted. The symbolic regression with a correlation (r) of 0.8338% a mean absolute error of 1.53 m and RMSE of 9.96% proved to be more efficient than the other traditional models. The test data maintained the same training pattern. The ANN with r of 0.8559 and RMSE of 9.85% presented better results than the other methods, showing that artificial intelligence models have the potential to complement and replace empirical models in forest modeling. Finally, it was concluded that symbolic regression, as well as ANN, has a greater flexibility and a higher accuracy than traditional models in estimating the total height of trees.

Keywords: Empirical models. Genetic programming. Artificial Neural Networks

INTRODUÇÃO

A medição da altura em um povoamento florestal é fundamental para estimar o volume e qualificar a produtividade de um determinado local, sendo o volume fundamental na fase de planejamento da produção de qualquer empresa florestal (RIBEIRO et al., 2010). Por ser uma prática difícil e onerosa, na prática, costuma-se medir o diâmetro de todas as árvores da parcela e a altura de algumas delas e, por

meio dos pares altura-diâmetro mensurados, estima-se as alturas das demais árvores da parcela (BINOTI et al., 2017). Essa relação conhecida como relação hipsométrica, busca estimar as alturas das árvores por meio da relação do diâmetro da árvore a 1,30m do solo e da altura total (VIANNA et al., 2016). Essa prática resulta na redução nos custos do inventário florestal e em uma operacionalização mais eficaz (RIBEIRO et al., 2010).

Fatores como idade e diâmetro são as principais entradas de modelos empíricos de regressão para estimar a altura das árvores (XIONG et al., 2016). Porém, quando a relação hipsométrica não possui uma correlação muito forte, devido à alta variabilidade das alturas para uma mesma classe de diâmetro, a inclusão de variáveis qualitativas nos modelos pode ser de grande importância, possibilitando a obtenção de estimativas que se aproximem ao máximo da realidade (MARTINS et al., 2016). Nesse sentido, as técnicas de aprendizagem de máquina (AM) têm sido empregadas com sucesso na área de mensuração florestal.

É notável que a maior parte dos estudos que abordam AM na modelagem da altura das árvores são limitados a aplicação de RNA. Portanto, outras técnicas de inteligência artificial, ainda são pouco ou nada estudadas nesse contexto, como é o caso da Regressão simbólica (RS), método baseado em computação evolutiva que visa encontrar soluções ótimas dentro de um conjunto infinito de expressões matemáticas por meio do cruzamento dos indivíduos. Sua aplicação nas ciências florestais tem demonstrado êxito no sensoriamento remoto (GHOSH; BEHERA; PARAMANIK, 2020), na gestão florestal (BOURQUE; BAYAT; ZHANG, 2019) e na hidráulica (KUMAR et al., 2014).

Considerando o potencial dessa técnica na melhoria das estimativas da altura das árvores, o presente trabalho teve como objetivo avaliar o desempenho na predição de altura por meio da utilização da Regressão Simbólica com foco na determinação de modelos hipsométricos que apresentem melhores resultados que os modelos tradicionalmente utilizados para essa finalidade.

MATERIAIS E MÉTODOS

Caracterização da área de estudo e base de dados

Os dados são provenientes de 2.109 parcelas de inventário florestal, distribuídas em 862 talhões, com plantios clonais de *Eucalyptus spp.,* localizados na região norte do estado de Minas Gerais. Segundo a classificação de Köppen (ALVARES et al., 2013), a região do estudo possui clima Aw (tropical com inverno seco) e Cw (subtropical com inverno seco). A temperatura média anual varia entre 18 e 27ºC, a precipitação média anual é igual a 1100 mm e a altitude média é de 567 m.

A base de dados é composta por 3 projetos, com 57 materiais genéticos pertencentes a 10 espécies, implantados em seis espaçamentos: 3,00 x 1,00 m; 3,00 x 2,00 m; 3,00 x 3,00 m; 4,00 x 2,00 m; 6,00 x 1,00 m; e 7,00 x 1,00 m. O plantio tinha idades variando entre 2 e 14 anos, estando as parcelas em maior proporção nas idades entre 3 e 7 anos.

Para a avaliação da capacidade de generalização dos modelos, realizou-se o particionamento dos dados. Na qual foi feita a divisão da base de dados em dois conjuntos independentes, sendo um para treinamento e outro para validação dos modelos. A porcentagem utilizada para a divisão foi de 30% dos dados para validação e 70% para treinamento dos algoritmos e ajuste dos modelos de regressão. Para

que a separação resultasse em dois grupos de amplitude e variação similares, a divisão foi feita dentro de cada parcela.

Regressão Simbólica

A definição dos parâmetros do algoritmo de regressão simbólica foi realizada após execuções iniciais com foco em tempo de processamento e qualidade de resultados e após revisões de problemas citados por outros autores (BARMPALEXIS et al., 2011; CHAABENI; NEDHI, 2021).

A população inicial foi escolhida por um método em que as soluções iniciais apresentassem uma boa diversidade no que diz respeito ao tamanho dos indivíduos. Cada indivíduo foi gerado considerando o seguinte conjunto de funções: soma (+); subtração (-); multiplicação (*); divisão (÷); exponenciação (exp); logaritmo (log); seno (sen); cosseno (cos); tangente (tan); e tangente hiperbólica (tanh). O tamanho da população é o número total de soluções testadas em cada geração, sendo uma solução representada por uma equação obtida aleatoriamente (no caso da população inicial) ou pelo cruzamento de soluções anteriores, a quantidade de gerações são as iterações do algoritmo. O tipo de seleção é o método utilizado para selecionar as soluções da população atual que serão utilizadas para cruzamento. A taxa de *crossover* é a probabilidade se realizar o cruzamento de duas soluções da população de pais, a taxa de mutação é a probabilidade de se alterar um ponto específico da solução após o *crossover*. Para seleção dos indivíduos mais aptos usa-se a função de *fitness*. E a profundidade e o comprimento máximo limitam o tamanho da árvore quanto à quantidade de operações matemáticas inseridas no modelo. Os valores de todos estes parâmetros se encontram na Tabela 1.

Tabela 1 - Parâmetros da regressão simbólica

Parâmetro	Valor
Tamanho da população	100
Gerações	10000
Método de inicialização	*Probabilistic*
Taxa de *crossover*	0,90
Taxa de mutação	0,15
Função *fitness*	Erro médio quadrático
Conjunto de funções	+, -, x, ÷ , exp, log, sen, cos, tan, tanh
Elitismo	1
Tipo de seleção	Classificação geral
Profundidade	8
Comprimento	20

Fonte: Elaborada pela autora, 2021

As variáveis consideradas como entradas foram Idade e DAP, consideradas como variáveis quantitativas, e Projeto, Espécie e Espaçamento, consideradas como variáveis qualitativas. A variável de saída foi a altura total.

O algoritmo de RS foi executado utilizando-se o software HeuristicLab versão 3.3.16, desenvolvido por membros do *Heuristic and Evolutionary Algorithms Laboratory* (WAGNER et al, 2014). A melhor equação obtida e seus parâmetros estimados foram utilizados para estimar a altura das árvores.

Regressão Linear

Quatro modelos hipsométricos de regressão linear foram testados, optando-se por aqueles que mais tem destaque na literatura consultada (SOARES et al., 2004). São eles o modelo de Curtis (Equação 1), o modelo de Trorey (Equação 2), o modelo da Linha Reta (Equação 3) e o modelo de Stoffels (Equação 4).

$$\mathbf{Ln(H) = \beta_0 + \beta_1 * (1/DAP)} \quad (1)$$

$$\mathbf{H = \beta_0 + \beta_1 * DAP + \beta_2 * DAP^2} \quad (2)$$

$$\mathbf{H = \beta_0 + \beta_1 * DAP} \quad (3)$$

$$\mathbf{Ln(H) = \beta_0 + \beta_1 * Ln(DAP)} \quad (4)$$

Em que: H = altura total da árvore (m); DAP = diâmetro a 1,30 m de da árvore altura (cm); βs = parâmetros a serem estimados.

Redes Neurais Artificiais

A metodologia empregada neste trabalho para a definição de configurações de RNA apropriada para estimação da altura de árvores foi baseada na metodologia utilizada no trabalho de Martins et al. (2016).

Com o software NeuroForest (BINOTI; BINOTI; LEITE, 2013), duzentas RNAs do tipo *Multilayer Perceptron* foram treinadas. A camada de entrada das RNAs possuía 22 neurônios, sendo um neurônio para cada variável quantitativa Idade e DAP e um para cada classe das variáveis categóricas Projeto, Espécie e Espaçamento. Considerou-se apenas uma única camada oculta contendo 11 neurônios com a função de ativação do tipo sigmóide. A camada de saída foi construída com apenas um neurônio, correspondendo à variável a altura total.

O treinamento foi realizado com o algoritmo *Resilient Propagation*. A execução foi interrompida quando o processamento alcançou 3000 ciclos ou um erro quadrático médio inferior a 0,0001.

Avaliação dos modelos

Para avaliar a qualidade das estimativas, foram calculadas a correlação (r) (Equação 5), quadrado médio do resíduo (MSE) (Equação 6), média do erro absoluto (MAE), (Equação 7), e raiz quadrada do erro médio (RMSE) (Equação 8) e análise gráfica dos resíduos.

$$\mathbf{r = \frac{S_{y\hat{y}}}{S_y S_{\hat{y}}}} \quad (5)$$

$$\mathbf{MSE = \sum_{i}^{n} \frac{(y - \hat{y})^2}{n}} \quad (6)$$

$$\mathbf{MAE} = \sum_{\mathbf{i}}^{\mathbf{n}} \frac{|\mathbf{y} - \hat{\mathbf{y}}|}{\mathbf{n}} \quad (7)$$

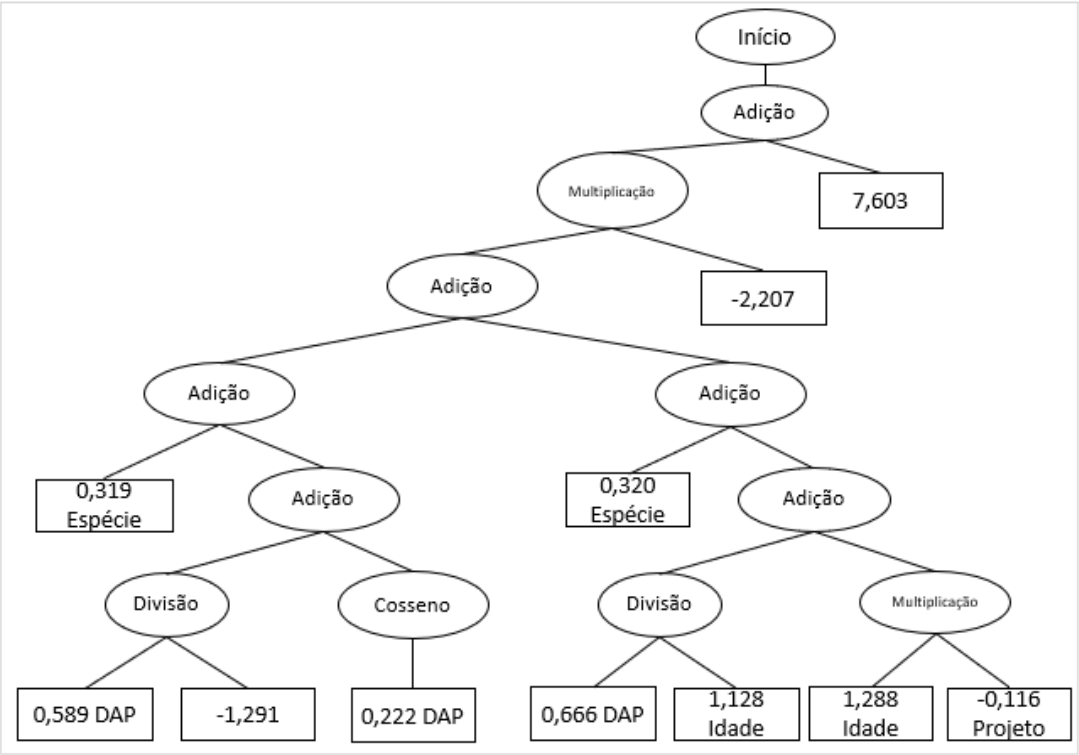
$$\mathbf{RMSE\%} = \frac{\mathbf{1}}{\bar{\mathbf{y}}} \sqrt{\sum_{\mathbf{i}}^{\mathbf{n}} \frac{(\mathbf{y}-\hat{\mathbf{y}})^{\mathbf{2}}}{\mathbf{n}}} * \mathbf{100} \quad (8)$$

Em que: Syŷ= Covariância; SySŷ= desvio padrão; n= número de amostras; y=altura observada; ŷ=altura estimada.

RESULTADOS

A Regressão Simbólica apresenta como resultado uma árvore hierárquica de busca (FIGURA 1) que pode ser convertida facilmente em uma expressão matemática (Equação 9).

Figura 4 - Representação da árvore hierárquica de busca



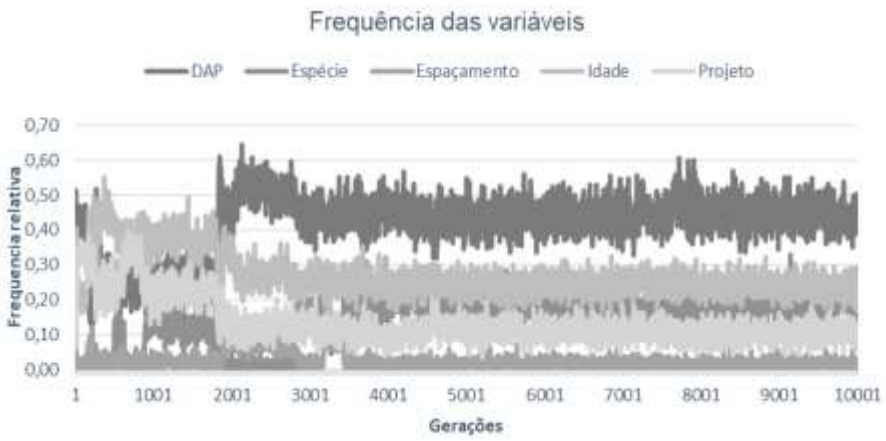
Fonte: Elaborada pela autora, 2021

$$H = \left(\left(\left(C_0.(E = S03) + \left(C_1.\frac{DAP}{C_2} + \cos\left(C_3.DAP\right)\right)\right) + \left(C_4.(E = S04) + \left(C_5.\frac{DAP}{C_6}.Id + C_7.Id.C_8.(P = P03)\right)\right)\right).C_9 + C_{10}\right) \quad (9)$$

Onde: H= altura total (m); C_s = parâmetros da equação; E= espécie; Id= Idade (anos); DAP= diâmetro a altura do peito (cm); P=projeto.

A Figura 2 mostra a frequência relativa das variáveis com o passar das gerações. Observa-se a partir da geração 2000 há uma relativa constância das variáveis.

Figura 2 - Frequência das variáveis em cada geração



Fonte: Elaborada pela autora, 2021

Os resultados mostraram que as variáveis de maior impacto para o modelo de regressão simbólica foram DAP, idade, projeto e espécie, respectivamente (TABELA 2). As estimativas de cada parâmetro se encontram na Tabela 3.

Tabela 2 - Impacto relativo de cada variável

Variável	Relevância no modelo
Diâmetro	0,612
Idade	0,167
Projeto	0,019
Espécie	0,007
Espaçamento	0,000

Fonte: Elaborada pela autora, 2021

Tabela 3 - Parâmetros do modelo

Parâmetros	**Estimativa**
C_0	0,319
C_1	0,589
C_2	-1,291
C_3	0,222
C_4	0,320
C_5	0,666
C_6	1,128
C_7	1,288
C_8	-0,116
C_9	-2,207
10	7,603

Fonte: Elaborada pela autora, 2021

Os dados mostram que a correlação das alturas reais com as estimadas variou de 76% a 86%. Isso reflete o desenvolvimento de modelos moderadamente bons, já a raiz quadrada do erro médio tem uma pequena variação (10% a 12%). A regressão simbólica com uma correlação de 0,8254 um erro médio absoluto de 1,55 m e um RMSE de 10,72% para os dados de treino se mostrou mais eficiente que os demais modelos, porém, menos eficiente que a RNA, que apresentou uma correlação de 0,8516. Os dados de validação mantiveram o mesmo padrão de treinamento (TABELA 4).

Como pode ser visto, a correlação é claramente maior quando se usa os modelos de inteligência artificial do que os modelos empíricos, ainda maior com a RNA e o MSE é menor do que o observado em todos modelos empíricos.

Enquanto o desempenho do conjunto de treinamento fornece uma ideia da capacidade de generalização do modelo, os conjuntos de validação formam um conjunto independente de dados para determinar a previsibilidade (KUMAR et al., 2014). Indicando assim, que a regressão simbólica pode ser usada de forma eficaz para modelagem de altura total.

Tabela 4 - Valores das estatísticas para cada um dos modelos.

Modelo	Treino				Validação			
	r	MSE	MAE	RMSE%	r	MSE	MAE	RMSE%
RS	0,8254	4,13	1,55	10,72%	0,8338	3,94	1,53	10,51%
RNA	0,8516	3,56	1,42	9,96%	0,8559	3,46	1,42	9,85%
Curtis	0,7605	5,57	1,81	12,45%	0,7735	5,27	1,80	12,11%
Trorey	0,7765	5,15	1,75	11,97%	0,7851	4,96	1,73	11,75%
Linear Simples	0,7753	5,17	1,75	11,99%	0,7823	5,02	1,74	11,82%
Stoffels	0,7761	5,17	1,75	12,00%	0,7846	4,98	1,74	11,77%

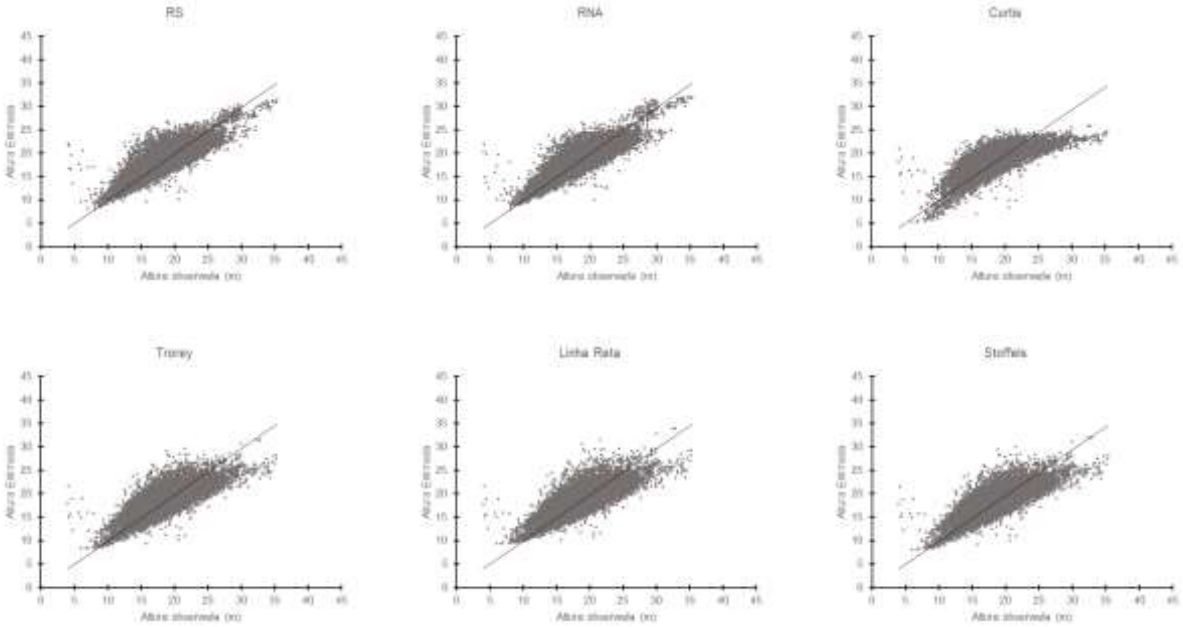
Nota: r: correlação entre valores observados e valores estimados; MSE: média do erro quadrático; MAE: média do erro absoluto; RMSE%: raiz quadrada do erro médio em porcentagem; RS: regressão simbólica; RNA: rede neural artificial.

Fonte: Elaborada pela autora, 2021

As distribuições das alturas das árvores observadas *versus* as estimadas sugerem que o modelo de regressão simbólica e RNA estimam mais razoavelmente as alturas totais das árvores (FIGURA 3).

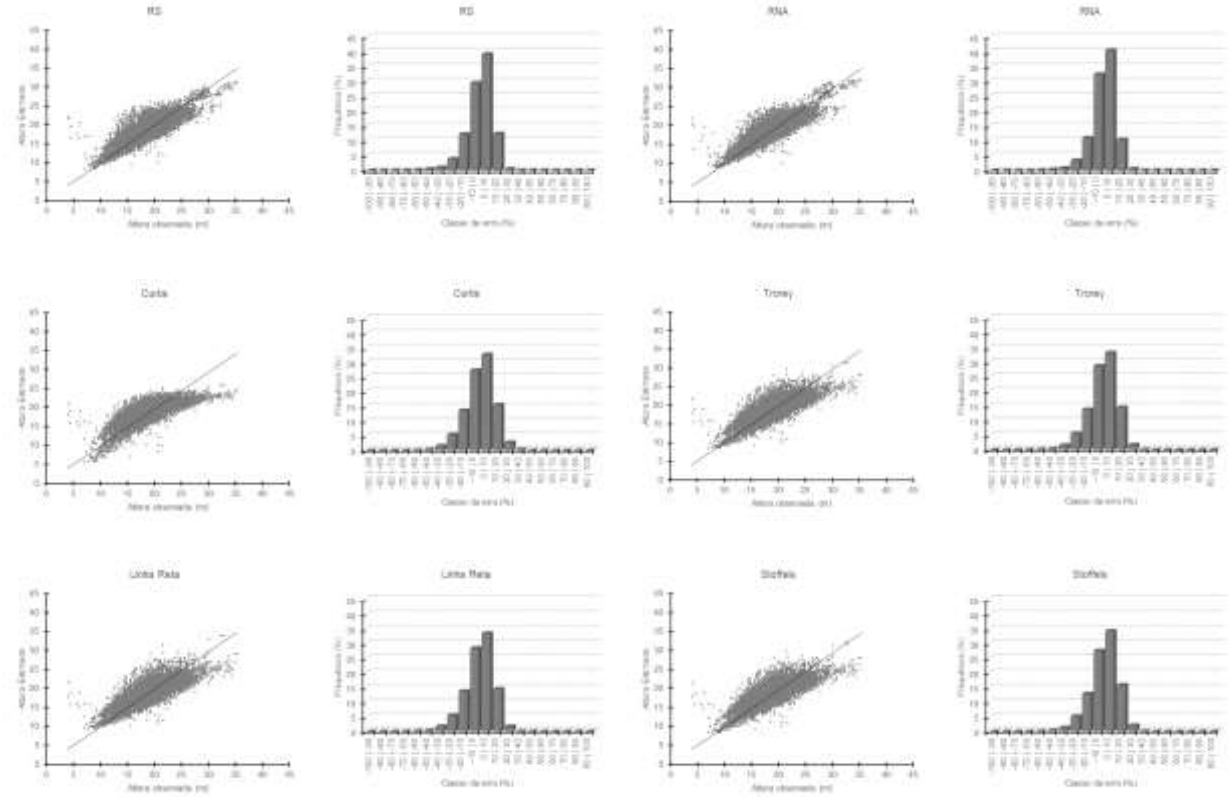
Nas Figura 4 e 5, são mostradas as correlações entre os valores de altura observados e previstos e a distribuição do erro por classes para os dados de treinamento e validação, respectivamente. A distribuição entre as alturas totais estimadas e observadas das árvores é razoável. Nos modelos tradicionais observa-se uma tendência de subestimativa das árvores mais altas o que não ocorre nos modelos de ML. Na regressão simbólica assim como na RNA observa-se que os erros se encontram majoritariamente nas duas classes centrais (-10% a 10%), enquanto nos modelos tradicionais os erros de dividem por mais classes.

Figura 3 - Altura estimada *versus* altura observada e distribuição de resíduos para os dados.



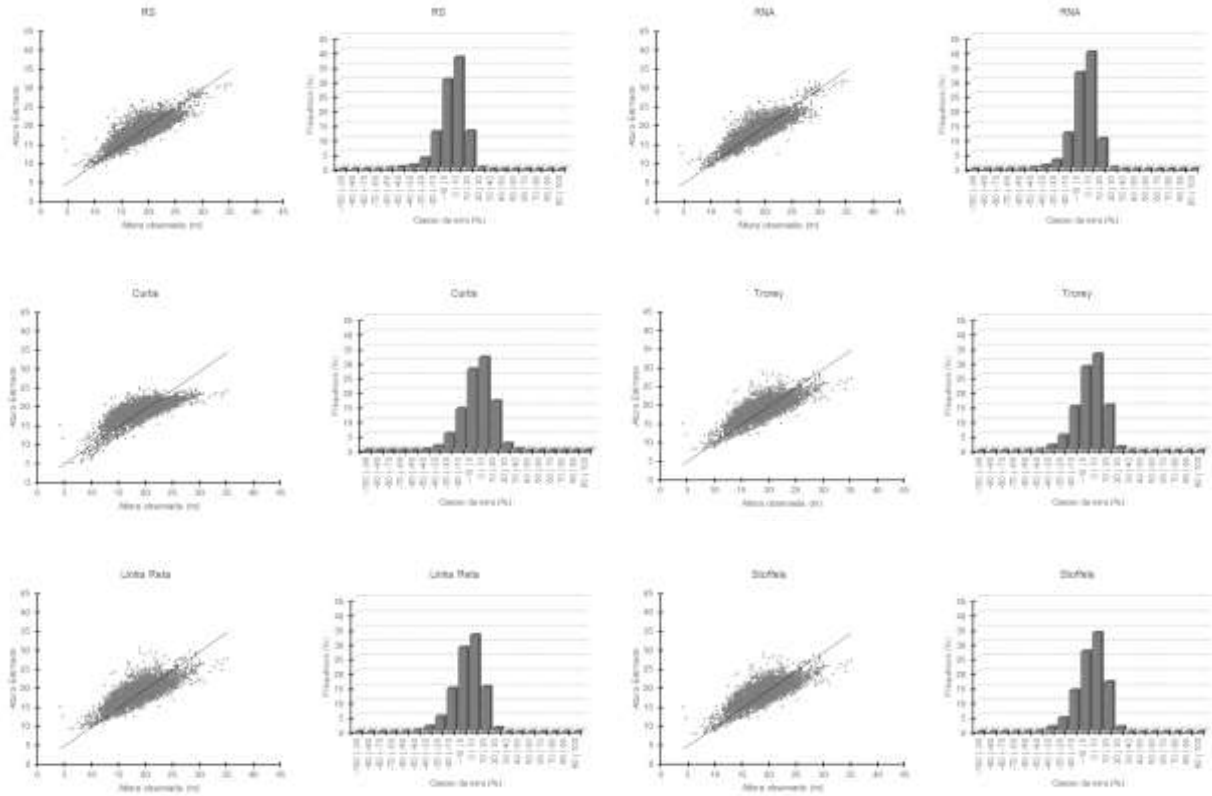
Fonte: Elabora pela autora, 2021

Figura 4 - Altura estimada versus altura observada e distribuição do erro para os dados de treinamento



Fonte: Elabora pela autora, 2021

Figura 5 - Altura estimada *versus* altura observada e distribuição do erro para os dados de validação.



Fonte: Elabora pela autora, 2021

DISCUSSÃO

O manejo florestal requer modelos precisos e capazes de prever características importantes da floresta, como diâmetros e alturas de árvores. Todos os métodos de modelagem que preveem o desempenho da floresta, como modelos de regressão e modelos de inteligência artificial, têm seus próprios pontos fortes e fracos. Embora os modelos de regressão tradicionais sejam capazes de fornecer fórmulas específicas, e isso possa facilitar a compreensão dos relacionamentos entre as variáveis nesses modelos, eles têm muitas limitações, incluindo a independência do intervalo de suposições estatísticas, como uma distribuição normal de dados, independência de variáveis e assim por diante (BAYAT et al. 2019; OU et al., 2019; TAKAYAMA et al., 2019).

A precisão e a capacidade de modelar relações complexas entre variáveis são as principais características de um modelo bom e adequado. Qualquer tipo de modelo de diâmetro-altura pode nem sempre ser adequado para todos os tipos de condições onde uma espécie de árvore em particular pode ser encontrada, porque as condições do local podem afetar a relação diâmetro-altura (BAYAT et al. 2020).

Uma das vantagens das técnicas de inteligência artificial na modelagem é que, em muitos casos, elas não têm as mesmas limitações dos modelos empíricos. Por exemplo, algumas suposições (normalidade dos dados e outras) podem afetar a qualidade dos modelos empíricos. Outras vantagens das técnicas de inteligência artificial são a capacidade de trabalhar com variáveis qualitativas e a relativa exatidão e precisão desses modelos (VIEIRA et al., 2018).

De acordo com os resultados deste estudo (Tabela 2), as técnicas de inteligência artificial apresentaram erros mais baixos e estimativas mais precisas da altura das árvores na área de estudo, o que está alinhado com os resultados de outros estudos (BAIAT et al., 2020; DIAMANTOPOULOU; OZÇELIL, 2012; VIEIRA et al., 2018).

Lee et al. (2018) usaram três novas técnicas de aprendizado de máquina, incluindo *Support Vetor Regression* (SVR), *modified regression trees* (RT) e *Randon Forest* (RF) na Coréia do Sul para prever as alturas das áreas florestais e concluíram que esses três modelos eram capazes de estimar bem a altura dos talhões, e as estimativas desses três modelos não foram estatisticamente significantes.

Finalmente, Bourque et al. (2019) na floresta de Kheyrud usaram a programação genética para determinar a relação entre altura e diâmetro na floresta de faias e selecionar as variáveis mais ambientais. Nesses estudos, eles provaram a capacidade das abordagens de IA para previsão. Apesar das diferenças nos tipos de métodos e nas técnicas de inteligência artificial usadas em todos esses estudos, ou nos números e tipos de entradas de modelos, todos eles mostram a superioridade dos métodos de inteligência artificial sobre modelos de regressão. Sendo assim, as técnicas de IA podem substituir os modelos empíricos.

CONCLUSÃO

Modelos de inteligência artificial têm o potencial de complementar e substituir modelos empíricos na modelagem florestal.

A regressão simbólica, assim como a RNA têm uma precisão maior do que os modelos tradicionais na estimativa da altura total das árvores.

A regressão simbólica e a RNA possuem maior flexibilidade, logo são mais capazes de modelar interfaces complexas e não lineares.

REFERÊNCIAS

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. M.; SPAROVEK, G.; Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728. 2013.

BARMPALEXIS, P; KACHRIMANIS, K.; TSAKONAS, A.; GEORGARAKIS, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. **Information Sciences**, v. 107, n. 1, p. 75-82. 2011.

BAYAT, M.; GHORBANPOUR, M.; ZARE, R.; JAAFARI, A.; PHAM, B.T. Application of artificial neural networks for predicting tree survival and mortality in the Hyrcanian forest of Iran. **Computers and Electronics in Agriculture**, v.164, 2019.

BAYAT, M.; BETTINGER, P.; HEIDARI, S.; KHALYANI, A. H.; JOURGHOLAMI, M.; HADIMI, S. K. Estimation of Tree Heights in an Uneven-Aged, Mixed Forest in Northern Iran Using Artificial Intelligence and Empirical Models. **Forests**, v. 11, n. 324, 2020.

BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G. (2013) NeuroForest Star. Patente: Programa de Computador. Número do registro: 13410-5, data de registro: 30/04/2013, título: "NeuroForest Star", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial.

BINOTI, D. H. B. Estimation of height of eucalyptus trees with Neuroevolution of augmenting topologies (NEAT). **Revista Árvore**, v. 41, n.3, p. e410314, 2017.

BOURQUE, C. P. A.; BAYAT, M.; ZHANG, C. An assessment of height–diameter growth variation in an unmanaged Fagus orientalis-dominated forest. **European Journal of Forest Research**, v.138, p. 607–621, 2019.

DIAMANTOPOULOU, M.J.; OZCELIK, R. Evaluation of different modeling approaches for total tree-height estimation in Mediterranean Region of Turkey. **Forest Systems**, v.21, n.3, p. 383-397, 2012.

GHOSH, S. M.; BEHERA, M. D.; PARAMANIK, S. Canopy height estimation using sentinel series images through machine learning models in a Mangrove Forest. **Remote sensing**, v.12. n.9, 2020.

LEE, J.; IM, J.; KIM, K.-M.; QUACKENBUSH, L.J. Machine Learning Approaches for Estimating Forest Stand Height Using Plot-Based Observations and Airborne LiDAR Data. **Forests**, v. 9, n.5, p.268-284, 2018.

MARTINS; E. R.; BINOTI, M. L. M.; LEITE, H. G.; BINOTI, D. H. B. DUTRA, G. C. Configuração de redes neurais artificiais para estimação da altura total de árvores de eucalipto. **Revista Brasileira de Ciências Agrárias (Agrária)**, v.11, n.2, p.117-123, 2016.

OU, Q.; LEI, X.; SHEN, C. Individual Tree Diameter Growth Models of Larch–Spruce–Fir Mixed Forests Based on Machine Learning Algorithms. **Forests**, v.10, n.2, p.187-207, 2019.

RIBEIRO, A. et al. Estratégias e metodologias de ajuste de modelos hipsométricos em plantios de *Eucalyptus* sp. **Cerne,** v. 16, n. 1, p. 22-31, 2010.

SOARES, C. P. B.; PAULA NETO, F.; SOUZA, A. L. **Dendrometria e inventário florestal**. Viçosa: Editora UFV, 2011. 272p.

SOARES, T. S.; SCOLFORO, J. R. S.; FERREIRA, S. O.; MELLO, J. M. Uso de diferentes alternativas para viabilizar a relação hipsométrica no povoamento florestal. **Revista Árvore**, v.28, n.6, p. 845-854, 2004.

TAKAYAMA, T.; UMEZAWA, T.; KOMURO, N.; OSAWA, N. A regression model-based method for indoor positioning with compound location fingerprints. Geo-Spat. **Information Science**, v. 22, p.107–113, 2019.

WAGNER, S. et al. Architecture and Design of the HeuristicLab Optimization Environment. In **Advanced Methods and Applications in Computational Intelligence**, Topics in Intelligent Engineering and Informatics Series, Springer, pp. 197-261. 2014.

VIANNA, L. M. et al. Modelos hipsométricos para um povoamento de *Eucalyptus urophylla* x *Eucalyptus grandis* no município de Vitória da Conquista-BA. **Enciclopédia Biosfera**, v. 13, n. 24, p. 746–754, 2016.

VIEIRA, G. C.; MENDONÇA, A. R.; SILVA, G. F.; ZANETTI, S. S.; SILVA, M. M.; SANTOS, A. R. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. **Science of the Total Environment,** v. 619, p. 1473–1481, 2018.

XIONG, B.M.; WANG, Z.X.; LI, Z.Q.; ZHANG, E.; TIAN, K.; LI, T.T.; LI, Z.; SONG, C.L. Study on the correlation among age, DBH and tree height of the Pseudotsuga sinensis in Qizimei Mountain Nature Reserve. **Forest Resources Management**., v. 4, p. 41–46, 2016.

4.4 Artigo 4 - Modelagem hipsométrica via regressão simbólica com variável ambiental

RESUMO

O objetivo deste artigo é estimar a altura total das árvores em plantios de eucalipto com adição de varáveis ambientais, usando o método de regressão simbólica via programação genética. A área de estudo está localizada na região norte do estado de Minas Gerais, compreendendo uma área de plantio clonal de *Eucalytus ssp*. A base de dados foi composta por dados dendrométricos provenientes de inventário florestal e dados climáticos oriundos de estações meteorológicas. O período estudado compreendeu os anos de 2008 a 2018. Para o estudo foi selecionado um clone amplamente distribuído, implantados em seis espaçamentos, com idade variando entre 3 e 11 anos. Das estações foram obtidos os dados de precipitação pluviométrica ($mm.ano^{-1}$) e temperatura média mensal (ºC). A correlação foi de 0,92 e 0,91 e o RMSE foi de 5,25% e 5,49% para os dados sem e com variáveis ambientais, respectivamente. As tendências de subestimativa observada no modelo sem precipitação persistiram também no modelo com precipitação. A altura não apresentou forte correlação com a precipitação e temperatura, logo a inclusão das variáveis ambientais não gerou estimativas de altura mais precisas.
Palavras- chave: Precipitação. Programação genética. Temperatura

ABSTRACT

The objective of this paper is to estimate the total height of trees in eucalyptus plantations with the addition of environmental variables, using the symbolic regression method by genetic programming. The study area is located in the northern region of the state of Minas Gerais, comprising an area of clonal plantation of *Eucalytus* ssp. The database was composed of dendrometric data from a forest inventory and climatic data from meteorological stations. The studied period comprised the years from 2008 to 2018. For the study, a widely distributed clone was selected, planted in six spacings, with ages ranging from 3 to 11 years. Rainfall (mm.year-1) and average monthly temperature (ºC) data were obtained from the stations. The correlation was 0.92 and 0.91 and the RMSE was 5.25% and 5.49% for the data without and with environmental variables, respectively. The underestimation trends observed in the model without precipitation persisted in the model with precipitation as well. Height did not show a strong correlation with precipitation and temperature, so the inclusion of environmental variables did not generate more accurate height estimates.
Key words: Precipitation. Genetic programming. Temperature

INTRODUÇÃO

No manejo florestal, gerar equações e métodos confiáveis para estimar a altura do povoamento é uma tarefa muito importante. Sua medição ou estimativa é amplamente utilizada para calcular o volume, incremento em altura e em alguns casos, pode indicar a qualidade produtiva de um local (SILVA et al, 2012).

O estudo de modelos, procedimentos e equipamentos para medir a altura das árvores é muito importante porque a medição da altura das árvores é considerada uma parte importante no custo do inventário florestal (BINOTI et al, 2013).

A relação hipsométrica (relação entre Dap e altura) depende de fatores ambientais e de características do povoamento, como: capacidade produtiva, idade, genótipo e espaçamento do plantio (CURTIS, 1967). O crescimento do eucalipto no Brasil tem uma das maiores taxas do mundo, com estudos que comprovam sua sensibilidade às variáveis de solo e de clima (STAPE, 2002). Variáveis ambientais como excedente hídrico, precipitação, temperatura média e déficit de pressão de vapor, afetam diretamente o crescimento do eucalipto (FERREIRA, 2009).

O desenvolvimento de modelos associando variáveis ambientais e produção florestal foi realizado em alguns trabalhos como pode-se citar Snowdon et al. (1999), Maestri (2003) e Ferreira (2009), para modelagem do crescimento e da produção e Ferraz Filho et al. (2011) e Castro Neto (2015) para modelagem de altura dominante. Esses trabalhos se diferenciam pelas variáveis ambientais utilizadas, método de seleção e forma de inclusão das mesmas.

Métodos de inteligência artificial (IA) têm se destacado nos últimos anos nas ciências florestais. Um deles ainda não tão usado é a regressão simbólica via programação genética (PG) que é um método de encontrar a melhor equação de ajuste através da aplicação de operações genéticas, com base na teoria da seleção natural de Darwin em que indivíduos que se adaptam melhor ao seu ambiente têm uma maior possibilidade de sobreviver e de passar as suas características genéticas para os seus descendentes, gerando automaticamente soluções que naturalmente resolvem melhor o problema investigado (SPINOZA; POZO, 2003). A PG pode ser considerada como uma extensão dos algoritmos genéticos (AG) em que a definição fixa do problema (principal limitação da AG) é evitada com a ajuda de árvores de comprimento variável em vez de indivíduos de tamanho fixo (KOZA, 1994).

Sendo assim, visando aumentar a precisão das estimativas de altura o objetivo deste artigo é estimar a altura total das árvores em plantios de eucalipto com adição de variáveis ambientais, usando o método de regressão simbólica via programação genética.

MATERIAIS E MÉTODOS

Localização

A área de estudo está localizada na região norte do estado de Minas Gerais, compreendendo uma área de plantio clonal de *Eucalytus ssp*. A precipitação média anual na região é de 1100 mm. A temperatura média é de 22,5 ºC, com pequena amplitude. Segundo a classificação de Köppen (ALVARES et al., 2013) o clima da região é classificado em dois tipos:

Aw: tropical com inverno seco. Apresenta estação chuvosa no verão, de novembro a abril, e nítida estação seca no inverno, de maio a outubro (julho é o mês mais seco). A temperatura média do mês mais frio é superior a 18ºC. As precipitações são superiores a 750 mm anuais, atingindo 1800 mm.

Cwa: subtropical com inverno seco (com temperaturas inferiores a 18ºC) e verão quente (com temperaturas superiores a 22ºC).

Base de dados

A base de dados foi composta por dados dendrométricos provenientes de inventário florestal e dados climáticos oriundos de estações meteorológicas. O período estudado compreendeu os anos de 2008 a 2018.

Para o estudo foram selecionados todos os talhões pertencentes a um mesmo clone amplamente distribuído em toda a área, implantados em seis espaçamentos (3 m x 1 m; 3 m x 2 m; 3 m x 3 m; 4 m x 2 m; 6 m x 1m e 7 m x 1 m). Com idade variando entre 3 e 11 anos.

Dados Climáticos

Na área de estudos existem duas estações meteorológicas automáticas, sendo uma localizada em Três Marias e a outra localizada em Pirapora. Cada talhão foi associado a uma estação meteorológica em função da distância. Para tanto, determinou-se o centro geométrico de cada talhão e calculou-se sua distância em relação a cada uma das estações. A menor distância definiu a associação entre a estação e o talhão.

Das estações foram obtidos os dados mensais de precipitação pluviométrica e temperatura média. A precipitação foi somada e feita a média anual em $mm.ano^{-1}$.

Regressão simbólica

Para o estudo se utilizou o software HeuristicLab versão 3.3.16, desenvolvido por membros do *Heuristic and Evolutionary Algorithms Laboratory* (WAGNER et al, 2014).

A RS utiliza os conceitos da PG para o processamento dos dados e criação de equações que melhor descrevem as relações entre as variáveis dependente e independentes. Assim, os parâmetros iniciais para o algoritmo de regressão simbólica via programação genética foram definidos após a experimentação inicial, de maneira que fosse assumido um compromisso entre velocidade e desempenho (BARMPALEXIS et al., 2011) e se encontram na tabela 1.

Tabela 1 - Valores dos parâmetros utilizados na regressão simbólica

Parâmetro	Valor
Tamanho da população	100
Gerações	10000
Taxa de *crossover*	0,90
Taxa de mutação	0,15
Função *fitness*	Erro médio quadrático
Conjunto de funções	+, -, x, ÷ , exp, log, sen, cos, tan, tanh
Elitismo	1
Tipo de seleção	Classificação geral
Profundidade	8
Comprimento	20

Fonte: Elaborada pela autora, 2021

Os dados disponíveis foram separados em 70% para treinamento dos modelos e 30% para validação. As variáveis consideradas como entradas foram Idade e DAP e Idade, DAP, Precipitação e Temperatura para os modelos sem e com variáveis ambientais, respectivamente. A variável de saída foi a altura total.

Avaliação

A avaliação da correlação entre as variáveis do ambiente teve seu efeito avaliado pela matriz de correlação de Pearson.

Inicialmente foi feito um ajuste sem adição das variáveis ambientais e em seguida foi realizado outro ajuste com a adição destas o para fins de comparação.

Para avaliar a qualidade das estimativas, foram calculadas a correlação(r) (Equação 1), média do erro quadrático (MSE) (Equação 2), média do erro absoluto (MAE) (Equação 3), Raiz quadrada do erro médio em porcentagem (EMP) (Equação 4).

$$\boldsymbol{r = \frac{S_{y\hat{y}}}{S_y S_{\hat{y}}}} \quad (1)$$

$$\boldsymbol{MSE = \sum_{i}^{n} \frac{(y - \hat{y})^2}{n}} \quad (2)$$

$$\boldsymbol{MAE = \sum_{i}^{n} \frac{|y - \hat{y}|}{n}} \quad (3)$$

$$\mathbf{RMSE\% = \frac{1}{\bar{y}} \sqrt{\sum_{i}^{n} \frac{(y - \hat{y})^2}{n}} * 100} \quad (4)$$

Em que: SQR= Soma de quadrado dos resíduos; SQT= Soma de quadrados totais; n= número de amostras; y=altura observada; ŷ=altura estimada.

RESULTADOS

As estatísticas descritivas das variáveis ambientais e biométricas estão apresentadas na Tabela 2. A tabela dá uma ideia da variação dos dados utilizados neste estudo.

Tabela 2 - Valores médios, mínimos e máximos das variáveis do povoamento

Variável	Média	Mínimo	Máximo
DAP (cm)	13,4	5,0	20,2
Altura (m)	21,4	4,1	32,1
Precipitação (mm/ano)	1056,4	726,4	1207,2
Temperatura média (ºC)	22,5	22,1	24,6

Fonte: Elaborada pela autora, 2021

Inicialmente o modelo foi ajustado sem adição das variáveis ambientais. O resultado desse ajuste é mostrado pela equação 4. E as estatísticas encontradas na Tabela 3.

$$H = (\log(((\log(c_0 . DAP . log(\, c_1 . Id) + c_2) + ((\sin(\, c_3 . Id)) + c_4) + (c_5 . Id \,.\log(\, c_6 . Id)))) . c_7 + c_8) \quad (4)$$

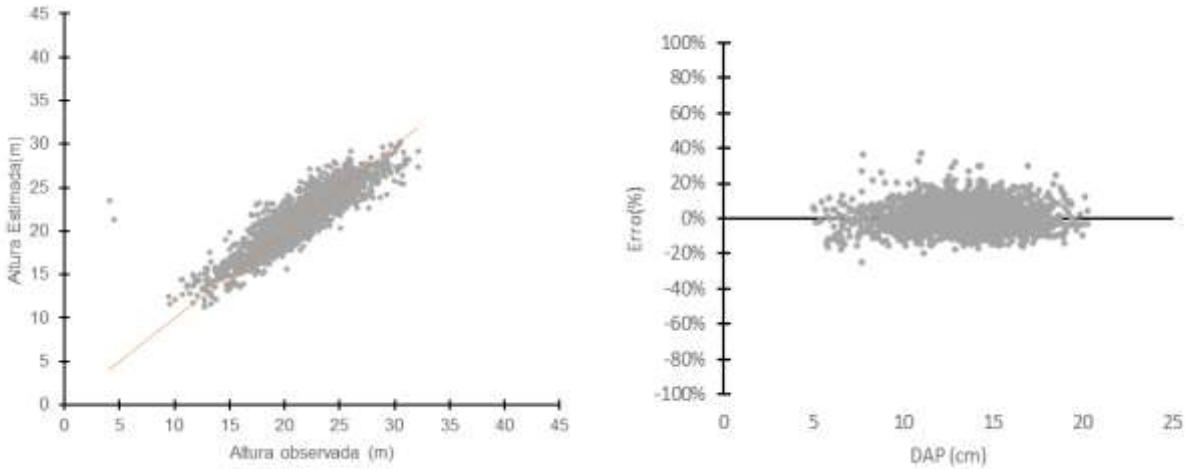
Tabela 3 - Parâmetros e estatísticas do ajuste, sem variáveis ambientais.

Parâmetro	Estimativa	Estatística	
C_0	0,6869	Treinamento	Validação
C_1	2,0354	r =0,89	r=0,92
C_2	14,639	MSE=2,59	MSE=1,97
C_3	1,1772	MAE = 1,20m	MAE=1,12 m
C_4	15,278	RMSE=5,59%	RMSE=5,25%
C_5	2,1934		
C_6	2,4548		
C_7	27,821		
C_8	-88,972		

Nota: r: correlação entre valores observados e valores estimados; MSE: média do erro quadrático; MAE: média do erro absoluto; RMSE%: raiz quadrada do erro médio em porcentagem.
Fonte: Elaborada pela autora, 2021

O erro médio absoluto (MAE), que significa o erro médio para a estimativa de altura em função do DAP foi de 1,20 m para o treino e 1,12 m para validação. A correlação foi de 0,89 para o treino e 0,92 para validação. O gráfico da Figura 1 apresenta a relação entre a altura real e estimada e análise de resíduos.

Figura 1 - Relação entre a altura real e estimada e gráfico de resíduos sem inclusão das variáveis ambientais.



Fonte: Elaborada pela autora, 2021

Na Tabela 4 é apresentada a correlação entre as variáveis. Todas as variáveis apresentaram correlações significativas embora somente o DAP e a idade tenham apresentado correlação forte (>0,60) com a altura.

Tabela 4 - Matriz de correlação das variáveis consideradas como entrada (DAP, Idade, Precipitação e Temperatura Média) e saída (Altura) do modelo de regressão simbólica.

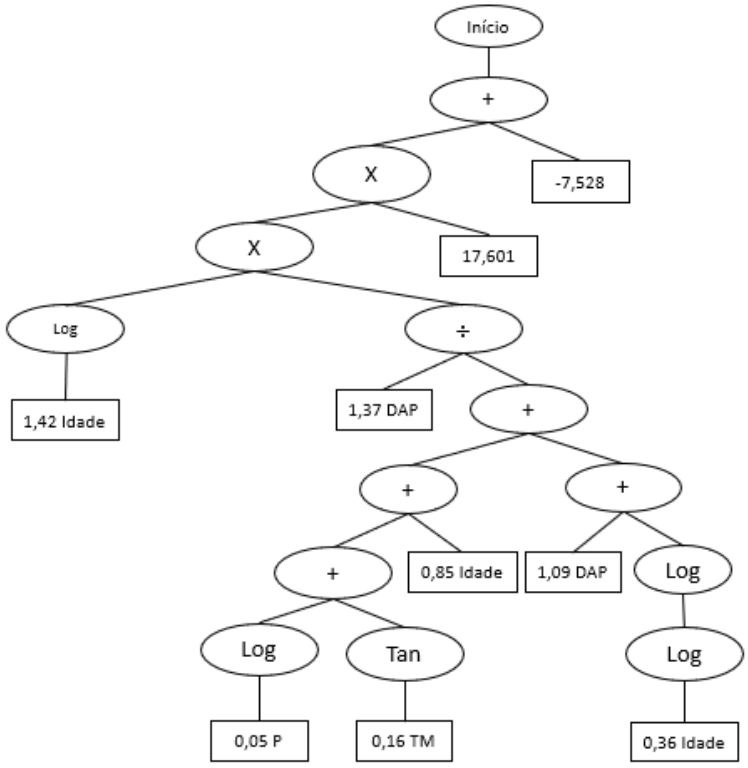
	DAP	Altura	Idade	Precipitação	Temp. média
DAP	1,00	0,83*	0,40*	0,26*	-0,28*
Altura	0,83*	1,00	0,61*	0,35*	-0,40*
Idade	0,40*	0,61*	1,00	0,53*	-0,52*
Precipitação	0,26*	0,35*	0,53*	1,00	-0,87*
Temp. média	-0,40*	-0,40*	-0,52*	-0,87*	1,00

Nota: * Significativo a 5% de probabilidade

Fonte: Elaborada pela autora, 2021

O resultado da regressão simbólica com a inclusão das variáveis ambientais pode ser visto Figura 2 que pode ser convertido na equação 5. E na tabela 6 são apresentados os parâmetros e as estatísticas do modelo.

Figura 2 - Representação da árvore hierárquica de busca da regressão



Fonte: Elaborada pela autora, 2021

$$H = ((\log{(c_0 . Id)} . \frac{c_1 . DAP}{((\log(c_2 . P) + \tan{(c_3 . TM)} + c_4 . Idade) + c_5 . DAP + \log{(\log{c_6 . Id)}}))} . c_7 + c_8) \quad (5)$$

Onde: H= Altura estimada em m; C_s= parâmetros do modelo; P= precipitação em mm/ano; DAP= diâmetro a altura do peito em cm; Id= Idade em anos; TM= temperatura média em ºC.

Tabela 6 - Parâmetros e estatísticas do ajuste do modelo com precipitação

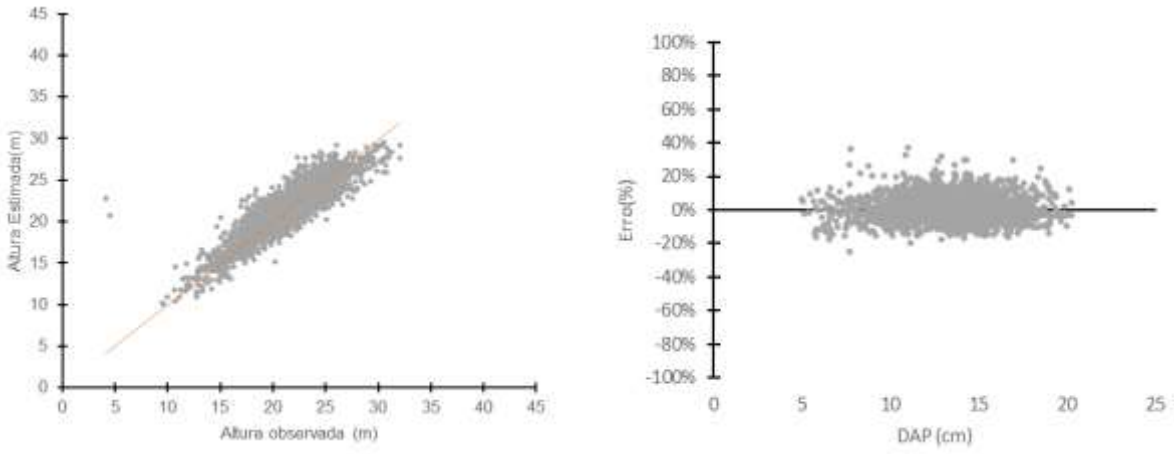
Parâmetro	Estimativa	Estatística	
C_0	1,419	Treino	Validação
C_1	1,375	r= 0,89	r=0,91
C_2	0,046	MSE=2,68	MSE=2,11m
C_3	0,161	MAE= 1,23m	MAE=1,16
C_4	0,845	RMSE=5,70%	RMSE=5,49%
C_5	1,089		
C_6	0,356		
C_7	17,601		
C_8	-7,528		

Nota: r: correlação entre valores observados e valores estimados; MSE: média do erro quadrático; MAE: média do erro absoluto; RMSE%: raiz quadrada do erro médio em porcentagem.
Fonte: Elaborada pela autora, 2021

Nota-se que as estatísticas de treino aqui apresentadas são praticamente iguais àquelas apresentadas na Tabela 4. Para o treino o coeficiente a correlação foi de 0,89 assim como no modelo anterior e o erro médio absoluto (MAE) foi de 1,23 m, enquanto no primeiro foi de 1,20 m. A validação também obteve dados semelhantes aos anteriores.

O gráfico apresentado na Figura 3 apresenta a relação entre a altura real e a estimada. E a análise gráfica de resíduos. Observa-se que não houve nenhuma mudança significativa na dispersão dos dados.

Figura 3 - Relação entre a altura real e a estimada e análise gráfica de resíduos com o modelo com variáveis ambientais



Fonte: Elaborada pela autora, 2021

DISCUSSÃO

No Figura 1 pode-se perceber que há uma pequena tendência de subestimativa para as árvores maiores. Na Tabela 4 demonstra-se que não houve ganhos em precisão com a adição das variáveis

ambientais. Resultados diferentes dos encontrados por Maestri (2003) que utilizou precipitação e temperatura média como modificadores de um modelo e conseguiu um ganho de 6,9% em relação ao modelo original sem as variáveis ambientais. Isso pode ter ocorrido devido a pequena correlação (menor que 50 %) da variável altura com as variáveis ambientais. Comparando-se os dois gráficos (Figuras 1 e 3) também não é possível perceber nenhuma alteração significativa, uma vez que a subestimativa observada no primeiro persistiu no segundo.

Em vários estudos que usaram variáveis ambientais houve um aumento de precisão das estimativas dos modelos como é o caso de Ferreira (2009) que utilizou variáveis ambientais para estimar índice de sítio e área basal em modelos de crescimento e produção e em ambos os casos se melhorou a precisão dos modelos. E Ferraz Filho et al. (2009) que inseriram variáveis climáticas (precipitação e radiação solar) no parâmetro de inclinação do modelo de Chapman e Richards e obtiveram estimativas de projeções de altura dominante mais precisas.

Então, como observado as variáveis ambientais sempre costumam mostrar influência sobre as estimativas dos modelos, apesar de não ter sido o caso do presente trabalho. Isso pode ter se dado devido ao fato da proximidade dos talhões deste estudo, pois as duas estações meteorológicas selecionadas não eram tão distantes, não havendo assim uma grande variação nas variáveis selecionadas. Já que a influência da temperatura e precipitação sobre a distribuição e o crescimento das florestas já é demonstrada pela alta correlação existente entre esses aspectos e as classificações climáticas, como as feitas por Merriam, Köppen e Thornthwaite (SPURR; BARNES, 1973).

CONCLUSÃO

A altura não apresentou forte correlação com a precipitação e a temperatura, logo a inclusão das variáveis ambientais não gerou estimativas de altura mais precisas.

REFERÊNCIAS

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. M.; SPAROVEK, G.; Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728. 2013.

BARMPALEXIS, P; KACHRIMANIS, K.; TSAKONAS, A.; GEORGARAKIS, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. **Information Sciences**, v. 107, n. 1, p. 75-82. 2011.

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.

CASTRO NETO, F. de. **Uso de variáveis climáticas para classificação de sítios em povoamentos de eucalipto**. 2015. 135 p. Dissertação (Mestrado em Engenharia Florestal)-Universidade Federal de Lavras, Lavras, 2015.

CURTIS, R. O. Height diamenter and height diamenter age equationsfor second growth Douglas-fir. **Forest Science**, Washington, v. 13, n. 4, p. 356-375, 1967.

FERRAZ FILHO, A. C.; SCOLFORO, J. R. S.; FERREIRA, M. Z.; MAESTRI, R.; ASSIS, A. L.; OLIVEIRA, A. D.; MELLO, J. M. Dominant height projection model with the addition of environmental variables. **Cerne**, Lavras, v. 17, n. 3, p. 427- 433, 2011.

FERREIRA, M. Z. **Modelagem da influência de variáveis ambientais no crescimento e produção de *Eucalyptus* sp**. 2009. 101 p. Tese (Doutorado em Engenharia Florestal) - Universidade Federal de Lavras, Lavras, 2009.

KOZA, J. R. Genetic prorgamming as means for programming computers by natural selection, **Stat. Comput.** v.4, p.87–112, 1994.

MAESTRI, R. **Modelo de crescimento e produção para povoamentos clonais de *Eucalyptus grandis* considerando variáveis ambientais**. 2003. 143 p. Tese (Doutorado em Engenharia Florestal) - Universidade Federal do Paraná, Curitiba, 2003.

SILVA, G. F.; CURTO, R. de A.; SOARES, C. P. B.; PIASSI, L. de C. Avaliação de métodos de medição de altura em florestas naturais. **Revista Árvore**, Viçosa, v. 36, n. 2, p. 341-348, 2012.

SNOWDON, P.; JOVANOVIC, T.; BOOTH, T. H. Incorporation of índices of anual climatic variation into growth models for *Pinus radiata*. **Forest Ecology and Managment**, Amsterdam, v117, p.187-197, 1999.

SPINOZA, E.; POZO, A. Controlling the population size in genetic programming, in: E. CANTU-PAZ, J.A. FOSTER, K. DEB, L.D. DAVIS, R. ROY, U.M. O'REILLY, H.G. BEYER, R. STANDISH, G. KENDALL, S. WILSON, M. HARMAN, J. WEGENER, D. DASGUPTA, M.A. POTTER, A.C. SCHULTZ, K.A. DOWSLAND, N. JONOSKA, J. MILLER (Eds.), **Genetic and Evolutionary Computation-GECCO**, Springer, Berlin, Heidelberg, 2003, p. 1975–1985.

SPURR, S. H.; BARNES, B. V. **Forest ecology**. Ronald Press Company, 1973. 571 p.

STAPE, J. L. **Production ecology of clonal Eucalyptus plantations in northeastern Brazil**. 2002. 212 p. Thesis (Doctor of Philosophy) - Colorado State University, Fort Collins, 2002.

WAGNER, S. et al. Architecture and Design of the HeuristicLab Optimization Environment. In **Advanced Methods and Applications in Computational Intelligence**, Topics in Intelligent Engineering and Informatics Series, Springer, pp. 197-261. 2014.

5 CONSIDERAÇÕES FINAIS

A regressão simbólica se mostrou um método viável e eficiente para estimativas hipsométricas, apresentando superioridade aos modelos hipsométricos tradicionais. Quando comparada à outra técnica de inteligência artificial (RNA) atingiu resultados semelhantes, mas não superiores.

A adição de variáveis categóricas ao modelo de regressão simbólica ocasionou melhoras na precisão das estimativas, sendo as variáveis dap, idade, projeto e clone utilizadas em conjunto as que apresentaram os melhores resultados.

Por se tratar de um tema novo dentro das ciências florestais, recomenda-se ainda que novos estudos sejam realizados sobre o assunto, como o estudo de outras configurações e utilização de outras bases de dados, para que assim a técnica possa ser aprimorada e consolidada.