EVANDRO NUNES MIRANDA

HIPSOMETRIA: SELEÇÃO DE VARIÁVEIS E MINERAÇÃO DE DADOS POR MÉTODOS DE INTELIGÊNCIA COMPUTACIONAL

LAVRAS – MG

2020

EVANDRO NUNES MIRANDA

HIPSOMETRIA: SELEÇÃO DE VARIÁVEIS E MINERAÇÃO DE DADOS POR MÉTODOS DE INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia Florestal, área de concentração em Ciências Florestais, para a obtenção do título de Mestre.

Prof. Dr. Lucas Rezende Gomide

Orientador

Prof. Dr. Bruno Henrique Groenner Barbosa

Coorientador

LAVRAS – MG

2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Miranda, Evandro Nunes.
Hipsometria: seleção de variáveis e mineração de dados por métodos de inteligência computacional / Evandro Nunes Miranda. - 2020.
87 p.

Orientador(a): Lucas Rezende Gomide.
Coorientador(a): Bruno Henrique Groenner Barbosa.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Feature Selection. 2. Genetic Algorithm. 3. Florestas Nativas. I. Gomide, Lucas Rezende. II. Barbosa, Bruno Henrique Groenner. III. Título.

EVANDRO NUNES MIRANDA

HIPSOMETRIA: SELEÇÃO DE VARIÁVEIS E MINERAÇÃO DE DADOS POR MÉTODOS DE INTELIGÊNCIA COMPUTACIONAL

HYPSOMETRY: FEAUTURE SELECTION AND DATA MINING BY COMPUTER INTELLIGENCE METHODS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia Florestal, área de concentração em Ciências Florestais, para a obtenção do título de Mestre.

APROVADA em 31 de Janeiro de 2020.

Prof. Dr. Lucas Rezende Gomide	UFLA
Prof. Dra. Carolina Souza Jarochinski e Silva	UFLA
Dr. Henrique Ferraço Scolforo	SUZANO

Prof. Dr. Lucas Rezende Gomide

Orientador

Prof. Dr. Bruno Henrique Groenner Barbosa

Coorientador

LAVRAS – MG

2020

AGRADECIMENTOS

A presente dissertação de mestrado é resultado de muito esforço, estudo e empenho. Mas, só consegui chegar até aqui pois na minha vida existem pessoas que me acompanharam e me deram um apoio precioso para a realização deste sonho, por isso, expresso aqui minha gratidão a todas elas.

Primeiramente, quero agradecer a Deus por mais essa oportunidade que Ele tornou possível, por todas as bênçãos recebidas e por todas as que ainda irei vivenciar na minha existência!

Agradeço aos meus pais pelo incentivo, apoio, amor e compreensão perante minhas ausências nesse período. Obrigado por desejarem sempre o melhor para mim, pelos esforços diários para que eu supere os obstáculos e realize os meus sonhos. Vocês foram fundamentais para a concretização deste trabalho!

Agradeço também aos meus irmãos e a minha família por sempre torcerem por mim e desejarem o melhor para a minha vida, esse carinho é recíproco.

À minha namorada, por tudo o que você transformou na minha vida. Obrigado pelo teu carinho, tua ajuda, tua alegria, tua atenção, tua vibração com as minhas conquistas e teu ombro em cada momento difícil que você ajudou a atravessar. Sem você essa conquista não teria o mesmo gosto. E também a sua família, que me trata como como se fosse membro dela. Obrigado meu amor. Te amo.

Minha gratidão especial ao Prof. Lucas Rezende Gomide, meu orientador, por todo incentivo, convívio, empenho e orientação que ampliaram ainda mais meu aprendizado e contribuíram de forma significativa não só para a consolidação deste trabalho, mas que com certeza será útil na minha carreira. Seu auxilio foi de fundamental importância. Agradeço por sempre acreditar no meu potencial.

Ao meu co-orientador Bruno Henrique Groenner Barbosa, por todo suporte que lhe coube, suas correções e incentivos, muito obrigado.

Desejo igualmente agradecer a todos os meus colegas do laboratório, cujo companheirismo esteve presente em todos os momentos. Agradeço também o corpo docente da instituição e aos técnicos administrativos que foram sempre prestativos e tornaram possível a concretização deste trabalho.

Agradeço também aos membros da banca de Defesa de Mestrado, que tão gentilmente aceitaram participar e contribuir para o desenvolvimento desta dissertação com conselhos e sugestões que foram de grande valia ao longo desse período.

Agradeço a Universidade Federal de Lavras (UFLA) pela oportunidade de continuar trilhando meu caminho em uma das melhores Universidades Federais do nosso país desde a graduação.

Agradeço também a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio para a realização do presente trabalho.

Por fim, a todos aqueles que contribuíram, direta ou indiretamente, para a realização desta dissertação, o meu sincero agradecimento.

RESUMO GERAL

O uso assertivo de variáveis dendrométricas impacta diretamente no planejamento florestal, deste modo, resultados mais precisos se faz necessário. A altura, se destaca entre as variáveis biométricas por ser um importante atributo usado comumente para os métodos de cálculo do volume e para a medida do incremento de altura e volume das árvores, etc. Assim, novas tecnologias e técnicas veem sendo implementadas nos últimos anos para auxiliar no seu cálculo. No contexto da estimativa da altura, modelos estatísticos tradicionais que tem uma boa resposta podem ser melhorados com técnicas de mineração de dados (*data mining*). Nessa dissertação, o princípio de *data mining* foi aplicado na seleção de variáveis (*feature selection*) nos dois capítulos para estimar a altura individual das árvores da bacia do Rio Grande - MG. No primeiro capítulo, o objetivo foi a seleção de variáveis dentro de modelos tradicionais da literatura, onde foram aplicadas possíveis combinações de variáveis como entrada em modelos não lineares, utilizando um algoritmo genético duplo, o primeiro seleciona e monta as combinações das variáveis, o segundo parametriza e ajusta o modelo construído. Os modelos gerados apresentaram um pequeno ganho nas estimativas, e a metodologia proposta se mostrou eficiente na busca por bons resultados, mas com dificuldades para encontrar bons resultados em problemas com muitas entradas. A proposta se mostrou robusta e pode ser aplicada a outros problemas. O segundo capítulo, buscou comparar métodos tradicionais de predizer a altura, com métodos de aprendizado de máquinas, na sua forma pura e híbrida. O modelo *Random Forest* (RF) com redução de variáveis se mostrou robusto, capaz de melhorar a resposta e reduzir o número de entradas no modelo de RF, apresentando melhores resultados aos demais. As técnicas que envolvem o uso de inteligência computacional se mostram eficazes na procura de bons resultados, com respostas superiores aos tradicionais, capazes de selecionar boas variáveis e estimar bons valores de altura.

Palavras chaves: *Feauture selection. Genetic Algotihm*. Florestas nativas.

GENERAL ABSTRACT

The assertive use of dendrometric variables has a direct impact on forest planning, so more accurate results are needed. Height stands out among biometric variables as it is an important attribute commonly used for volume calculation methods and for measuring height and volume increment of trees, etc. Thus, new technologies and techniques have been implemented in recent years to assist in their calculation. In the context of height estimation, traditional statistical models that have a good response can be improved with data mining techniques. In this dissertation, the principle of data mining was applied to feature selection in both chapters to estimate the individual height of the trees of the Rio Grande basin - MG. In the first chapter, the objective was to select variables within traditional literature models, where possible variable combinations were applied as input to nonlinear models, using a dual genetic algorithm, the first selects and assembles the variable combinations, the second parameterize and adjust the constructed model. The generated models presented a small gain in the estimates, and the proposed methodology proved to be efficient in the search for good results, but with difficulties to find good results in problems with many inputs. The proposal proved robust and can be applied to other problems. The second chapter sought to compare traditional methods of predicting height with machine learning methods in their pure and hybrid form. The Random Forest (RF) model with variable reduction proved to be robust, capable of improving the response and reducing the number of entries in the RF model, presenting better results to the others. Techniques that involve the use of computational intelligence are effective in the search for good results, with superior answers than traditional ones, capable of selecting good variables and estimating good height values.

Keywords: Feauture selection. Genetic Algotihm. Native forests.

LISTA DE FIGURAS

PRIMEIRA PARTE

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

ARTIGO 2

LISTA DE TABELAS

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

ARTIGO 2

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

O conhecimento de atributos e variáveis no setor florestal são uma importante ferramenta para análise e diagnósticos de parâmetros estocásticos do povoamento, sendo imprescindíveis no manejo do recurso florestal. A altura, por sua vez, é comumente usada como entrada de diferentes modelos dentro da modelagem no manejo florestal. A sua determinação em campo exige investimento, e pode ser estimada com o uso de modelos hipsométricos. Essa variável pode ainda expressar diversas interpretações produtivas e ecológicas, auxiliando ainda na estratificação de sítios e sua capacidade produtiva.

Os inventários florestais cada vez mais abrangem maiores extensões, exigindo um maior detalhamento de informações para caracterizar o povoamento. Todavia, este maior detalhamento fica susceptível a efeito de características de caráter não linear, difíceis de dimensionar em grandes bancos de dados (*big data*). Recentemente, com a crescente complexidade dos problemas florestais envolvendo grandes áreas, houve a necessidade do emprego de métodos robustos que lidassem com natureza linear ou não. Assim, o crescente avanço computacional permitiu o uso destas metodologias, deste modo, técnicas de aprendizagem de máquina (AM) e mineração de dados surgiram para suprir essa demanda. De acordo com Lidberg, Nilsson e Ågren (2019), o aprendizado de máquina é uma técnica de mineração de dados que identifica padrões dos conjuntos de dados, e os replica na previsão de novos dados. A utilização de métodos de inteligência se mostra extremamente eficiente na extração de conhecimento, como pode ser demonstrado em inúmeros trabalhos da literatura (HONG et al., 2018; POURRAHMATI et al., 2018; TUAN; DINH; LONG, 2019).

Os avanços nos métodos de coleta de dados, por exemplo sensores remotos, permitiu extrair um elevado número de variáveis que podem incrementar o poder preditivo de um modelo. Contudo, o ganho exponencial no número desses atributos gera o desafio de sua seleção. Os métodos de seleção de variáveis (*features selection)* são amplamente implementado na extração de informações estratégicas dos bancos de dados. Estes permitem diminuir a dimensionalidade e complexidade, potencializando os resultados e suas explicações (GHAEMI; FEIZI-DERAKHSHI, 2016). Ao passo que, a implementação de técnicas de seleção de variáveis, que podem ser divididos em métodos de *wrapper* (filtro), pode servir como uma ferramenta para identificar e remover atributos desnecessários, irrelevantes e redundantes,

melhorando o desempenho de previsão do modelo, reduzindo sua complexidade, e proporcionando uma solução mais simples (HONG et al., 2018).

Diante disso, objetiva-se nesta dissertação empregar algoritmos da área de aprendizagem de máquina, em destaque o algoritmo *random forest* e métodos meta-heurística (algoritmo genético) em problemas envolvendo a seleção de variáveis explicativas na altura das árvores. O capítulo 2 teve como objetivo a seleção de variáveis e sua alocação em modelos não lineares de regressão. O princípio do método consistiu na combinação ou não de variáveis, em conjunto com operadores matemáticos, utilizando um algoritmo genético para isto. Ele funciona como um filtro (*wrapper*) para seleção das variáveis em modelos não lineares da literatura. De forma posterior, utiliza outro algoritmo genético para parametrizar e ajustar a construção dos modelos.

No capítulo 3 objetiva-se realizar a modelagem da altura em função de métodos tradicionais, com modelos clássicos de regressão, e abordagens com aprendizagem de máquina na seleção de variáveis e ajuste hipsométrico da vegetação nativa na bacia do Rio Grande, Minas Gerais, utilizando grande quantidade de dados de diferentes fontes, resoluções e formatos.

2 REVISÃO DE LITERATURA

2.1 Hipsometria

A altura individual das árvores é uma variável fundamental a ser obtida em povoamentos florestais, comumente utilizada para determinar volume, carbono e outros importantes atributos da floresta. Esta variável ainda apresenta uma série de interpretações de produtividade e biológicas, que ajudam comutar na estratificação da dos plantios.

A altura constitui-se uma importante característica da árvore e pode ser medida ou estimada (SILVA et al., 2012). Existem uma série de maneiras de quantificar a altura, algumas delas sendo: através de medições com hipsômetros; subindo na árvore (para cubagem em florestas nativas); com trena (no caso de árvores abatidas); com varas de medição, e etc. Outra forma muito usual é através de estimativas, a qual consiste em utilizar relações hipsométricas (relações entre altura e diâmetro) para determinar a altura de árvores não medidas, visto a atividade cara e onerosa em medir a altura de todas as árvores que compõem a população (SCOLFORO, 2006).

A obtenção da altura por relações diâmetro altura se dá principalmente pelo uso de modelos de regressão, e é amplamente diversificado nas inúmeras áreas da ciência, para diferentes contextos e aplicações. Os modelos tentam explicar a relação entre uma variável resposta Y e uma variável explicativa X. Conforme Schneider, Schneider e Souza (2009), o uso da análise de regressão sobre o aspecto de ferramenta, está amplamente ligado na solução de grande parte dos problemas florestais, principalmente na obtenção de estimativas dos parâmetros da floresta, ao qual utiliza-se relações biométricas que possibilitam a obtenção de valores estimados de forma indireta através de equações de regressão.

A determinação da altura por meio de relações hipsométricas é influenciado por uma série de características do povoamento, como os quais: espécie, posição sociológica, idade, tamanho de copa, densidade, sítio e práticas silviculturais (CURTO et al., 2014; MACHADO et al., 2008). Por isso a importância da determinação correta de qual equação hipsométrica utilizar.

Embora consolidadas as equações hipsométricas, novas práticas comerciais vêm sendo utilizadas na determinação da altura, como o uso de modelos de aprendizado de máquina (BINOTI; BINOTI; LEITE, 2013; CAMPOS et al., 2017; VENDRUSCOLO et al., 2015) e o uso de sensores aerotransportados (ALEXANDER; KORSTJENS; HILL, 2018; GARCÍA et al., 2018; LEE et al., 2018; SULLIVAN et al., 2017). A melhor acurácia das estimativas da

altura é um preponderante no planejamento florestal, dada sua importância nos aspectos estocásticos e biológicos dos povoamentos florestais.

2.2 Aprendizagem de máquina

Os aumentos consideráveis no número de observações dos conjuntos de dados nas últimas décadas exigiram esforços da comunidade científica na implementação de novas tecnologias que auxiliassem na resolução de problemas. Deste modo, surgiram métodos de solução a partir de regras programadas em máquinas com auxílio de regras estabelecidas por especialistas (CAMACHO et al., 2018).

Estes métodos de aprendizado avançaram consideravelmente nos últimos anos, e devido ao aumento da complexidade dos problemas e grande volume de dados gerados em diferentes setores da sociedade, tornaram seu uso independente, reduzindo a intervenção humana, se tornando uma ferramenta sofisticada graças aos avanços computacionais. As regras transmitidas pelos especialistas, eram usadas como técnicas capazes de criar-se por si própria, uma hipótese ou função, que através de experiência passada era capaz de resolver o problema a se tratar. Essas hipóteses, que se baseavam em experiência adquirida, deu-se o nome de aprendizagem de máquina (FACELLI et al., 2015).

Existem várias técnicas de aprendizagem de máquina, as quais podem ser divididas em cinco sub áreas que compartilham uma mesma semelhança: possuem inspiração nos seres vivos, através de aspetos biológicos, como a inteligência humana, o comportamento dos animas, as leis da física, a teoria evolutiva, entre outras (BRAGA et al., 2015). As áreas são subdivididas como (BRAGA et al., 2015; KONAR, 2005): redes neurais artificiais; computação evolutiva; inteligência de enxames; sistemas imunológicos artificiais e sistemas *fuzzy*.

O avanço das tecnologias e inovações das técnicas de inteligência computacional vem em emergente crescimento, obtendo sucesso em resolução de problemas complexos, nas mais distintas áreas de conhecimento, como por exemplo, na meteorologia (COUTINHO; SILVA; DELGADO, 2016), mineração de dados (HONG et al., 2018), otimização de sistema de energia, (RAHMAN; MOHAMAD-SALEH, 2018), seleção e classificação de dados (GONÇALVES e GONÇALVES et al., 2016; LOPATIN et al., 2016; NAGHIBI; POURGHASEMI; DIXON, 2016; YOUSSEF et al., 2016), aplicações financeiras (NGAI; XIU; CHAU, 2009; NGAI et al., 2011), dentre outros. Rahman e Mohamad-Saleh (2018) falam que os métodos de IA se transformaram nas últimas décadas, com a aplicação de inúmeras técnicas IA, tornando-se um imenso campo de pesquisa e trazendo uma série de benefícios as áreas estudadas.

2.2.1 Algoritmo Genético

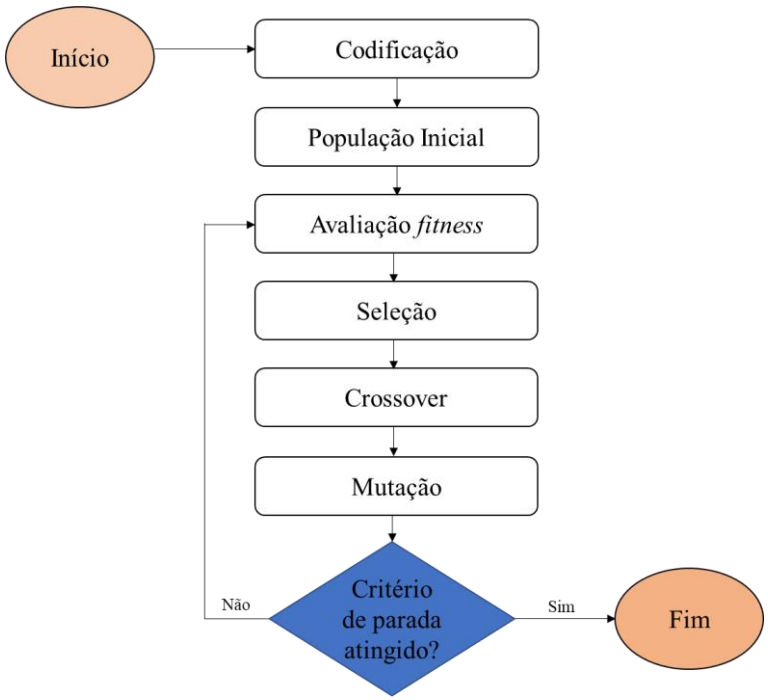
O algoritmo genético (AG) foi proposto inicialmente na década de 60 por John Holland, e foram amplamente estudados, experimentados e aplicados em várias disciplinas de engenharia (GARG, 2016). O AG é um método de otimização que supostamente simula a evolução biológica, que imita o comportamento de evolução coletiva do grupo. Assim, cada indivíduo representa uma solução aproximada do espaço de busca do problema.

O AG é inspirado no processo genético de organismos biológicos (YU; XU, 2014), e é um algoritmo de busca global. De acordo com Yang et al. (2008), em comparação com métodos tradicionais (método de busca exaustiva direta e o método de busca direcionado por gradiente) para otimização de funções, uma das principais vantagens do AG é a robustez na busca de soluções ótimas globais, particularmente em otimização multimodal e problemas multiobjetivo.

Das et al. (2018) falam que o AG considera várias soluções individuais, formando uma população, e fazem testes de convergência no escopo geral do espaço de pesquisa, levando a uma maior possibilidade de encontrar a solução ideal global. Ele utiliza um escalar simples como medida de precisão, o que não requer o uso de informações derivadas, e, portanto, são fáceis de usar e implementar. Logo, considerando as regras de sobrevivência/reprodução do mais apto, o AG explora continuamente novas e melhores soluções sem pré-suposições, como a continuidade e unimodalidade (KAO; ZAHARA, 2008). O algoritmo é muito útil na resolução de problemas de otimização, por funcionar corretamente mesmo que os parâmetros de entrada sejam ligeiramente alterados ou que tenha a presença de ruído razoável (DAS et al., 2018).

Uma definição complementar é dita por Pezzella; Morganti e Ciaschetti, (2008), onde o algoritmo genético segue o paradigma da evolução, assim, a partir de uma população inicial, o algoritmo aplica operadores genéticos para produzir descendentes (na terminologia de busca local, corresponde a explorar a vizinhança), que são presumivelmente mais adequados que seus ancestrais. A cada geração (iteração), cada novo indivíduo (cromossomo) corresponde a uma solução. A sobrevivência/reprodução dos indivíduos na população são promovidas pela eliminação de características indesejáveis e baixa adaptação ao problema (KALSI; KAUR; CHANG, 2018). A estrutura geral do AG pode ser descrita da seguinte maneira: codificação, população inicial, avaliação, operadores de seleção, *crossover* e mutação, geração de descendentes e critério de parada (Figura 1).

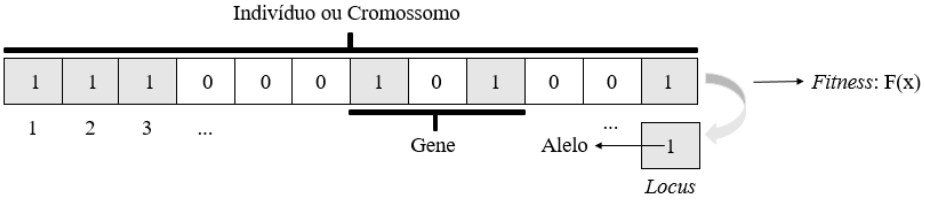
Figura 1 - Fluxograma da estrutura de um Algoritmo Genético Simples.



Fonte: Adaptado de Abdulhamed, Tawfeek e Keshk (2018).

Para a resolução de problemas de otimização, uma importante parte do AG é a codificação, afim de tornar o algoritmo mais apropriado para uma aplicação específica. Deste modo, cada solução é geralmente codificada como uma cadeia binária, real ou outra, chamada cromossomo. Linden (2012) comenta que a representatividade cromossomial mais simples e mais usada pela comunidade é a binária, isto é, um cromossomo ou indivíduo é nada mais do que uma sequência de bits (cada gene portanto, sendo um bit), mas pode assumir valores inteiros, reais, etc. O cromossomo é formado por um conjunto de genes, e podem ser um simples bit ou um pequeno grupo sequencial de bits adjacentes. A posição do gene no cromossomo denomina-se *locus* e pode expressar uma parte da resposta do problema. A figura 2 a seguir exemplifica um cromossomo.

Figura 2 - Cromossomo com seus respectivos genes, locus e fitness.



Fonte: Do autor (2020).

Os genes que compõem os cromossomos descrevem a solução do algoritmo, e a ordem em que aparecem, interferem diretamente nas respostas que virão com a evolução do AG. Cada cromossomo representa uma solução para o problema. A representação cromossomial é fundamental devido a capacidade de tradução da informação do problema, tornando de maneira viável a decodificação pelo computador. Deste modo, quanto mais adequada ao problema, maior a qualidade da resposta da função objetivo (*fitness*). A representação cromossomial é arbitrária, ficando a cargo do pesquisador a melhor maneira de se adequar ao problema, assim, algumas regras gerais devem ser seguidas (LINDEN, 2012): a) a representação deve ser a mais simples possível; b) caso tenha soluções proibidas ao problema, elas não devem ter representatividade e; c) caso tenha restrições ao problema, elas devem estar implícitas dentro da representação.

Um conjunto de cromossomos são formados de forma aleatória na etapa inicial do algoritmo genético para formar sua população inicial, de acordo com a codificação dada pelo autor. A população inicial pode ser considerada dentre as etapas do algoritmo genético, umas das mais importantes, uma vez que é por ela que todo o processamento passa, ou seja, participando diretamente na qualidade da resposta e no tempo de convergência (MICHALEWICZ, 1996). Usualmente, a geração de uma população depende de um gerador aleatório P (KALSI; KAUR; CHANG, 2018). Linden (2012) diz que o tamanho da população inicial influencia diretamente no desempenho do algoritmo. Caso seja muito pequeno, não haverá condições para que ocorra variabilidade genética dentro da população, o que acarretará em soluções que fogem do ótimo. Em casos de população muito grande, acarretará em tempo de processamento muito alto, o que leva a uma busca exaustiva.

Com a população gerada, é preciso avaliar seus indivíduos, assim, uma função de avaliação, denominada de *fitness* (aptidão ou objetivo), é utilizada para medir o desempenho dos indivíduos, associando a cada um deles uma nota. Segundo Ghamisi e Benediktsson (2015), o mérito de cada cromossomo é avaliado usando a função *fitness*, ou seja, a função de aptidão é basicamente uma função objetiva que é usada para resumir, como uma única figura de mérito, determinada solução capaz de atingir os objetivos definidos. Os valores contidos no cromossomo retornam um valor numérico, cujo significado é uma métrica da qualidade da solução obtida. Como os AGs são técnicas de otimização, a função *fitness* deve ser tal que se o cromossomo C_1 representa uma solução melhor do que o cromossomo C_2, assim, a avaliação de C_1 é superior à resposta de C_2. Na etapa de seleção geralmente é a função de *fitness*, que calcula a adequação de cada cromossomo individual na população, em termos de número real (KALSI, KAUR e CHANG, 2018).

O conjunto de cromossomos produzido para formar a população inicial, após avaliar, uma nova população deve ser produzida por diferentes operadores do AG (NEMATI; BRAUN; TENBOHLEN, 2018). A reprodução da população é obtida pela aplicação iterativa de um conjunto de operadores estocásticos, que geralmente consiste em Mutação, Cruzamento e Seleção (KALSI; KAUR; CHANG, 2018). Deste modo, cromossomos aptos são selecionados para a geração de novos cromossomos, se dá nome de seleção dos pais. O processo de seleção utiliza uma população de possíveis soluções para a resolução de problemas, os quais evoluem de acordo com operadores probabilísticos concebidos a partir de metáforas biológicas. Desse modo, há uma tendência de que, na média, os indivíduos representem soluções cada vez melhores à medida que o processo evolutivo continua (BARBOZA et al., 2015).

A operação de seleção consiste em escolher um indivíduo melhor da população inicial que pode ser aplicado à iteração subsequente (YU; XU, 2014). A seleção de pais consiste na simulação do processo de seleção natural que atua sobre as espécies biológicas, em que os pais mais capazes geram mais filhos, entretanto também permitindo que os pais menos aptos também possam gerar descendentes (PEREIRA; ALVES; MARINHO, 2018). Na literatura encontram-se diversos métodos para simular o mecanismo de seleção dos pais, os mais difundidos são método do torneio; elitista; método de amostragem estocástica uniforme; seleção local; seleção por ranking; seleção truncada; método da roleta viciada, conforme descrito por Pereira, Alves e Marinho (2018).

O operador de reprodução ou *Crossover* é uma operação de troca entre cromossomos dos indivíduos pais e produz novos indivíduos (YU; XU, 2014). Dentre os três operadores fundamentais, o operador de crossover é o mecanismo de evolução mais importante, pois ajuda muito o algoritmo a escapar de ótimos locais, produzindo alguns novos cromossomos candidatos (CHUANG; CHEN; HWANG, 2015). O objetivo do operador é gerar novos indivíduos, cada vez mais aptos, a solucionar um dado problema (PEREIRA; ALVES; MARINHO, 2018). Dentre os operadores de *crossover* mais estudados são: *crossover* de um ponto; *crossover* de dois pontos; *crossover* uniforme; *crossover* baseado em maioria (LINDEN, 2012; MITCHELL, 1996; PEREIRA; ALVES; MARINHO, 2018). Essas regras preservam a viabilidade de novos indivíduos, sendo que a taxa de cruzamento determina qual proporção dos cromossomos parentais deve ser recombinada (NEMATI; BRAUN; TENBOHLEN, 2018).

A mutação é uma outra etapa e permite gerar um novo indivíduo, este operador é usado para alterar os indivíduos aleatoriamente e manter uma diversidade genética através das gerações. A configuração da taxa de mutação é mais crítica que a da taxa de cruzamento. Se ela estiver definido com um valor alto de porcentagem, a busca das soluções se transformará em

uma pesquisa aleatória primitiva e, se estiver definido muito baixo, a diversidade da pesquisa será agravada (NEMATI; BRAUN; TENBOHLEN, 2018). Então, a mutação é aplicada na população para aumentar a aleatoriedade dos indivíduos e diminuir a possibilidade de ficar preso no ótimo local (GHAMISI; BENEDIKTSSON, 2015). A mutação na sua forma usual escolhe aleatoriamente alguns indivíduos da população, considerando uma dada taxa de mutação, e altera seus genes a um intervalo (NEMATI; BRAUN; TENBOHLEN, 2018). A taxa de cromossomos mutados será constante e uniforme para todas as gerações. Eventualmente, através dos operadores AGs (seleção, *crossover* e mutação), uma boa solução será encontrada combinando diferentes soluções possíveis (CHEN et al., 2010; BALIEIRO et al., 2014), ou seja, espera obter-se em média resultados melhores à medida que as gerações vão sendo geradas.

A existência de um critério de parada é de fundamental importância para o processo de finalização do algoritmo. Deste modo, Gomide (2009), discorre sobre a importância do critério de parada, a sua escolha deve ser realizada de forma que maximize o algoritmo. Rodrigues et al. (2004) falam das dificuldades e dos problemas enfrentados na confecção de um AG, dentre os problemas está estabelecer o critério de parada. A dificuldade se dá devido à complexidade do algoritmo em avaliar a qualidade da solução em dado instante da busca, são míopes, ou seja, não há critério no algoritmo para ele definir quando está próxima a solução ótima. Alguns mecanismos ajudam a estabelecer um critério de parada, e pode-se citar os seguintes (RODRIGUES et al., 2004): a) o número máximo de iterações; b) o tempo máximo de processamento e; c) a estabilização da função *fitness*.

2.2.2 *Random Forest*

O *Random Forest* (RF) é um algoritmo de aprendizado de máquina muito utilizado para problemas de regressão e classificação, foi implementado pelo matemático Breiman (2001). Ele permite uma modelagem flexível de interações em altas dimensões, criando um grande número de árvores de regressão e calculando a média de suas previsões (WAGER; ATHEY, 2018). As árvores são criadas desenhando um subconjunto de amostras de treinamento através da substituição (uma abordagem de ensacamento), ou seja, algumas podem ser selecionadas novamente, enquanto outras não (BELGIU; DRĂGU, 2016).

Para o ajuste do modelo, o algoritmo requer três parâmetros a serem definidos para produzir as árvores de decisão: *ntree* (número de árvores treinadas na floresta), *nodesize* (tamanho do nó do terminal de destino) e *mtry* (número de recursos aleatórios usados para

dividir um nó da árvore) (O'BRIEN; ISHWARAN, 2019). Ele se divide em dois níveis de randomização, o primeiro é a agregação de *bootstrap* ("ensacamento"), onde um subconjunto aleatório de dois terços das observações são usadas para treinar as árvores, e o terço restante dos dados (observações "fora do saco") são excluídos para validação (WOZNICKI et al., 2019). O segundo nível é referente a cada nó das árvores de decisão, onde de forma randômica é selecionado um número de variáveis, e a variável que apresentar a melhor divisão é selecionada para aumentar a árvore nesse nó (WOZNICKI et al., 2019). Para a avaliação de cada nó, é utilizado métricas. Para problemas de regressão a métrica de avaliação é o *mean squared error* (MSE, em português erro quadrático médio), já para problemas de classificação é utilizado o critério de Gini com o objetivo de minimizar a impureza para alcançar subgrupos homogêneos dos dados (WOZNICKI et al., 2019). A árvore de decisão cresce até onde o parâmetro de decisão de término é decidido pelo usuário (*nodesize*). A árvore de decisão que recebeu o menor valor de MSE ou que recebeu mais votos na classificação é a resposta final do algoritmo (BELGIU; DRĂGU, 2016).

Embora o algoritmo tenha muitos benefícios, e se ajuste com grandes bancos de dados e muitos atributos,o RF apresenta pouca precisão para conjuntos de dados complexos (por exemplo, conjuntos de dados grandes e conjuntos de dados com interações variáveis complexas) (SPEISER et al., 2019), na maioria das vezes, o conjunto de dados inclui muitos recursos com diferentes qualidades que podem influenciar o desempenho dos classificadores ou da regressão (KUMAR; SHAIKH, 2017). Para driblar esse problema é importante selecionar o melhor conjunto de atributos que melhore o desempenho do modelo, aumente a eficiência computacional e diminua os requisitos de armazenamento (KUMAR; SHAIKH, 2017). Métodos de *feature selection* ajudam a reduzir a complexidade e incompreensibilidade dos resultados, e a reduzir a dimensionalidade dos conjuntos de dados (GHAEMI; FEIZI-DERAKHSHI, 2016). Deste modo, uma série de métodos veem sendo utilizados para reduzir esta dimensionalidade para modelos de RF, dentre os mais comuns, métodos recursivos de redução de variáveis (ABDOH; ABO RIZKA; MAGHRABY, 2018), PCA (GEETHA et al., 2019), ou na pior das hipóteses, via tentativas manuais de seleção das melhores. E mais recentemente a hibridização entre algoritmo genético e *Random Forest* (GARF), que veem trazendo ótimos resultados na seleção de variáveis e melhorando a capacidade preditiva e de classificação do RF (ALIČKOVIĆ; SUBASI, 2017; CERRADA et al., 2016; HONG et al., 2018; NAGHIBI; AHMADI; DANESHI, 2017; PAUL et al., 2017).

2.3 *Data mining*

A partir do avanço computacional e tecnológico, bem como a facilidade de obtenção de dados, viu-se a necessidade de busca por informações ocultas dentro dos volumosos bancos de informações, com isso surgiram técnicas conhecidas como *data mining* (mineração de dados). Ngai et al. (2011) atribuem ao *data mining* uma importante ferramenta para descobrir as verdades escondidas por trás dos conjuntos de informações. Assim, o *data mining* é uma técnica de descoberta de informações incorporadas em grandes bases de dados, ou conjuntos de dados, é uma análise matemática que identifica padrões e tendências dos atributos e descobre relações complexas para prever resultados. Martins et al. (2018) definem a mineração de dados como uma estratégia de extração de conhecimento, que pode ser expressa em termos de padrões ou regras, características importantes do conjunto de dados em questão, portanto, a ferramenta fornece um meio para entender melhor os recursos implícitos nos dados brutos, o que é fundamental em um processo de tomada de decisão.

A mineração de dados transforma dados em conhecimento útil, e pode ser definida como o processo de selecionar, explorar e modelar grandes quantidades de dados para descobrir padrões previamente desconhecidos. O seu estudo está inserido nas raízes da análise estatística tradicional, assim como nas ciências da inteligência artificial/aprendizado de máquina, com o objetivo de se beneficiar de ambas (MORO; RITA; VALA, 2016). Existem diferentes métodos de *data mining*, alguns dos principais são: Redes Neurais artificiais, *Random Forest*, Algoritmos Genéticos (AGs) e a própria estatística clássica.

As Redes Neurais Artificiais (RNAs) são técnicas de inteligência computacional que simulam o comportamento de uma rede neural biológica, reproduzindo algumas de suas funções. As RNAs são instrumentos adequados na mineração de dados, amplamente utilizada por se adequar na construção de modelos não-lineares e complexos (ESMAEILY et al., 2018). São formadas por um número finito de camadas com diferentes elementos computacionais chamados neurônios (ASILTÜRK; ÇUNKAŞ, 2011). Tais neurônios, interligados em uma ou mais camadas, podem possuir muitas conexões, aos quais são associadas à pesos, esses pesos passam por processo de aprendizagem e armazenam os melhores resultados com a passar das iterações (GALVÃO; DE FÁTIMA MARIN, 2009). Um método robusto que busca e aprende padrões através de repetições dos dados, e atinge o ótimo global através da iteratividade.

Outra técnica disseminada para mineração de dados é o *Random Forest* (RF). O algoritmo consiste em muitas árvores de classificação individuais onde cada árvore é um

classificador que recebe um certo peso para sua classificação. Durante a construção do modelo, a representação é dada por três argumentos, nós internos, nós externos e ramificações. Os nós internos estão sempre vinculados às funções de decisão para decidir qual nó será dividido, caso o ganho de informação seja positivo, o nó se divide, caso contrário, o nó não se dividirá (FRAIWAN et al., 2012; QI et al., 2018). Os nós de saída são nós que não precisam mais ser divididos, e são conhecidos como terminais ou nós folhas. Para problemas de mineração, classificação, dentre outros problemas, um rótulo de classe será atribuído a cada nó externo para classificar amostras que se enquadram nesse nó, e as ramificações são usadas para conectar nós internos e nós externos (QI et al., 2018). As saídas de classificação de todas as árvores são usadas para determinar a saída de classificação geral, que é feita escolhendo o modo (a saída com mais votos) de todos os resultados de classificação de árvores (FRAIWAN et al., 2012). O *Random Forest* têm mostrado um excelente desempenho em vários problemas práticos e estão entre os métodos de regressão de uso geral mais precisos disponíveis (BIAU; DEVROYE, 2010).

Usualmente, na mineração de dados se aplica muito o uso de meta-heurísticas com princípios na computação evolucionária, dentre elas, uma das mais usuais está Algoritmo Genético (AG). Ele é baseado em fundamentos da evolução natural da teoria de Charles Darwin e da genética natural (LI et al., 2019). O AG é bastante utilizado para resolver problemas complexos de otimização, classificação e tendências de informações. Um AG simples possui ao menos operações genéticas básicas: seleção, cruzamento e mutação. Na seleção, algumas soluções da população são selecionadas como pais; no cruzamento, os pais são cruzados para produzir descendentes; e em mutação, a prole pode ser alterada de acordo com as regras de mutação. A meta-heurística apresenta uma forte robustez, adaptabilidade e paralelismo implícito, ele pode executar de maneira rápida e eficaz a otimização global, que desempenha um papel importante na tecnologia de mineração de dados (LI et al., 2019).

Uma quarta via diz respeito a estatística clássica, uma ferramenta que aplica modelos para análise e interpretação de dados. Das mais usuais em estatística aplicadas na mineração de dados está a regressão logística (RL), de natureza binária, método desenvolvido na década de 60. A regressão logística multinomial (MLR) é uma alternativa à regressão logística binomial (HOX; MOERBEEK; VAN DE SCHOOT, 2017; RÉMY; MARTIAL; CLÉMENTIN, 2018). A MLR tem como vantagem a relação não linear entre a variável dependente e cada variável independente e por ser usada em situações em que não há ordenação dos valores *K* das variáveis dependentes (RÉMY; MARTIAL; CLÉMENTIN, 2018).

Há diferentes métodos para o descobrimento de informações ocultas, todas com uma série de vantagens e desvantagens, portanto não é possível afirmar qual é o melhor. Por outro lado, dentro da área florestal a temática é relativamente nova, embora hoje em dia possa-se encontrar muitos trabalhos na área, ocasionado principalmente ao avanço da tecnologia e facilidade na obtenção das variáveis. Estudos com essa temática no setor florestal podem ser encontrados em trabalhos de Arpaci et al. (2014), Margono et al. (2014), Sanquetta et al. (2015), Pinheiro et al. (2016), Pourtaghi et al. (2016), Pourghasemi et al. (2017), Carvalho et al. (2017) e Hong et al. (2018).

O *data mining* demonstra ser uma importante ferramenta na tomada de decisões e na seleção de caracteres, e a sua capacidade de predizer bons resultados com boa ou ótima qualidade, principalmente em bancos de dados volumosos e com grande quantidade de variáveis trazem uma melhoria de técnicas difundidas da literatura, trazendo uma série de vantagens para as áreas às quais são aplicadas (HONG et al., 2018; REIS et al., 2019; SILVEIRA et al., 2019).

2.4 Seleção de variáveis

A mineração de dados é uma importante etapa na descoberta de conhecimentos, mas o elevado número de variáveis em um banco de dados se torna um problema. Muitos destes atributos são totalmente irrelevantes ou redundantes, desta forma, uma metodologia de *selection feature* (seleção de recursos ou seleção de variáveis), ajuda a reduzir a complexidade e incompreensibilidade dos resultados, e a reduzir a dimensionalidade dos conjuntos de dados antes da mineração de dados (GHAEMI; FEIZI-DERAKHSHI, 2016).

A seleção de atributos se dá de inúmeras formas, desde a aplicação de técnicas de combinação de variáveis clássicas a métodos de inteligência. Abordagens clássicas, como *stepwise*, adota como critério de seleção a estatística F, mas também pode ser feito com o coeficiente de correlação linear múltipla, erro quadrático total e critério de informação de Akaike (ALVES; LOTUFO; LOPES, 2014). O método *stepwise* é feito de forma iterativa, adicionando (passo *forward*) e removendo variáveis (passo *backward*), a partir de um critério de seleção (ALVES; LOTUFO; LOPES, 2014). Trabalhos com essa metodologia, podem ser encontrados com facilidade na área florestal, um exemplo pode ser visto no estudo de Orellana e Figueiredo Filho (2017), onde selecionaram variáveis através do procedimento para gerar modelos de predição de parâmetros para projetar a distribuição diamétrica em florestas nativas com a função Weibull.

Partindo para métodos mais atuais, o uso da regressão PLS (*Partial Least Squares*) vem sendo amplamente utilizada em procedimentos de seleção de variáveis por sua capacidade de operar com grande número de variáveis correlacionadas e afetadas por ruído (ZIMMER; ANZANELLO, 2014). A PLS seleciona variáveis através da correlação, ela atribui importância a cada descritor com base em vetores informativos (vetor de correlação, vetor de regressão, e produtos entre ambos) e reduz os preditores a um conjunto menor de componentes não correlacionados (BIRCK; CAMPOS; DE MELO, 2016). O PLS produz um pequeno número de combinações lineares dos recursos originais e as utiliza como entrada para o modelo de regressão, porque se supõe que um pequeno número de recursos latentes (ocultos) sejam responsáveis pela maior parte da variação no conjunto de dados (SCOTTI et al., 2016).

Outro método bastante tradicional e conhecido de seleção de variáveis é o LASSO (Operador de retração e seleção absolutos mínimos), ele aborda a colinearidade, reduzindo os coeficientes de preditores correlacionados em direção a zero (KIM et al., 2016). Foi proposto por Tibshirani (1996) e conforme Wang et al. (2018) é um método inovador de seleção de variáveis para regressão, minimizando a soma residual dos quadrados sujeitos à soma do valor absoluto dos coeficientes menor que uma constante e é um método de regressão esparsa que regulariza o parâmetro sob suposição esparsa. Foi originalmente introduzido no contexto de mínimos quadrados.

Partindo para o uso de inteligência computacional na seleção de recursos, existem inúmeras meta-heurísticas e heurísticas que fazem esse papel. Os algoritmos evolutivos baseados em populações, como algoritmos genéticos (AG), enxames de partículas (PSO) e otimização de colônias de formigas têm sido utilizados com grande sucesso, uma vez que são capazes de encontrar soluções adequadas sem explorar todo o espaço de busca (VIEIRA et al., 2012; HONG et al., 2018). A seleção de variáveis deixa mais assertiva a resposta, diminui a complexidade de possuir um grande banco de dados e a evitar ruído derivados da complexidade das iterações existentes.

3 CONSIDERAÇÕES FINAIS

Sobre o aspecto preditivo de atributos florestais, a estimativa da altura é umas das mais importantes no setor florestal para um bom planejamento, dada sua importância para a quantificação de volume, carbono, dentre outras. Para isso, a inclusão de dados de sensoriamento remoto, de gradientes geográficos e de solo são uma importante etapa na modelagem do atributo, aumentam a capacidade preditiva, além de dá um poder explicativo a variável resposta. Para isso, a implementação de técnicas de seleção de variáveis mais eficazes e robustas são necessárias. O algoritmo genético na seleção de variáveis permite varrer um espaço de busca maior, e também de associar variáveis que dão boas respostas e que provavelmente seriam eliminadas por métodos tradicionais de seleção. O método permite expandir o conhecimento dos dados, atribuindo aspectos importantes na classificação e predição da altura.

REFERÊNCIAS

ABDOH, S. F.; ABO RIZKA, M.; MAGHRABY, F. A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. **IEEE Access**, Piscataway, v. 6, p. 59475–59485, 2018.

ABDULHAMED, A. A.; TAWFEEK, M. A.; KESHK, A. E. A genetic algorithm for service flow management with budget constraint in heterogeneous computing. **Future Computing and Informatics Journal**, Egypt, v. 3, n. 2, p. 341–347, 2018.

ALEXANDER, C.; KORSTJENS, A. H.; HILL, R. A. Int J Appl Earth Obs Geoinformation In fl uence of micro-topography and crown characteristics on tree height estimations in tropical forests based on LiDAR canopy height models. **Int J Appl Earth Obs Geoinformation**, Enschede, v. 65, p. 105–113, 2018.

ALIČKOVIĆ, E.; SUBASI, A. Breast cancer diagnosis using GA feature selection and Rotation Forest. **Neural Computing and Applications**, London, v. 28, n. 4, p. 753–763, 2017.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Applied and Computational Mathematics**, São Carlos, v. 1, p. 1–6, 2014.

ARPACI, A. et al. Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. **Applied Geography**, Sevenoaks, v. 53, p. 258–270, 2014.

ASILTÜRK, I.; ÇUNKAŞ, M. Modeling and prediction of surface roughness in turning operations using artificial neural network and multiple regression method. **Expert Systems with Applications**, Oxford, v. 38, n. 5, p. 5826–5832, 2011.

BALIEIRO, A. et al. A multi-objective genetic optimization for spectrum sensing in cognitive radio. **Expert Systems with Applications**, Oxford, v. 41, n. 8, p. 3640–3650, 2014.

BARBOZA, A. O. et al. Pesquisa bibliométrica em estratégia como prática: resultados exploratórios e comparação de fontes. **Revista Eletrônica Sistemas & Gestão**, São Domingos, v. 10, n. 4, p. 561–574, 2015.

BELGIU, M.; DRĂGU, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 114, p. 24–31, 2016.

BIAU, G.; DEVROYE, L. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. **Journal of Multivariate Analysis**, New York, v. 101, n. 10, p. 2499–2518, 2010.

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

BIRCK, M. G.; CAMPOS, L. J.; DE MELO, E. B. Estudo computacional de 1H-imidazol-2-il-pirimidina-4,6-diaminas para a identificação de potenciais precursores de novos agentes antimaláricos. **Quimica Nova**, São Paulo, v. 15, p. 1–8, 2016.

BRAGA, R. et al. Ferramentas para desenvolvimento de sistemas baseados em Inteligência computacional - um mapeamento sistemático. In: XII Simpósio Brasileiro de Automação Inteligente (SBAI), 2015, Natal. **Anais...** Natal: SBAI, 2015, p. 384.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.

CAMACHO, D. M. et al. Next-Generation Machine Learning for Biological Networks. **Cell**, Cambridge, v. 173, n. 7, p. 1581–1592, 2018.

CAMPOS, B. P. F. et al. Predição da altura total de árvores em plantios de diferentes espécies por meio de redes neurais artificiais. **Pesquisa Florestal Brasileira**, Colombo, v. 36, n. 88, p. 375, 2017.

CARVALHO, M. C. et al. Modelagem do nicho ecológicos de espécies arbóreas em uma área tropical brasileira. **Cerne**, Lavras, v. 23, n. 2, p. 229–240, 2017.

CERRADA, M. et al. Fault diagnosis in spur gears based on genetic algorithm and random forest. **Mechanical Systems and Signal Processing**, New York v. 70–71, p. 87–103, 2016.

CHEN, S. et al. Genetic algorithm-based optimization for cognitive radio networks. **33rd IEEE Sarnoff Symposium 2010, Conference Proceedings**, Piscataway, 2010, p. 1-6.

CHUANG, Y. C.; CHEN, C. T.; HWANG, C. A real-coded genetic algorithm with a direction-based crossover operator. **Information Sciences**, London, v. 305, n. 1, p. 320–348, 2015.

COUTINHO, E. R.; SILVA, R. M.; DELGADO, A. R. S. Using computational intelligence technique for the meteorological data prediction | Utilização de técnicas de inteligência computacional na predição de dados meteorológicos. **Revista Brasileira de Meteorologia**, Rio de Janeiro, v. 31, n. 1, p. 24–36, 2016.

CURTO, R. D. A. et al. Relações hipsométricas em floresta estacional semidecidual. **Revista de Ciências Agrarias**, Recife, v. 57, n. 1, p. 57–66, 2014.

DAS, A. K. et al. Prediction of fine particulate matter chemical components with a spatio-temporal model for the Multi-Ethnic Study of Atherosclerosis cohort. **Applied Energy**, New York, v. 26, n. 5, p. 499–523, 2018.

ESMAEILY, H. et al. Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes. **Iranian Biomedical Journal**, Tehran, v. 22, n. 5, p. 303–311, 2018.

FACELLI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2015.

FRAIWAN, L. et al. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. **Computer Methods and Programs in Biomedicine**, Amsterdam, v. 108, n. 1, p. 10–19, 2012.

GALVÃO, N. D.; DE FÁTIMA MARIN, H. Técnica de mineração de dados: Uma revisão da literatura. **ACTA Paulista de Enfermagem**, São Paulo, v. 22, n. 5, p. 686–690, 2009.

GARCÍA, M. et al. Int J Appl Earth Obs Geoinformation Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery. **Int J Appl Earth Obs Geoinformation**, Enschede, v. 66, n. December 2017, p. 159–173, 2018.

GARG, H. A hybrid GSA-GA algorithm for constrained optimization problems. **Information Sciences**, New York, v. 478, p. 292–305, 2016.

GEETHA, R. et al. Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. **Journal of Medical Systems**, Dordrecht, v. 43, n. 9, 2019.

GHAEMI, M.; FEIZI-DERAKHSHI, M. R. Feature selection using Forest Optimization Algorithm. **Pattern Recognition**, Oxford, v. 60, p. 121–129, 2016.

GHAMISI, P.; BENEDIKTSSON, J. A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. **IEEE Geoscience and Remote Sensing Letters**, New York, v. 12, n. 2, p. 309–313, 2015.

GOMIDE, L. R. **Planejamento florestal espacial**. 2009. 256p. Tese (Doutorado em Engenharia Florestal) - Universidade Federal do Paraná, 2009.

GONÇALVES E GONÇALVES, W. et al. Classificação de estratos florestais utilizando redes neurais artificiais e dados de sensoriamento remoto. **Ambiente & Água - An Interdisciplinary Journal of Applied Science**, Taubaté, v. 11, n. 3, 2016.

HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, New York, v. 630, p. 1044–1056, 2018.

HOX, J. J.; MOERBEEK, M.; VAN DE SCHOOT, R. **Multilevel Analysis**. Third edition. | New York, NY: Routledge, 2017.

KALSI, S.; KAUR, H.; CHANG, V. DNA Cryptography and Deep Learning using Genetic Algorithm with NW algorithm for Key Generation. **Journal of Medical Systems**, Dordrecht, v. 42, n. 1, 2018.

KAO, Y. T.; ZAHARA, E. A hybrid genetic algorithm and particle swarm optimization for multimodal functions. **Applied Soft Computing Journal**, Londres, v. 8, n. 2, p. 849–857, 2008.

KIM, S. Y. et al. Prediction of fine particulate matter chemical components with a spatio-temporal model for the Multi-Ethnic Study of Atherosclerosis cohort. **Journal of Exposure Science and Environmental Epidemiology**, New York, v. 26, n. 5, p. 520–528, 2016.

KONAR, A. **Computational Intelligence: Principles, Techniques and Applications**. New York: Springer-Verlag Berlin Heidelberg, 2005.

KUMAR, S. S.; SHAIKH, T. Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest. In: **2017 International Conference on Computer and Applications (ICCA).** Doha:IEEE, 2017. p. 227-231.

LEE, J. et al. Machine learning approaches for estimating forest stand height using plot-based observations and Airborne LiDAR data. **Forests**, Basel, v. 9, n. 5, 2018.

LI, W. et al. Data mining optimization model for financial management information system based on improved genetic algorithm. **Information Systems and e-Business Management**, Heidelberg, n. 0123456789, 2019.

LIDBERG, W.; NILSSON, M.; ÅGREN, A. Using machine learning to generate high-resolution wet area maps for planning forest management: A study in a boreal forest landscape. **Ambio**, Stockholm, p. 1-12, 2019.

LINDEN, R. **Algoritmos Genéticos**. 3ª Edição, Rio de Janeiro: Ciência Moderna, 2012.

LOPATIN, J. et al. Comparing Generalized Linear Models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. **Remote Sensing of Environment**, New York, v. 173, p. 200–210, 2016.MACHADO, S. DO A. et al. Comportamento da relação hipsométrica de Araucaria angustifolia no capão da Engenharia Florestal da UFPR. **Pesquisa Florestal Brasileira**, Colombo, n. 56, p. 5–16, 2008.

MARGONO, B. A. et al. Primary forest cover loss in indonesia over 2000-2012. **Nature Climate Change**, London, v. 4, n. 8, p. 730–735, 2014.

MARTINS, D. et al. Making a state-of-the-art heuristic faster with data mining. **Annals of Operations Research**, Basel, v. 263, n. 1–2, p. 141–162, 2018.

MICHALEWICZ, Z. **Genetic Algorithms + Data Structures = Evolution Programs**. 3. ed. New York: Springer-Verlag Berlin Heidelberg, 1996.

MITCHELL, M. **An Introduction to Genetic Algorithms**. 1. ed. Santa Fé: MIT Press Cambridge, MA, 1996.

MORO, S.; RITA, P.; VALA, B. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. **Journal of Business Research**, New York, v. 69, n. 9, p. 3341–3351, 2016.

NAGHIBI, S. A.; AHMADI, K.; DANESHI, A. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. **Water Resources Management**, Dordrecht, v. 31, n. 9, p. 2761–2775, 2017.

NAGHIBI, S. A.; POURGHASEMI, H. R.; DIXON, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. **Environmental Monitoring and Assessment**, Dordrecht, v. 188, n. 1, p. 1–27, 2016.

NEMATI, M.; BRAUN, M.; TENBOHLEN, S. Optimization of unit commitment and economic dispatch in microgrids based on genetic algorithm and mixed integer linear programming. **Applied Energy**, New York, v. 210, p. 944–963, 2018.

NGAI, E. W. T. et al. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision Support Systems**, Amsterdam, v. 50, n. 3, p. 559–569, 2011.

NGAI, E. W. T.; XIU, L.; CHAU, D. C. K. Application of data mining techniques in customer relationship management: A literature review and classification. **Expert Systems with Applications**, Oxford, v. 36, n. 2 PART 2, p. 2592–2602, 2009.

O'BRIEN, R.; ISHWARAN, H. A random forests quantile classifier for class imbalanced data. **Pattern Recognition**, Oxford, v. 90, p. 232–249, 2019.

ORELLANA, E.; FIGUEIREDO FILHO, A. Uso do método da predição de parâmetros para projetar a distribuição diamétrica em florestas nativas com a função Weibull. Ciência Florestal, Santa Maria, v. 27, n. 3, p. 981–991, 2017.

PAUL, D. et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. **Computerized Medical Imaging and Graphics**, Amsterdam, v. 60, p. 42–49, 2017.

PEREIRA, A. M.; ALVES, A. C. B.; MARINHO, R. P. Estratificação de solo multicamada através da função kernel e do algoritmo genético. **Revista de Engenharia e Tecnologia**, Ponta Grossa, v. 10, n. 1, p. 229–247, 2018.

PEZZELLA, F.; MORGANTI, G.; CIASCHETTI, G. A genetic algorithm for the Flexible Job-shop Scheduling Problem. **Computers & Operations Research**, Oxford, v. 35, n. 10, p. 3202–3212, 2008.

PINHEIRO, T. F. et al. Forest degradation associated with logging frontier expansion in the Amazon: The BR-163 region in southwestern pará, Brazil. **Earth Interactions**, Boston, v. 20, n. 17, 2016.

POURGHASEMI, H. R. et al. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. **Science of the Total Environment**, New York, v. 609, p. 764–775, 2017.

POURRAHMATI, M. R. et al. Mapping lorey's height over Hyrcanian forests of Iran using synergy of ICESat/GLAS and optical images. **European Journal of Remote Sensing**, Florence, v. 51, n. 1, p. 100–115, 2018.

POURTAGHI, Z. S. et al. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. **Ecological Indicators**, New York, v. 64, p. 72–84, 2016.

QI, C. et al. Prediction of open stope hangingwall stability using random forests. **Natural Hazards**, Dordrecht, v. 92, n. 2, p. 1179–1197, 2018.

RAHMAN, I.; MOHAMAD-SALEH, J. Hybrid bio-Inspired computational intelligence techniques for solving power system optimization problems: A comprehensive survey. **Applied Soft Computing Journal**, Londres, v. 69, p. 72–130, 2018.

REIS, A. A. et al. Volume estimation in a Eucalyptus plantation using multi-source remote sensing and digital terrain data: a case study in Minas Gerais State, Brazil. **International Journal of Remote Sensing**, London, v. 40, n. 7, p. 2683–2702, 2019.

RÉMY, N. M.; MARTIAL, T. T.; CLÉMENTIN, T. D. The prediction of good physicians for prospective diagnosis using data mining. **Informatics in Medicine Unlocked**, London, v. 12, p. 120–127, 2018.

RODRIGUES, F. L. et al. Metaheurística algoritmo genético para solução de problemas de planejamento florestal com restrições de integridade. **Revista Árvore**, Viçosa, v. 28, n. 2, p. 233–245, 2004.

SANQUETTA, C. R. et al. Comparison of data mining and allometric model in estimation of tree biomass. **BMC Bioinformatics**, London, v. 16, n. 1, p. 1–9, 2015.

SCHNEIDER, P. R.; SCHNEIDER, P. S. P.; SOUZA, C. A. M. **Análise de Regressão aplicada à Engenharia Florestal**. 2. Edição ed. Santa Maria: FACOS, 2009.

SCOLFORO, J. R. S. **Biometria florestal:** modelos de crescimento e produção florestal. Lavras: UFLA/FAEPE, 2006.

SCOTTI, M. T. et al. Variable-selection approaches to generate QSAR models for a set of antichagasic semicarbazones and analogues. **Chemometrics and Intelligent Laboratory Systems**, New York, v. 154, p. 137–149, 2016.

SILVA, G. F. DA et al. Avaliação de métodos de medição de altura em florestas naturais. **Revista Arvore**, Viçosa, v. 36, n. 2, p. 341–348, 2012.

SILVEIRA, E. M. O. et al. Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 78, p. 175–188, 2019.

SPEISER, J. L. et al. A comparison of random forest variable selection methods for classification prediction modeling. **Expert Systems with Applications**, Oxford, v. 134, p. 93–101, 2019.

SULLIVAN, F. B. et al. Forest Ecology and Management Comparison of lidar- and allometry-derived canopy height models in an eastern deciduous forest. **Forest Ecology and Management**, New York v. 406, p. 83–94, 2017.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Oxford, v. 58, n. 1, p. 267–288, 1996.

TUAN, N. T.; DINH, T. T.; LONG, S. H. Height-diameter relationship for Pinus koraiensis in Mengjiagang Forest Farm of Northeast China using nonlinear regressions and artificial neural network models. **Journal of Forest Science**, Prague, v. 65, n. 4, p. 134–143, 2019.

VENDRUSCOLO, D. G. S. et al. Estimativa da altura de eucalipto por meio de regressão não linear e redes neurais artificiais. Revista Brasileira de Biometria, Lavras,. v. 33, n. 4, p. 555–569, 2015.

VIEIRA, S. M. et al. **Metaheuristics for feature selection: application to sepsis outcome prediction**. Evolutionary Computation (CEC), 2012 IEEE Congress on. **Anais**...Brisbane: 2012

WAGER, S.; ATHEY, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. **Journal of the American Statistical Association**, New York, v. 113, n. 523, p. 1228–1242, 2018.

WANG, S. et al. Predicting ship fuel consumption based on LASSO regression. **Transportation Research Part D: Transport and Environment**, New York, v. 65, p. 817–824, 2018.

WOZNICKI, S. A. et al. Development of a spatially complete floodplain map of the conterminous United States using random forest. **Science of the Total Environment**, New York, v. 647, p. 942–953, 2019.

YANG, H. et al. Optimal sizing method for stand-alone hybrid solar-wind system with LPSP technology by using genetic algorithm. **Solar Energy**, Amsterdam, v. 82, n. 4, p. 354–367, 2008.

YOUSSEF, A. M. et al. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. **Landslides**, Heidelberg, v. 13, n. 5, p. 839–856, 2016.

YU, F.; XU, X. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. **Applied Energy**, Amsterdam, v. 134, p. 102–113, 2014.

ZIMMER, J.; ANZANELLO, M. J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Production**, Rio de Janeiro, v. 24, n. 1, p. 84–93, 2014.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - MINERAÇÃO DE DADOS E SELEÇÃO DE VARIÁVEIS PARA MODELOS HIPSOMETRICOS NÃO LINEARES

DATA MINING AND FEAUTURE SELECTION FOR NONLINEAR HYPSOMETRIC MODELS

Artigo formatado conforme a NBR 6022 (ABNT, 2003) e adaptado as exigências do Manual de Normalização de Trabalhos Acadêmicos da UFLA.

RESUMO

A mineração de dados é um método útil no contexto de manipulação da informação em grandes bancos de dados, sendo aplicada a vários contextos científicos. A extração de variáveis para explicar um dado comportamento é sua principal razão. Assim, o objetivo do trabalho partiu do pressuposto de selecionar variáveis explicativas para estimar a altura das árvores com modelos de regressão com o auxílio do algoritmo genético (AG). Nesse contexto, um conjunto de 5608 árvores e 150 variáveis (povoamento, sensores remotos, ambientais) foram utilizados para a construção de modelos hipsométricos não lineares, considerando o uso do AG. Um total de três modelos não lineares foram utilizados como base para a inclusão de variáveis. Estes podiam variar de uma a três entradas, sendo possível ainda uma combinação simples de segunda ordem. Adotou-se ainda a possibilidade de utilização de operadores aritméticos para a formação de uma nova variável. Nesse sentido, aplicou-se uma normalização para controlar o efeito de escala. As variáveis foram divididas em três bancos de dados, com valores proporcionais de cada classe para verificar a influência do conjunto de dados na seleção dos mesmos. Foram realizados um total de 30 repetições para cada conjunto de dados e modelos, gerando 270 modelos. Para averiguar o desempenho da técnica, foram utilizados histogramas das variáveis selecionadas e análise do RMSE %. Ao final, apenas os 2 melhores modelos gerados foram selecionados e comparados com seus respectivos clássicos, com análises dos erros (RMSE, RMSE% e R^2) e análises gráficas. Os resultados mostraram que o método de seleção via algoritmo genético é capaz de apresentar as melhores variáveis aos modelos. Contudo, com o aumento no número de entradas no modelo, maior é a complexidade na obtenção de boas soluções. Este fato também está associado à forma matemática de cada modelo. Em geral, o diâmetro das árvores e variáveis de sensores foram as mais selecionadas, seguidas pelas ambientais. Os modelos otimizados apresentaram resultados similares aos encontrados nos originais, porém com explicação ecológica. Conclui-se que o AG foi capaz de selecionar variáveis que otimizasse os modelos de regressão e auxiliassem na explicação ecológica da estimativa da altura.

Palavras-chave: *Feauture selection*. Algoritmo genético. Manejo florestal.

ABSTRACT

Data mining is a useful method in the context of manipulating information in large databases and is applied to various scientific contexts. The extraction of variables to explain a given behavior is its main reason. Thus, the objective of the study was based on the assumption of selecting explanatory variables to estimate tree height with regression models with the aid of the genetic algorithm (GA). In this context, a set of 5608 trees and 150 variables (stand, remote sensors, environmental) were used for the construction of nonlinear hypsometric models, considering the use of GA. A total of three nonlinear models were used as the basis for the inclusion of variables. These could range from one to three entries, and even a simple second order combination is possible. The possibility of using arithmetic operators for the formation of a new variable was also adopted. In this sense, a normalization was applied to control the scale effect. The variables were divided into three databases, with proportional values of each class to verify the influence of the data set on their selection. A total of 30 repetitions were performed for each data set and models, generating 270 models. To verify the performance of the technique, we used histograms of the selected variables and analysis of the RMSE%. In the end, only the 2 best models generated were selected and compared with their respective classics, with error analysis (RMSE, RMSE% and R^2) and graphical analysis. The results showed that the genetic algorithm selection method is able to present the best variables to the models. However, with the increase in the number of entries in the model, the greater the complexity in obtaining good solutions. This fact is also associated with the mathematical form of each model. In general, tree diameter and sensor variables were the most selected, followed by environmental ones. The optimized models presented similar results to those found in the originals, but with ecological explanation. It was concluded that the GA was able to select variables that optimized the regression models and assisted in the ecological explanation of the height estimate.

Keywords: Feature selection. Genetic Algorithm. Forest management.

1 INTRODUÇÃO

A altura das árvores é uma importante variável auxiliar no entendimento dos processos ecológicos e de crescimento de um povoamento florestal. Amplamente difundida nos modelos de crescimento e produtividade, sendo igualmente necessária em métodos de classificação da capacidade produtiva local (FIGUEIREDO FILHO et al., 2010; KANDARE et al., 2017; NOORDERMEER et al., 2018; PARRESOL et al., 2017; TÉO et al., 2017). A sua mensuração é onerosa e passível de erros, principalmente em povoamentos nativos, e por isso há estudos vinculados a modelos hipsométricos que a descreve (UZOH, 2017). Embora na literatura os modelos de diâmetro-altura sejam os mais comumente usados (SHARMA; BREIDENBACH, 2015), outros métodos via inteligência computacional vem sendo aplicados recentemente (AHMED et al., 2015; GARCÍA et al., 2018; MATASCI et al., 2018; STABEN; LUCIEER; SCARTH, 2018; ZHAO et al., 2019).

A modelagem de um atributo não é uma tarefa simples, frequentemente se faz necessário selecionar um conjunto de variáveis, as mais significativas para explicar um evento. Atualmente, com os avanços tecnológicos na obtenção e construção de um banco de dados formado por inúmeras variáveis, torna-se pouco oneroso ao pesquisador a sua estruturação (WANG; WANG; CHANG, 2016). A redução da dimensionalidade dos conjuntos de dados, por sua vez, buscando superar problemas de colinearidade é uma tarefa concernente a mineração de dados (GHAEMI; FEIZI-DERAKHSHI, 2016). Essa técnica é hábil na identificação e escolha de atributos para os modelos (HONG et al., 2018). Considerando exclusivamente modelos lineares existem métodos como *stepwise* (NORYANI et al., 2019), *reversible jump* (PAN, HSIEN e TSAI, 2018), *partial least squares* (FERREIRA et al., 2017), métodos híbridos como a combinação entre o algoritmo genético e *partial least squares* (MEHMOOD; AHMED, 2016; ZIMMER; ANZANELLO, 2014), lasso (WANG et al., 2018), PCA (GORGANNEJAD et al., 2019), entre outras. Já envolvendo modelos não lineares, têm-se metodologias de construção usando programação genética (KHANDELWAL et al., 2017) e *reversible jump* (GALAGALI; MARZOUK, 2019). Contudo, independente do caminho a seguir, há um grande desafio lançado na seleção dessas variáveis.

Os problemas de seleção de variáveis são complexos pela natureza combinatória de possibilidades. No presente estudo, cerca de 150 variáveis foram extraídas e estas poderiam ser combinadas entre si através de operadores matemáticos (multiplicação, divisão, adição e subtração), gerando alguns milhões de opções de modelos. Decorrente dos fatos, há uma integração entre esta temática e a inteligência computacional, sendo o uso de meta-heurísticas

um caminho natural e mais assertivo (HONG et al., 2018; VIEIRA et al., 2012). Portanto, essa abordagem metodológica foi proposta com base na implementação de um algoritmo genético, o qual consiste na seleção e combinação de variáveis para a construção de modelos não lineares, tornando-se assim uma opção auxiliar à modelagem. Desse modo, o objetivo do estudo foi a seleção e combinação entre variáveis populacionais, ambientais e de sensores, para modelos de regressão não lineares explicativos da relação da altura das árvores; avaliar o desempenho da meta-heurística na estruturação dos modelos de regressão e ampliação das variáveis disponíveis para a seleção; assim como comparar os modelos otimizados com os modelos hipsométricos clássicos, e; gerar novas opções de modelos auxiliares na explicação ecológica dos efeitos fitogeográficos, considerando a incorporação de variáveis ambientais.

2 MATERIAL E MÉTODOS

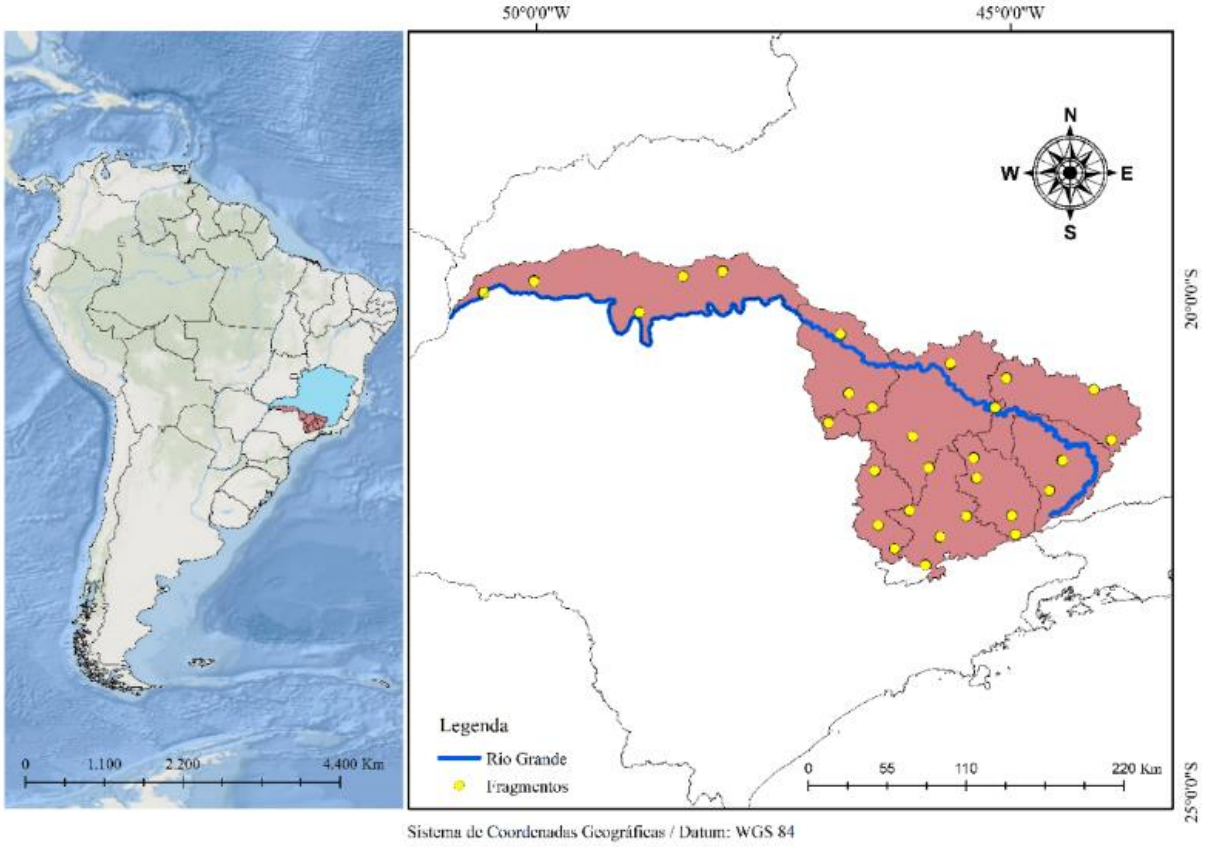
2.1 Área de estudo

A área de estudo está concentrada no Estado de Minas Gerais, na bacia hidrográfica do rio Grande, com área de 86.110 km², subdividida em 14 sub-bacias e abrange uma área total de aproximadamente 15% do território do estado (Figura 1). A bacia do rio Grande apresenta classificação climática de Köppen do tipo Cwa (verão quente), Cwb (verão temperado) e Aw (verão quente e chuvoso) (ALVARES et al., 2013). O tipo de solo predominante são os latossolos, com mosaicos de neossolo litólico, argissolo e cambissolo (CURI et al., 2008).

De acordo com a disponibilidade das áreas, em relação à composição da vegetação remanescente, recebem destaque as pertencentes ao domínio Mata Atlântica: Floresta Estacional Semidecidual; e ao domínio Cerrado: Campo, Campo cerrado, Cerradão e Cerrado *Sensu Stricto*. Essa cobertura vegetal nativa da bacia hidrográfica representa aproximadamente 16% do seu território (SCOLFORO et al., 2008).

Ao todo foram amostrados 29 fragmentos florestais nas diferentes fitofisionomias, variando ainda classes de clima e altitude, totalizando 1009 parcelas de 250m^2. As variáveis coletadas foram o diâmetro a altura do peito (DAP) e altura total (HT), bem como a identificação botânica para um total de 45.239 árvores. As árvores medidas possuíam um DAP superior a 5 cm e compuseram a classe de variáveis populacionais, como área basal e densidade (N/ha).

Figura 1 - Mapa de localização da bacia hidrográfica do rio Grande, com informações da altitude local e da distribuição das parcelas.



Fonte: Do autor (2020).

2.2 Variáveis ambientais

A obtenção das variáveis ambientais (Tabela 1) considerou distintas categorias, tais como: a) geográfica; b) solo, e; c) sensoriamento remoto. As variáveis que formam a classe geográfica foram obtidas das coordenadas geográficas latitude (Y) e longitude (X) do centroide das subparcelas. Já as informações que constituem a classe solo, são provenientes de amostras coletadas em apenas uma fração das parcelas (154 parcelas), sendo as demais descartadas das análises (855/1009 parcelas). Tais amostras foram obtidas em diferentes profundidades, sendo estas, 0 a 10 cm, 10 a 20 cm, 20 a 40 cm, 40 a 60 cm e 60 a 100 cm, caracterizando o solo em diferentes horizontes (“A”, “B”, “C”, “D” e “E”). Os valores de pH (acidez ou basicidade do solo), potássio, fósforo, cálcio, magnésio, alumínio, acidez potencial (H + Al), CTC (capacidade de troca de cátion) efetiva, CTC a pH 7, saturação por bases, saturação por alumínio, matéria orgânica, fósforo remanescente, zinco, ferro, manganês, cobre, boro,

enxofre, porcentual de argila, silte e areia foram retiradas diretamente da análise laboratorial dos solos.

A obtenção das variáveis ambientais que formam a classe sensoriamento remoto foram adquiridas de dados espectrais através do ambiente de sistemas de informações geográficas (SIG), a partir de imagens do satélite Landsat 8 OLI (30m de resolução) e MODIS (*Moderate Resolution Spectroradiometer*) (com resolução variável entre 250 a 1.000m), adquiridos dentro do intervalo de tempo do inventário florestal.

Foram extraídas do sensor Landsat 8 OLI 12 cenas para abranger toda a área do experimento, oriundas do *United States Geological Survey of Earth* (USGS/EROS), já com as devidas correções geométricas e radiométricas. Com a formação do mosaico foram calculados 7 índices de vegetação, utilizados para compor o banco de dados das variáveis independentes: NDVI - *Normalized Difference Vegetation Index* (ROUSE et al., 1973); NDMI - *Normalized difference moisture index* (WILSON; SADER, 2002); EVI - *Enhanced vegetation index* (JUSTICE et al., 1998); SAVI - *Soil-adjusted Vegetation Index* (HUETE, 1988); mSAVI - *Modified Soil-adjusted Vegetation Index* (QI et al., 1994); NBR - *Normalized Burn Ratio* (MILLER; THODE, 2007); NBR2 - *Normalized Burn Ratio 2* (MILLER; THODE, 2007). E a partir do modelo digital de elevação *Shuttle Radar Topography Mission* – SRTM, redimensionado para 100 m de resolução espacial, foram calculadas 16 variáveis morfométricas, utilizando a ferramenta *Terrain Analysis* do software SAGA GIS (v. 6.3.0). Do sensor MODIS foram extraídas variáveis relacionadas à temperatura da superfície da Terra (emis31, emis32, lstd, lstn), atividade fotossintética (fpar, lai), evapotranspiração (et, le, pet, ple), produtividade primária (gpp) e porcentagem de cobertura vegetal (treecover). Desse modo, foram obtidas variáveis amostradas de solos nos diferentes horizontes, variáveis geográficas, índices landsat e produtos modis num total de 147 atributos por parcela redimensionadas para hectare, que se completam com as 3 a nível de povoamento.

Tabela 1- Variáveis explicativas utilizadas na modelagem da altura das árvores.

(continua)

Classes		Variáveis	Sigla	Unidade
Geográficas		Latitude	Y	m
		Longitude	X	m
Solo*		Acidez ou basicidade do solo	pH	-
		Potássio	K	$mmol_c/dm^3$
		Fósforo	P	mg/dm^3
		Cálcio	Ca	$mmol_c/dm^3$
		Magnésio	Mg	$mmol_c/dm^3$
		Alumínio	Al	mg/dm^3
		Acidez potencial	H+Al	mg/dm^3
		CTC efetiva	t	mg/dm^3
		CTC a pH 7	T	mg/dm^3
		Saturação por Bases	V	mg/dm^3
		Saturação por Alumínio	m	mg/dm^3
		Matéria Orgânica	M.O.	mg/dm^3
		Fósforo Remanescente	P-Rem	mg/dm^3
		Zinco	Zn	mg/dm^3
		Ferro	Fe	mg/dm^3
		Manganês	Mn	mg/dm^3
		Cobre	Cu	mg/dm^3
		Boro	B	mg/dm^3
		Enxofre	S	mg/dm^3
		Relação da análise física (argila)	Argila	%
		Relação da análise física (silte)	Silte	%
		Relação da análise física (areia)	Areia	%
Sensoriamento Remoto	Morfométrica	Morfométrica	altitude	Categórica
		Analytical hillshading	analytical	Categórica
		Aspect	Aspect	Categórica
		Channel network base level	cn_base_le	Categórica
		Convergence index	conv_index	Categórica
		Cross sectional curvature	c_sec_curv	Categórica
		Diffuse insolation	dif_insol	Categórica
		Direct insolation	direct_ins	Categórica
		Flow accumulation	flow_accum	Categórica
		Longitudinal curvature	long_curv	Categórica
		LS factor	ls_factor	Categórica
		Relative slope	relative_s	Categórica
		Valley depth	valley_dep	Categórica
		Vertical distance	vert_dist	Categórica
		Wetness index	wet_index	Categórica
		Slope (%)	slope_perc	Categórica
	Espectral Landsat	Espectral Landsat	evi	Categórica
		Normalized Difference Vegetation Index	ndvi	Categórica
		Modified Soil-adjusted Vegetation Index	msavi	Categórica
		Normalized Burn Ratio	nbr	Categórica
		Normalized Burn Ratio 2	nbr2	Categórica
		Normalized difference moisture index	ndmi	Categórica
		Soil-adjusted Vegetation Index	savi	Categórica

Tabela 1 - Variáveis explicativas da altura das árvores a serem utilizadas na modelagem.

(conclusão)

Classes		Variáveis	Sigla	Unidade
Sensoriamento Remoto	Espectral MODIS	Emissivity bands 31	emis31	Categórica
		Emissivity bands 32	emis32	Categórica
		Global evapotranspiration	et	Categórica
		Fraction of photosynthetically active radiation	fpar	Categórica
		Latent heat flux	le	Categórica
		Land Surface Temperature day	lstd	Categórica
		Land Surface Temperature night	lstn	Categórica
		Potential global evapotranspiration	pet	Categórica
		Potential latent heat flux	ple	Categórica
		Percent Tree Cover	treecover	Categórica
		Gross Primary Production	gpp	Categórica
		Leaf Area Index	lai	Categórica

* Variáveis distribuídas nos horizontes “A”, “B”, “C”, “D” e “E” do solo.

2.3 Inclusão de variáveis nos modelos não lineares

Após a estruturação do banco de dados e a remoção de *outliers*, a mesma foi dividida em dois conjuntos: treino (80%) e validação (20%). A proposta foi garantir uma validação dos modelos e assim evitar efeitos aleatórios e pontuais na inclusão das variáveis, acarretando resultados enviesados. Assim, a fase de seleção de variáveis e sua inclusão nos modelos não lineares (Tabela 2) foi conduzida pelo uso do Algoritmo Genético. Nesse estágio, existia disponível um total de 150 variáveis de entrada, que foram combinadas entre si empregando operadores aritméticos (multiplicação, divisão, adição e subtração). Essa proposta permitiu ampliar o espaço amostral de novas possibilidades de variáveis aos modelos hipsométricos testados. Além disso, diante de tal complexidade do problema, adotou-se a estratégia de estruturar o banco de dados em três subdivisões progressivas, em que as variáveis foram divididas em conjuntos de 50, 100 e 150 variáveis, para avaliar o desempenho do método. A regra adotada foi estabelecida conforme sua natureza, sendo os atributos do povoamento e geográficos em todos os conjuntos; solo e sensores remotos adicionados de forma percentual aos conjuntos.

As variáveis numéricas foram normatizadas linearmente entre o intervalo de 0 e 1 com a finalidade de evitar a influência de alguma variável devido sua maior ou menor magnitude. A implementação do algoritmo genético foi estabelecida após testes de parametrização, envolvendo os operadores de seleção torneio, critério de parada (100 gerações), tamanho da população (100 indivíduos).

Tabela 2 - Modelos hipsométricos de povoamentos não lineares a serem testados.

Modelos	Fonte	Classes de entrada (ξ_i)	Fórmula
Logístico	Pearl e Reed (1920)	1	$Y_i = \frac{\theta_0}{\left(1 + \theta_1 exp\left(-\theta_2 \ \xi_{1i}\right)\right)} + \varepsilon_i$
Hoerl	Daniel e Wood (1980)	2	$Y_i = \theta_0 \theta_1^{\xi_{1i}} \xi_{2i}^{\ \theta_2} + \varepsilon_i$
Prodan	Prodan (1968)	3	$Y = \frac{\xi_{1i}}{\exp^{\theta_0 + \theta_1 \xi_{2i} + \theta_3 \xi_{3i}^2}} + \varepsilon_i$

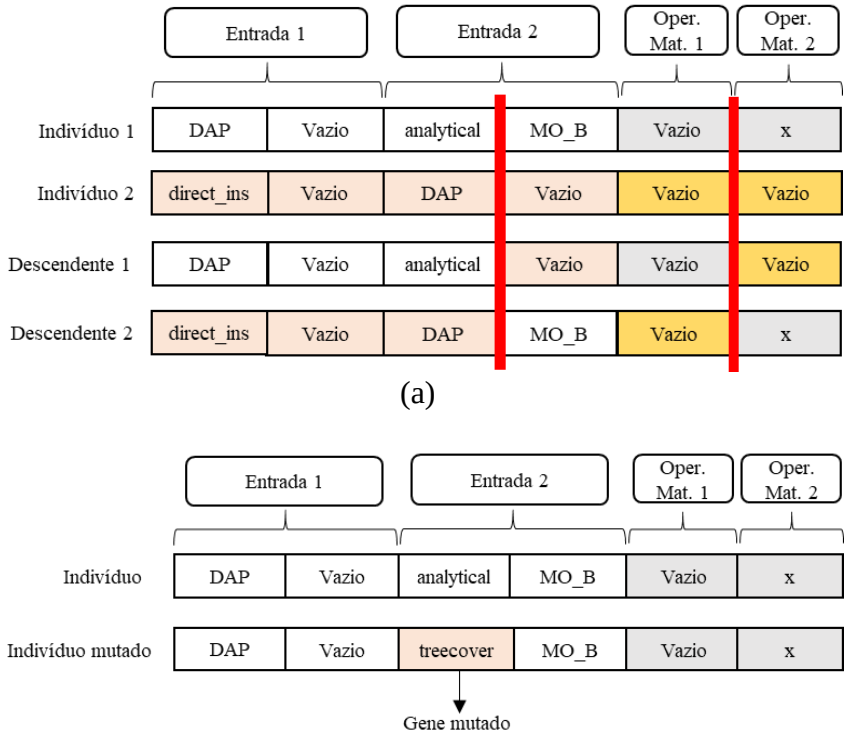
Y_i: altura (m) do indivíduo *i*; ξ_1, ξ_2 *e* ξ_3: combinação de variáveis 1, 2 e 3 *i* a serem inseridas; θ_0, θ_1, θ_2, e θ_3: parâmetros dos modelos; ε_i: erro aleatório da árvore *i*.

As variáveis numéricas foram normatizadas linearmente entre o intervalo de 0 e 1 com a finalidade de evitar a influência de alguma variável devido sua maior ou menor magnitude. A implementação do algoritmo genético foi estabelecida após testes de parametrização, envolvendo os operadores de seleção torneio, critério de parada (100 gerações), tamanho da população (100 indivíduos). Os operadores de *crossover* e mutação foram adaptados ao problema. No caso do *crossover*, adotou-se um operador de *crossover* específico com base no operador de um ponto de corte para criação dos descendentes. Em que, conforme a entrada de variáveis, utilizou-se um ponto de corte para os genes que contém as variáveis.

Posteriormente, de acordo com a posição de corte do cromossomo para as variáveis de entrada, pode ou não ter um segundo ponto de corte, o qual interfere na mudança do operador matemático, permitindo ter mais um ponto de corte específico para os genes do operador matemático para as combinações das variáveis (Figura 2a).

Para o operador genético de mutação, optou-se por utilizar a possibilidade de mudança de gene para as variáveis de entradas *i* dos genes do cromossomo, conferindo diversidade à população (Figura 2b). Diante tal contexto, de acordo com o gene escolhido para ser mutado caso seja uma célula vazia, ele implicará em uma segunda mutação. Esta mutação envolve o operador matemático referente a entrada em que o gene está locado no cromossomo. A taxa de cromossomos mutados será constante e uniforme para todas as gerações com probabilidade de 0,025.

Figura 2 - Estrutura dos operadores de *crossover* (a) e mutação (b) utilizados como esquema ilustrativo.



Em que: DAP: diâmetro a altura do peito; analytical: analítico; MO_B; matéria orgânica do horizonte B do solo; direct_ins: insolaração direta; x: multiplicação.

Fonte: Do autor (2020).

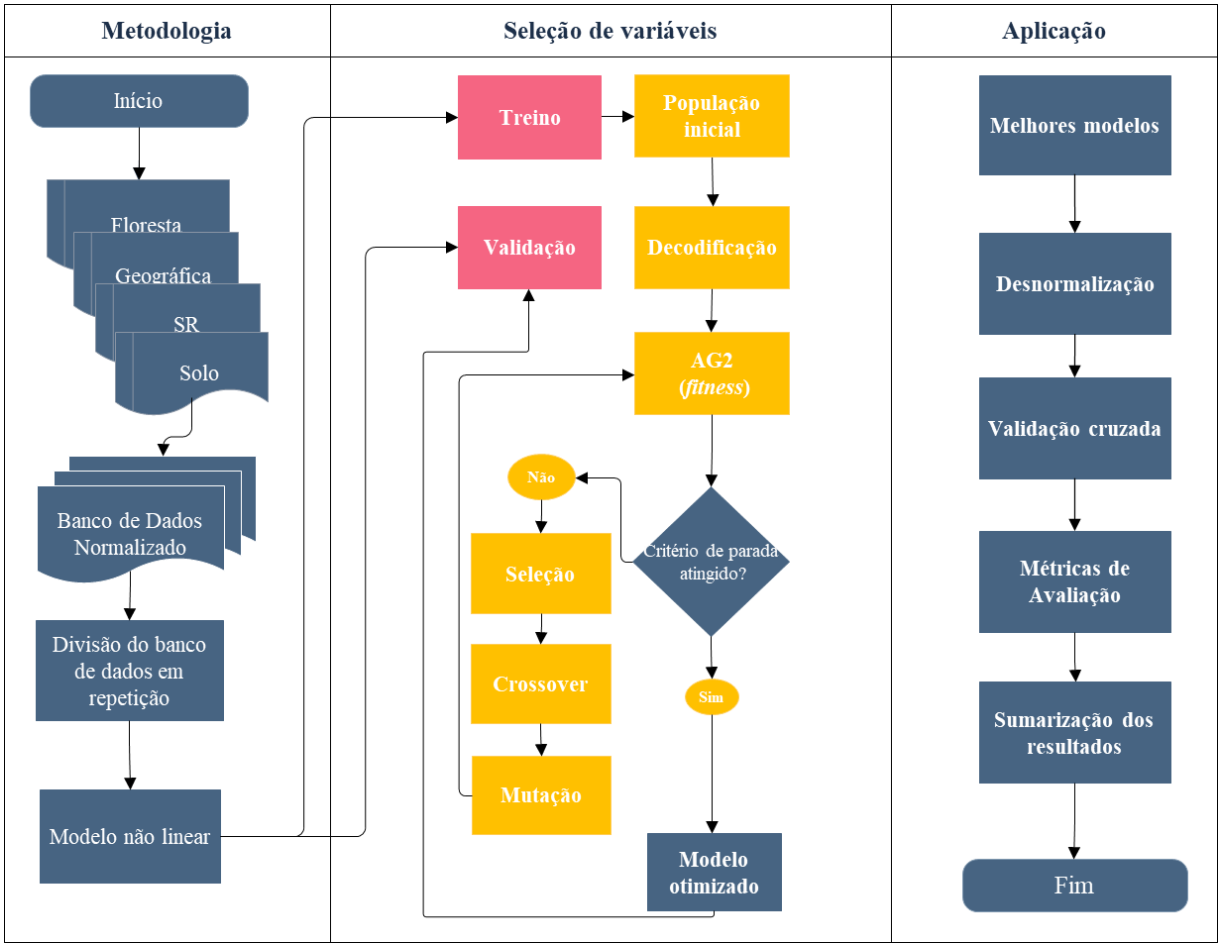
A função *fitness* (Equação 2) adotada foi baseada na soma quadrática dos resíduos, gerada pelo segundo algoritmo genético (MONTI, 2018) e desenvolvido para o ajuste do modelo de regressão não linear. Em que: h_i é a altura total da árvore observada e i ∈ {1,…, N}; e $\hat{h}_i$ é o valor estimado da altura total da árvore a partir do modelo de regressão não linear testado.

$$f = \sum_{i=1}^{n} (h_i - \hat{h}_i)^2 \qquad (2)$$

A implementação do algoritmo duplo foi desenvolvida em *software* R (*Version* 3.5.2 – © 2018 RStudio, Inc.), aplicado com paralelismo para otimização do tempo. O pacote utilizado foi o doParallel (OOI et al., 2019). A metodologia foi processada numa CPU com processador Intel (R) Core™ i7-6700 CPU @ 3.40 GHz, com memória instalada (RAM) de 16,0 GB. O fluxograma (Figura 3) descreve este processo metodológico, que envolve o uso de dois

algoritmos genéticos. Finalizado, os melhores modelos construídos pela metodologia foram novamente submetidos ao algoritmo genético já com as variáveis desnormalizadas para o cálculo dos seus parâmetros na escala normal.

Figura 3 - Processo metodológico utilizando dois algoritmos genéticos.



Fonte: Do autor (2020).

2.4 Critérios de avaliação dos modelos gerados

Uma forma de avaliar a contribuição das variáveis e suas combinações na capacidade preditiva foi pela visualização das mesmas em gráficos de histogramas, separando os modelos e suas entradas em classes de atributo conforme sua construção: a) Floresta; b) Ambiental; c) Sensor; d) Misto 1: Floresta + Ambiental; e) Misto 2: Floresta + Sensor; f) Misto 3: Ambiental + Sensor.

A partir dos dois melhores indivíduos gerados pelo algoritmo genético por modelo, estes foram comparados com sua forma clássica, ou seja, apenas com a inclusão do DAP. Uma vez definido os modelos, foi analisado pelo teste *t* a significância dos parâmetros ao nível de 5%.

Além disso, foram obtidas as análises de resíduos, como: RMSE, RMSE% e coeficiente de determinação (R^2). Posteriormente, avaliou-se a dispersão gráfica dos resíduos e também a avaliação gráfica da altura estimada versus DAP para as formas tradicionais dos modelos e o melhor indivíduo gerado por elas.

3 RESULTADOS

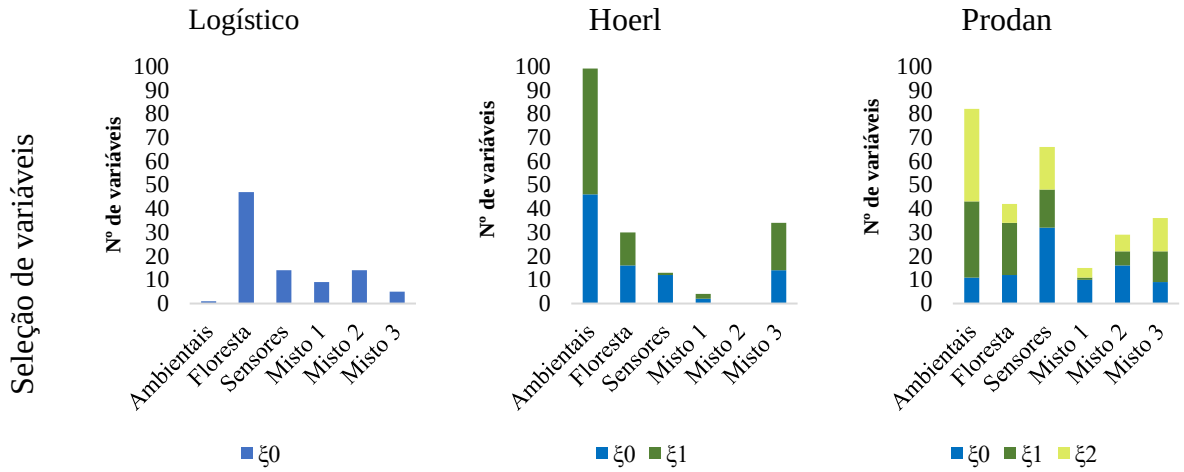
A seleção de variáveis e sua inclusão nos modelos seguiu um padrão conforme a natureza da variável, sendo estes associados as entradas do modelo de regressão não linear. Percebe-se ainda que o tipo de modelo hipsométrico expressa uma associação direta com o tipo de variável, uma vez que não seguem a mesma tendência de seleção. A Figura 4 traz a frequência das classes de variáveis escolhidas para os modelos hipsométricos, levando em consideração o melhor indivíduo para cada repetição nos três bancos de dados, totalizando os 90 melhores indivíduos para cada modelo.

Observa-se que a porcentagem de seleção das variáveis ambientais foram as mais recorrentes para os modelos de Hoerl (52,2%) e Prodan (58,8%), já que os mesmos apresentam mais de duas entradas. Esta tendência não foi observada no modelo Logístico, visto que o mesmo possui uma única entrada. Neste modelo, a variável floresta exerce uma forte pressão em sua seleção, já que os maiores valores de correlação entre altura estão nessa classe. Nesse caso, o valor chegou a 52,2% de uso dessa variável contra valores próximos de zero envolvendo as variáveis ambientais. Um outro fato que chama a atenção é que a medida que aumentaram-se as entradas do modelo, maior foi a participação das demais classes de variáveis em sua formação.

O AG selecionou na sua maioria variáveis individuais, principalmente para o modelo Logístico, mas com o aumento do número de entradas, ocorreram a maior presença de combinações entre atributos. O porcentual de variáveis combinadas foram 40%, 62% e 53% respectivamente, para os modelos Logístico, Hoerl e Prodan. Dentre as combinações, é possível que a maior presença do operador multiplicação para o modelo de Hoerl (53%), e da soma para o modelo de Prodan (32%).

A combinação entre classes de variáveis não foi uma opção interessante para os modelos testados com mais de duas entradas. Contudo, essa foi mais recorrente no modelo de Prodan, por apresentar um maior número de entradas. Também foi observado que à medida que se aumenta o número de entradas nos modelos, maior a dificuldade do algoritmo em encontrar boas soluções.

Figura 4- Histograma de frequência por classe de variáveis selecionadas e incorporadas nos modelos hipsométricos gerados pelo algoritmo genético independente do banco de dados.



Em que: ξ_i : variáveis independentes dos modelos.

Fonte: Do autor (2020).

A partir dos valores do RMSE (%) observa-se uma variação na qualidade da solução apresentada, e também que o número de entradas do modelo hipsométrico não interfere diretamente nas soluções ótimas. Porém, quanto maior a complexidade do modelo, maior será a probabilidade de soluções ruins, como pode ser observado nos valores máximos de RMSE (%) (Tabela 3). Entretanto, para as melhores soluções de ambos os modelos, pode ser observada a semelhança entre os valores mínimos de erro, tanto para o treino, quanto para a validação (Tabela 3). Diante dessa similaridade, considerando os três modelos, a seleção de variáveis pelo algoritmo genético é considerada robusta.

Tabela 3 - Estatística descritiva da raiz quadrada do erro médio porcentual (RMSE%) para a base de dados de treino e validação.

Modelos	**Dados**	**RMSE (%)**				
		Desvio Padrão	**Média**	**Máximo**	**Mínimo**	**CV**
Logístico	Treino	5.81	36.04	45.03	31.07	16.11
	Validação	5.83	36.05	46.50	29.81	16.18
Hoerl	Treino	3.74	43.94	48.43	32.78	8.51
	Validação	4.68	44.24	64.85	31.86	10.57
Prodan	Treino	4.73	40.72	48.05	32.30	11.62
	Validação	6.34	41.17	78.82	32.00	15.41

Em que: CV: coeficiente de variação.

Fonte: Do autor (2020).

Ao longo da convergência do algoritmo genético, nota-se que a quantidade de variáveis (50, 100 ou 150 variáveis) definida para cada banco de dados não interfere diretamente nos resultados encontrados, visto que os valores médios de RMSE (%) foram relativamente próximos entre si, independente do conjunto de dados e proporcionais entre eles. Para caracterizar que o banco de dados não interferiu nos resultados, verificou-se a proporção dos valores de RMSE com resultados inferiores a 40% em cada um deles, e também para cada modelo. Assim, observou-se uma constância dos valores superiores à 60, 40 e 35% referente a cada conjunto de dados, paras os modelos Logístico, Hoerl e Prodan, respectivamente.

A seleção/incorporação das variáveis e posterior ajuste do modelo não linear apresentou um tempo médio de 124 segundos, independente do modelo adotado. O tempo total envolvendo as repetições resultou no valor de 14.400 segundos ao todo.

Isolando os 2 melhores indivíduos do algoritmo genético por modelo e independente da repetição, tem-se uma análise comparativa. Contudo, a desnormalização das variáveis selecionadas e novos ajustes dos modelos foi realizada. Esse procedimento foi adotado para voltar à escala original das variáveis e com isso utilizar o modelo direto.

Assim, ao analisar as medidas de precisão referentes aos modelos genéricos e clássicos, observa-se que para a medida de precisão RMSE (%), os valores das métricas obtidas foram próximos entre si, mesmo alterando as variáveis de entrada (Tabela 3). Essa similaridade também ocorre para o coeficiente de determinação, com valores acima de 50%, menos para o modelo de Prodan original. Na sua maioria, os resultados mostraram ajustes satisfatórios, com a maioria das construções com precisões minimamente próximas umas às outras, com alguns resultados superiores aos originais.

O método adotado foi capaz de identificar não só as melhores variáveis e suas combinações matemáticas, como ainda sua posição mais significativa dentro do modelo. Essa característica é importante e por isso, mesmo obtendo valores similares para as métricas, seus valores foram próximos, o que garante expandir as relações funcionais em macrorregiões. O exemplo mais claro é o modelo Logístico que ficou associado ao DAP com *direct_ins* e *analytical*, para cada um dos 2 melhores modelos.

Na sua maioria, variáveis que associam informações de luminosidade apresentaram boas respostas, como pode ser visto na seleção das variáveis *direct_ins* e *analytical* nos três modelos (Tabela 4). Outras variáveis com bom retrospecto de soluções positivas podem ser observadas pela seleção das variáveis ambientais t_A e Zn_D, nos modelos de Hoerl e Prodan, respectivamente (Tabela 4).

Tabela 4 - Métricas de avaliação para os 2 melhores modelos genéricos Logístico, Hoerl e Prodan e suas formas originais para dados da validação cruzada. Análise descritiva dos modelos testados e suas estatísticas na predição das alturas das árvores.

	Modelos	Dados	θ_0	θ_1	θ_2	RMSE (%)	R^2 (%)
Logístico	$Y_i = \frac{\theta_0}{\left(1 + \theta_1 exp\left(-\theta_2 (DAP * direct\_ins)_i\right)\right)}$	Treino	15.0913*	2.9470 *	-0.0363*	25.38	55.18
		Validação				25.39	54.91
	$Y_i = \frac{\theta_0}{\left(1 + \theta_1 exp\left(-\theta_2 (DAP * analytical)_i\right)\right)}$	Treino	15.3436*	2.8785*	-0.1507*	25.78	53.75
		Validação				25.8	53.42
	$Y_i = \frac{\theta_0}{\left(1 + \theta_1 exp\left(-\theta_2 (DAP)_i\right)\right)}$	Treino	15.1658*	3.0387*	-0.1352*	26.01	52.93
		Validação				26.02	52.60
Hoerl	$Y_i = \theta_0 \theta_1^{direct\_ins_i} DAP_i^{\theta_2}$	Treino	1.7484*	1.1183*	0.5089*	25.39	55.13
		Validação				25.41	55.13
	$Y_i = \theta_0 \theta_1^{Zn\_D_i} DAP_i^{\theta_2}$	Treino	2.5627*	1.0029*	0.5206*	25.98	53.01
		Validação				26.00	53.01
	$Y_i = \theta_0 \theta_1^{DAP_i} DAP_i^{\theta_2}$	Treino	2.0931*	0.9903*	0.6576*	25.89	55.13
		Validação				25.91	55.13
Prodan	$Y = \frac{DAP_i}{\exp^{\theta_0 + \theta_1 DAP_i + \theta_3 direct\_ins_i^2}}$	Treino	0.0041*	0.0331*	-0.0151*	25.84	53.51
		Validação				25.86	53.18
	$Y = \frac{DAP_i}{\exp^{\theta_0 + \theta_1 DAP_i + \theta_3 (t\_A - analytical)_i^2}}$	Treino	-0.2007*	0.0323*	0.0003*	26.36	51.62
		Validação				26.38	51.27
	$Y = \frac{DAP_i}{\exp^{\theta_0 + \theta_1 DAP_i + \theta_3 DAP_i^2}}$	Treino	-0.3539*	0.0542*	-0.0006*	25.95	44.85
		Validação				25.98	44.85

Em que: DAP: diâmetro à altura do peito; t_A: CTC efetiva do dorizonte do solo na camada A; Zn_D: Zinco do horizonte do solo na camada D; analytical: analytical hillshading; direct_ins: direct insolation; longitudin: longitudinal curvature; Y_i: altura (m) do indivíduo i; θ_0, θ_1, θ_2, e θ_3: parâmetros dos modelos. θ_i: parâmetros estimados; RMSE: raiz quadrada do erro médio; RMSE%: raiz quadrada do erro médio percentual; R^2: coeficiente de determinação percentual; *significativo ao nível de probabilidade de 95%; ^{ns}: não significativo ao nível de probabilidade de 95%.

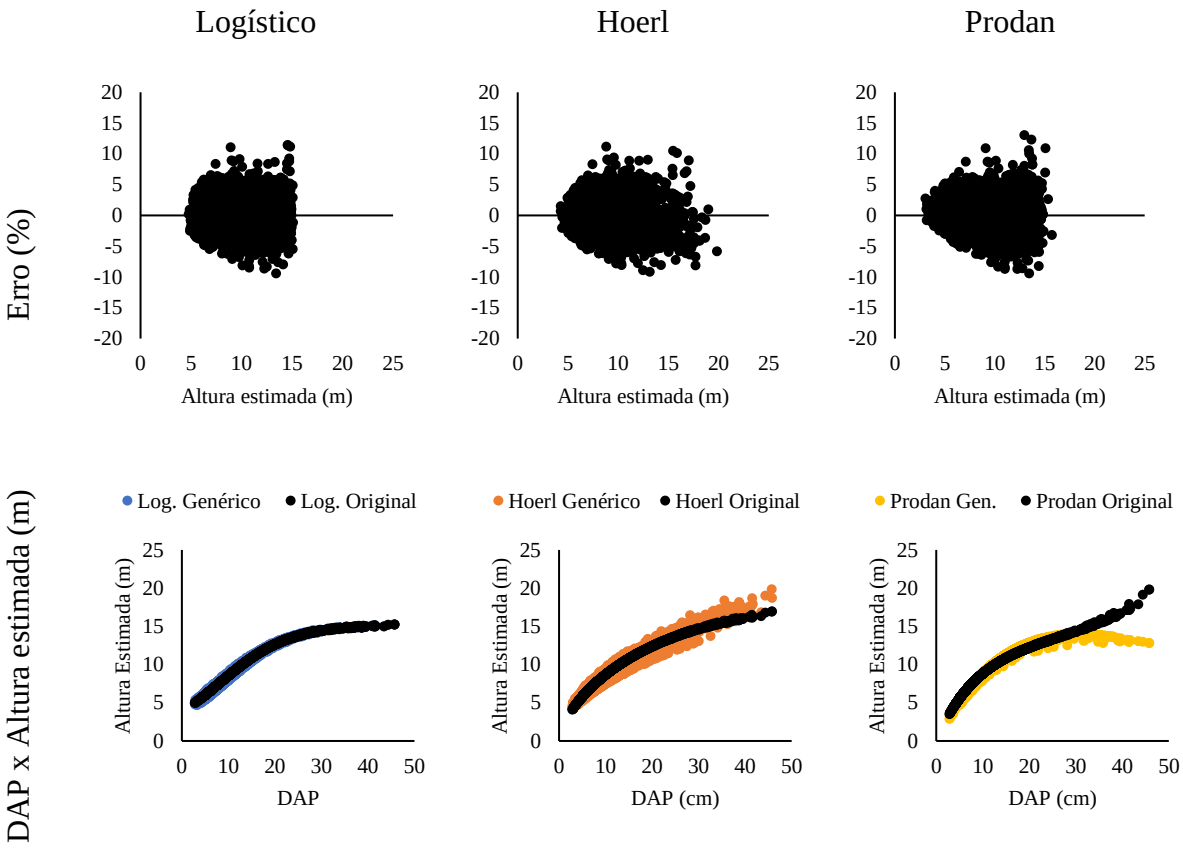
Fonte: Do autor (2020).

A inclusão de novas variáveis nos modelos originais não afetou a capacidade preditiva dos modelos, mantendo o efeito biológico das construções clássicas, essa afirmação argumentativa pode ser observada pelos valores similares dos coeficientes aos dos modelos originais, o que mostra robustez nas variáveis selecionadas (Tabela 4).

De forma complementar aos critérios de avaliação dos modelos anteriormente avaliados, avaliou-se os gráficos de dispersão dos resíduos para os melhores modelos gerados para Parabólico, Hoerl e Prodan (Figura 5). É notório que em todos os modelos os erros se encontram no intervalo entre 10 e -10 (m). Os modelos gerados da construção dos originais Prodan e Logístico apresentaram menor uniformidade em termos de distribuição de resíduos do que o modelo construído de Hoerl. Diferente dos modelos de uma e três entradas, os modelos genéricos de Hoerl, com duas entradas, foi levemente superior ao demonstrar uma distribuição

dos resíduos mais balanceada, com menor tendenciosidade em superestimar e subestimar a altura total das árvores (Figura 5).

Figura 5 - Gráfico de resíduos para os melhores modelos gerados por Parabólico, Hoerl e Prodan, e suas formas tradicionais em função de DAP para dados de validação.



Fonte: Do autor (2020).

Os modelos genéricos construídos a partir das formulações dos modelos Logístico e Prodan, com a implementação e associação de novas variáveis, não representaram um ganho relativamente significativo, tanto em aspectos de predição, quanto por métricas de erro, como pode ser observado na tabela 4 e figura 5. Já para os modelos genéricos de Hoerl, a inclusão de novas variáveis representou um ganho nas medidas de precisão, embora não represente um ganho alto, mas quando observado nos gráficos de dispersão dos erros (Figura 5), veem-se que os mesmos apresentam uma melhor dispersão, capazes de predizer indivíduos com até 20 m, já para o modelo clássico, o máximo alcançado pelo modelo foi de 15 m.

Na figura 5 observamos a relação da altura estimada pelos modelos não lineares com o diâmetro à altura do peito (DAP). Pode-se observar que os modelos genéricos seguiram a tendência da relação sigmoidal de modelos clássicos que se baseiam somente na relação diâmetro-altura (D-H). Para os melhores modelos observa-se uma similaridade entre eles, o de

melhorar esse comportamento sigmoidal da relação DAP-H, as vezes achatando suas extremidades em busca de uma boa resposta (Figura 5).

4 DISCUSSÃO

A construção de um modelo de regressão atende um pressuposto básico na ciência, no primeiro momento apresenta uma natureza preditiva para a obtenção de valores desconhecidos. Essa característica quantitativa é necessária em previsões para os diversos propósitos. Uma segunda opção é a capacidade explicativa do modelo de regressão, afim de contribuir na construção de hipóteses ecológicas e de processos envolvendo crescimento e produtividade. Os resultados derivados dos esforços computacionais da presente pesquisa empatam com a forma clássica dos modelos hipsométricos, utilizando apenas o diâmetro como variável de entrada. Contudo, examinando as opções apresentadas de modelos gerados tem-se a inclusão de variáveis ambientais e de sensores remotos como explicativas da altura total das árvores. Assim, o conhecimento adequado do comportamento da variável altura em povoamentos florestais é importante para definir estratégias de manejo, as quais exigem previsões precisas e confiáveis para confidenciá-las ao planejamento florestal, tanto para florestas nativas quanto para florestas plantadas. Relacionar a altura a variáveis ambientais pode proporcionar inferências sobre seu comportamento diante de possíveis mudanças climáticas (FIGUEIREDO et al., 2016; MAO et al., 2017; VALLET; PEROT, 2016; VIZCAÍNO-PALOMAR et al., 2016). Essa abordagem é atual e nos faz refletir sobre a capacidade das espécies recorrentes nesses ambientes naturais em dar condições adequadas a sua resiliência. Sabe-se que a altura das árvores forma o dossel da floresta, sendo um ponto importante para o processo de seleção de indivíduos. O chamado efeito filtro é uma condição necessária e evolutiva para o estabelecimento e recomposição da área por diferentes espécies (COSTA-SAURA et al., 2017; GARCÍA-HERNÁNDEZ et al., 2019; JAKOVAC et al., 2016). Além disso, a altura das árvores está associada a área de copa e produção de assimilados (GLATTHORN et al., 2017; HOLIŠOVÁ et al., 2016; ZHANG et al., 2018), sendo reflexo direto das condições ambientais locais. Pela lógica, quanto menor a área foliar e a altura das árvores, menor a capacidade de incorporar CO2 atmosférico, resultando em menores taxas de crescimento (POORTER et al., 2018; PRADO-JUNIOR et al., 2016; WARING; LANDSBERG; LINDER, 2016).

Em relação à estrutura dos modelos e seus desdobramentos matemáticos, a abordagem adotada considerando o algoritmo genético na seleção de variáveis torna-se viável. O método apresentou uma série de vantagens, sendo a principal delas a agilidade em se trabalhar com um

grau elevado de combinações de variáveis. A cada geração do processo o algoritmo genético foi capaz de identificar regiões do espaço amostral **Ω** com excelentes ótimos locais. Fato comprovado pela igualdade estatística entre modelos clássicos e suas formas genéricas. O segundo ponto foi desconsiderar a geração de parâmetros iniciais requisitadas pelos algoritmos clássicos de ajuste, como levenberg-marquardt e newton, já se adotou um segundo algoritmo genético específico para este fim.

O comportamento dos coeficientes do modelo não linear ao se associarem com as variáveis geradas pelo algoritmo genético tiveram um comportamento adequado. Os valores encontrados para os melhores modelos seguiram valores similares aos modelos clássicos, o que os torna robustos e confiáveis. O estudo de modelos não lineares relacionados à altura-diâmetro também foi alvo em outros trabalhos realizados em outras macrorregiões do Brasil, como o estudo de Barbosa et al. (2019). As combinações das variáveis quando associados ao diâmetro, apresentaram o mesmo comportamento das relações de altura-diâmetro, se aproximam de uma forma sigmoidal (PENG; ZHANG; LIU, 2001; GÓMEZ-GARCÍA et al., 2015; COSTA; SCHRODER; FINGER, 2016; FERRAZ FILHO et al., 2018).

Nossos modelos de altura individual das árvores selecionaram e combinaram variáveis de caráter do povoamento, solo, geográficas e de sensor para explicar seu comportamento através de uma metodologia de inteligência computacional, diferente de abordagens tradicionais da literatura. As melhores variáveis alternativas às populacionais foram as classes sensor e ambiental, embora erros de previsão ainda ocorreram, os seus resultados foram relativamente bons, permanecendo o RMSE % abaixo de 30% nas melhores respostas. O problema das estimativas não está na metodologia, a qual mostrou-se muito robusta e aplicável, mas sim relacionado a complexidade dos dados trabalhados que derivam de uma alta diversidade de espécies tropicais. A composição florística em estandes florestais naturais, está associado a aspectos específicos de cada espécie, cada qual apresenta características fisiológicas e funcionais intrínsecas únicas, com valores diferentes de densidade da madeira, exigência de luz (tolerante à sombra *vs.* exigente à luz), variação radial específica, que provavelmente determinam a taxa de crescimento específica da espécie (MENSAH et al., 2018a; MENSAH; SEIFERT; GLÈLÈ KAKAÏ, 2016). Os resultados encontrados de RMSE % e R^2 vão de encontro à valores médios encontrados em florestas naturais no sul do Brasil, com *Araucaria angustifolia* (COSTA et al., 2018; COSTA; SCHRODER; FINGER, 2016). O mesmo não pode ser observado para florestas da região amazônica, que apresenta valores de erro menores (BARBOSA et al., 2019; CASSOL et al., 2018), e por mais que a Amazônia apresente alta diversidade de espécies, as condições climáticas da região são relativamente diferentes das

encontradas na região da bacia do rio Grande, com períodos mais uniformes de chuva e temperatura, ao contrário do encontrado no sudeste brasileiro, que impacta diretamente nos estandes florestais. Atributos ambientais e condições climáticas heterogêneas ao longo da bacia atuam como sinal filogenético no crescimento heterogênico entre espécies e indivíduos de árvores.

A inclusão de variáveis ambientais em modelos clássicos, devem ser entendidas como a representação mais precisa e realista de atributos específicos locais. A seleção e combinação de variáveis pode ser atribuído à modelos de efeito misto, dada a magnitude do algoritmo em selecionar variáveis explicativas. Sharma, Vacek e Vacek (2016) e Fu et al. (2017), apresentaram resultados ruins ao atribuírem a relação altura-diâmetro a modelos não lineares de efeito misto em povoamentos florestais naturais, contudo, abordagens com seleção e combinação de variáveis surgem como uma zona de escape para melhores resultados, uma vez esclarecida a capacidade do algoritmo genético em associar um grupo de variáveis a uma boa resposta. Mensah et al. (2018) afirmam que há evidências crescentes de que nenhuma forma geral de função se encaixa melhor na alometria altura-diâmetro em macroescalas, porém, a inclusão de variáveis ambientais dá um caráter explicativo aos modelos empíricos.

As variáveis que tiveram melhor resposta no estudo foram as variáveis da classe sensor, como *direct insolation* (insolação direta) e *analytical hillshading* (o ângulo entre a superfície e os feixes de luz recebidos), que quando associadas ao DAP tiveram um ganho de resposta nos modelos. Pode-se constatar que tais variáveis influenciam geograficamente o porte das árvores nos povoamentos. Potapenko, Kunah e Fedushko (2019) em seu estudo sobre poluição do solo associam diretamente a variável de insolação direta a processos biológicos em regimes ambientais. Por ela está intrinsicamente associado a fatores edafoclimáticos, esta informação corrobora com o crescimento da altura das árvores, identificando um gradiente bem definido na região de estudo. Por outro lado, valores altos do atributo *analytical hillshading* indicam menor estatura de seus indivíduos, a variável representa o sombreamento do terreno, assim, áreas que possuem maior área de sombra possui indivíduos menores.

Variáreis que refletem a luminosidade incidida na floresta caracterizam bem o porte das árvores. Deste modo, atributos que representem essa característica costumam apresentar uma alta correlação com a altura do dossel, como pode ser observado em trabalhos que procuraram estimar a altura do dossel com índices de área foliar (MCDOWELL et al., 2002; QU et al., 2018; YUAN et al., 2013).

Entre as variáveis ambientais consideradas, o t_A (CTC a pH 7 do horizonte do solo na camada A) e Zn_D (Zinco do horizonte do solo na camada D) foram as que tiveram melhores

resultados. Solos com maiores valores de t_A, apresentam maior troca catiônica, e, por consequência, maior quantidade de nutrientes, o que favorece o crescimento dos indivíduos do povoamento. O AG foi capaz de identificar essas variações entre as parcelas e importa-las ao modelo, associando a área a indivíduos com menor ou maior altura. Fricker et al. (2019) também observaram que o pH do solo influencia diretamente as escalas em que os indivíduos vão se apresentar, maiores alturas para árvores coníferas em solos ácidos (pH baixo), e árvores de carvalho de baixa estatura para solos mais básicos (pH alto).

Como sabemos, as espécies florestais apresentam uma série particularidades e necessidades quanto a nutrição de micronutrientes, com valores diferentes para cada espécie (BÜNDCHEN et al., 2013). O zinco, que foi selecionado pelo método, é um nutriente essencial para as plantas, e está entre os metais pesadosmais móveis no sistema solo/planta, principalmente em solos ácidos como florestas, e ele pode ser um fator limitante no crescimento das árvores (MOSQUERA-LOSADA; LÓPEZ-DÍAZ; RIGUEIRO-RODRÍGUEZ, 2009; RIGUEIRO-RODRÍGUEZ; MOSQUERA-LOSADA; FERREIRO-DOMÍNGUEZ, 2012; SAUERBECK, 1991). Assim, áreas com solos com menor taxa de zinco tinham como característica o menor porte das árvores, e a seleção desta variável funcionou como um filtro na regressão do modelo, caracterizando estes indivíduos a uma menor altura. A seleção destas variáveis ambientais de solo dá um caráter explicativo aos modelos, filtrando o potencial de crescimento e de produtividade local.

A precisão da seleção de variáveis conseguiu se comportar bem, independente da dimensão testada frente ao número de variáveis (50, 100 e 150). Os resultados obtidos demonstraram que este número não influenciou a geração de modelos de qualidade, reduzindo a dimensionalidade do banco de dados. Os resultados condizem com outras heurísticas e meta-heurísticas de seleção difundidas na área de *data mining*. Pohjankukka et al. (2018) em seu estudo, demonstraram a capacidade do AG em selecionar variáveis para predizer a altura em povoamentos florestais da Finlândia. Hong et al. (2018) utilizaram o algoritmo genético para selecionar os melhores atributos que explicassem a incidência de incêndios florestais na China, o algoritmo foi capaz de reduzir o número de variáveis que explicassem o problema e melhorar a resposta.

Outras meta-heurísticas também apresentam resultados positivos, por exemplo, Ghaemi e Feizi-Derakhshi (2016) encontraram respostas satisfatórias ao aplicarem o algoritmo FOA (algoritmo de otimização florestal) para selecionar recursos. Desta forma, a aplicabilidade

de técnicas de seleção de variáveis se mostra eficiente na melhoria dos resultados, quando há um elevado número de atributos disponíveis.

O algoritmo genético produz combinações lineares latentes entre as variáveis e as usa nas construções dos modelos clássicos de regressão não linear, nesse quesito ela se assemelha a *partial least squares* (PLS). A PLS seleciona um conjunto reduzido de variáveis através de correlação de combinações lineares dos recursos originais e as utiliza como entrada para modelos de regressão (SCOTTI et al., 2016). Existem outros métodos bastante usuais que também realizam a retração e seleção de variáveis, dentre os mais comuns estão LASSO (operador de seleção e retração menos absoluto) (MORENO et al., 2017), *stepwise* (RUIZ et al., 2018) e métodos metaheurísticos como enxame de partículas (PSO) (JAIN; JAIN; JAIN, 2018) e otimização de colônias de formigas (SHUNMUGAPRIYA; KANMANI, 2017). Contudo, o uso do algoritmo genético torna-se mais interessante, pois ele em sua construção no código e operadores realiza a combinação ou não entre variáveis. Não requisitando nesse caso que as entregue como input inicial do processo.

O método como esperado expressou maior dificuldade de processamento (maior tempo) e parametrização, mas o uso do paralelismo entre os núcleos do processador, reduziu substancialmente o tempo de processamento. O algoritmo genético com função de diminuição do erro (soma de quadrados do erro) se consolidou como uma boa alternativa para seleção de variáveis. Primeiro, o algoritmo AG mostrou-se sensível à medida que o algoritmo evoluía, diminuindo o erro e selecionando as melhores respostas para os modelos. Segundo, a restrição do número de entradas nos modelos de saídas se mostrou eficiente, uma e duas entradas obtendo as melhores respostas, mas encontrou dificuldade de encontrar boas respostas no modelo de três entradas. Assim, o método deve ser utilizado com ressalvas quando há um número elevado de entradas no modelo, já que nem sempre foi capaz de obter bons resultados.

5 CONCLUSÃO

Uma abordagem com algoritmo genético para seleção e combinações de variáveis populacionais, ambientais e de sensor foi utilizada em modelos de regressão não linear para selecionar a melhor configuração de atributos para explicar a altura individual das árvores. A inclusão de variáveis ambientais em modelos de regressão clássicos foi entendida como a representação realista e representativa da área no ajuste dos modelos. As variáveis que tiveram as melhores respostas foram as da classe sensor, como *direct insolation* e *analytical hillshading*, e da classe ambiental, como pH e Zn, que quando associadas ao diâmetro à altura do peito

inserem uma explicação ecológica nos modelos de regressão. Os resultados com a inclusão de variáveis tiveram respostas similares aos encontrados nos modelos originais. Modelos com menos entradas tiveram melhores respostas, e o aumento dos números de entradas associa a respostas com modelos menos otimizados. O AG foi capaz de encontrar relações complexas de associação de variáveis, os quais por métodos clássicos seriam triviais e cansativos de se fazer. Estudos posteriores devem ser feitos em outras linhas do manejo florestal para gerar novas opções de modelos que possam auxiliar na explicação ecológica dos efeitos fitogeográficos.

REFERÊNCIAS

AHMED, O. S. et al. Characterizing stand-level forest canopy cover and height using Landsat time series, samples of airborne LiDAR, and the Random Forest algorithm. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 101, p. 89–101, 2015.

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

BARBOSA, R. I. et al. Allometric models to estimate tree height in northern amazonian ecotone forests. **Acta Amazonica**, Manaus, v. 49, n. 2, p. 81–90, 2019.

BÜNDCHEN, M. et al. Nutritional status and nutrient use efficiency in tree species of subtropical forest in southern Brazil. **Scientia Forestalis/Forest Sciences**, Piracicaba, v. 41, n. 98, p. 227–236, 2013.

CASSOL, H. L. G. et al. Improved tree height estimation of secondary forests in the Brazilian Amazon. **Acta Amazonica**, Manaus, v. 48, n. 3, p. 179–190, 2018.

COSTA-SAURA, J. M. et al. Environmental filtering drives community specific leaf area in Spanish forests and predicts relevant changes under future climatic conditions. **Forest Ecology and Management**, Amsterdam, v. 405, n. September, p. 1–8, 2017.

COSTA, E. A. et al. Height-Diameter Models for *Araucaria angustifolia* (Bertol.) Kuntze in Natural Forests. **Journal of Agricultural Science**, Ottawa, v. 10, n. 8, p. 133, 2018.

COSTA, E. A.; SCHRODER, T.; FINGER, C. A. G. Relação altura-diâmetro para *Araucaria angustifolia* (Bertol.) kuntze no sul do Brasil. **Cerne**, Lavras, v. 22, n. 4, p. 493–500, 2016.

CURI, N. et al. **Zoneamento ecológico-econômico do Estado de Minas Gerais: componentes geofísicos e biótico**. 1. ed. Lavras: Editora UFLA, 2008.

DANIEL, C.; WOOD, F. S. **Fitting Equations to Data: Computer Analysis of Multifactor Data**. 2nd. ed. New York: John Wiley & Sons, 1980.

FERRAZ FILHO, A. C. et al. Height-diameter models for *Eucalyptus* sp. plantations in Brazil. **Cerne**, Lavras, v. 24, n. 1, p. 9–17, 2018.

FERREIRA, A. R. L. et al. Assessing anthropogenic impacts on riverine ecosystems using nested partial least squares regression. **Science of the Total Environment**, New York, v. 583, p. 466–477, 2017.

FIGUEIREDO, E. O. et al. LIDAR-based estimation of bole biomass for precision management of an Amazonian forest: Comparisons of ground-based and remotely sensed estimates. **Remote Sensing of Environment**, New York, v. 187, p. 281–293, 2016.

FIGUEIREDO FILHO, A. et al. Evolution of the hypsometric relationship in *Araucaria angustifolia* plantations in the mid-south region of Paraná state. **Cerne**, Lavras, v. 16, n. 3, p. 347–357, 2010.

FRICKER, G. A. et al. More than climate? Predictors of tree canopy height vary with scale in complex terrain, Sierra Nevada, CA (USA). **Forest Ecology and Management**, New York, v. 434, n. November 2018, p. 142–153, 2019.

FU, L. et al. A generalized nonlinear mixed-effects height to crown base model for Mongolian oak in northeast China. **Forest Ecology and Management**, New York, v. 384, p. 34–43, 2017.

GALAGALI, N.; MARZOUK, Y. M. Exploiting network topology for large-scale inference of nonlinear reaction models. **Journal of the Royal Society Interface**, London, v. 16, n. 152, 2019.

GARCÍA-HERNÁNDEZ, M. DE LOS Á. et al. Effects of environmental filters on early establishment of cloud forest trees along elevation gradients: Implications for assisted migration. **Forest Ecology and Management**, New York, v. 432, n. August 2018, p. 427–435, 2019.

GARCÍA, M. et al. Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 66, p. 159–173, 2018.

GHAEMI, M.; FEIZI-DERAKHSHI, M. R. Feature selection using Forest Optimization Algorithm. **Pattern Recognition**, Amsterdam, v. 60, p. 121–129, 2016.

GLATTHORN, J. et al. Effects of forest management on stand leaf area: Comparing beech production and primeval forests in Slovakia. **Forest Ecology and Management**, New York, v. 389, p. 76–85, 2017.

GÓMEZ-GARCÍA, E. et al. Height-diameter models for maritime pine in portugal: A comparison of basic, generalized and mixed-effects models. **IForest**, Potenza, v. 9, n. Feb 2016, p. 72–78, 2015.

GORGANNEJAD, S. et al. Quantitative prediction of the aged state of Ni-base superalloys using PCA and tensor regression. **Acta Materialia**, Oxford, v. 165, p. 259–269, 2019.

HOLIŠOVÁ, P. et al. Comparison of assimilation parameters of coppiced and non-coppiced sessile oaks. **IForest**, Potenza, v. 9, n. 4, p. 553–559, 2016.

HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, New York, v. 630, p. 1044–1056, 2018.

HUETE, A. . A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, New York, v. 25, n. 3, p. 295–309, ago. 1988.

JAIN, I.; JAIN, V. K.; JAIN, R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. **Applied Soft Computing Journal**, Amsterdam, v. 62, p. 203–215, 2018.

JAKOVAC, C. C. et al. Land use as a filter for species composition in Amazonian secondary forests. **Journal of Vegetation Science**, Knivsta, v. 27, n. 6, p. 1104–1116, 2016.

JUSTICE, C. O. et al. The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. **IEEE Transactions on Geoscience and Remote Sensing**, New York, v. 36, n. 4, p. 1228–1249, 1998.

KANDARE, K. et al. Individual tree crown approach for predicting site index in boreal forests using airborne laser scanning and hyperspectral data. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 60, n. September 2016, p. 72–82, 2017.

KHANDELWAL, M. et al. Function development for appraising brittleness of intact rocks using genetic programming and non-linear multiple regression models. **Engineering with Computers**, New York, v. 33, n. 1, p. 13–21, 2017.

MAO, L. et al. Environmental landscape determinants of maximum forest canopy height of boreal forests. **Journal of Plant Ecology**, Beijing, v. 12, n. 1, p. 96–102, 2017.

MATASCI, G. et al. Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. **Remote Sensing of Environment**, New York, v. 209, p. 90–106, 2018.

MCDOWELL, N. et al. The relationship between tree height and leaf area: Sapwood area ratio. **Oecologia**, Berlin, v. 132, n. 1, p. 12–20, 2002.

MEHMOOD, T.; AHMED, B. The diversity in the applications of partial least squares: An overview. **Journal of Chemometrics**, New York, v. 30, n. 1, p. 4–17, 2016.

MENSAH, S. et al. Height – Diameter allometry in South Africa's indigenous high forests: Assessing generic models performance and function forms. **Forest Ecology and Management**, Amsterdam, v. 410, n. November 2017, p. 1–11, 2018.

MENSAH, S.; SEIFERT, T.; GLÈLÈ KAKAÏ, R. Patterns of biomass allocation between foliage and woody structure: The effects of tree size and specific functional traits. **Annals of Forest Research**, Arezzo, v. 59, n. 1, p. 49–60, 2016.

MILLER, J. D.; THODE, A. E. Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR). **Remote Sensing of Environment**, New York, v. 109, n. 1, p. 66–80, 2007.

MORENO, P. C. et al. Individual-tree diameter growth models for mixed Nothofagus second growth forests in southern Chile. **Forests**, Basel, v. 8, n. 12, p. 1–19, 2017.

MOSQUERA-LOSADA, M. R.; LÓPEZ-DÍAZ, M. L.; RIGUEIRO-RODRÍGUEZ, A. Zinc and copper availability in herbage and soil of a *Pinus radiata* silvopastoral system in Northwest Spain after sewage-sludge and lime application. **Journal of Plant Nutrition and Soil Science**, Weinheim, v. 172, n. 6, p. 843–850, 2009.

NOORDERMEER, L. et al. Direct and indirect site index determination for Norway spruce and Scots pine using bitemporal airborne laser scanner data. **Forest Ecology and Management**, Amsterdam, v. 428, p. 104–114, 2018.

NORYANI, M. et al. Material selection of natural fibre using a stepwise regression model with error analysis. **Journal of Materials Research and Technology**, Rio de Janeiro, v. 8, n. 3, p. 2865–2879, 2019.

OOI, Hong et al. **doParallel:** Foreach Parallel Adaptor for the 'parallel' Package, 2019. Disponível em: https://cran.r-project.org/web/packages/doParallel/index.html

PAN, J. C.; LEE, M. H.; TSAI, M. Y. Reversible jump Markov chain Monte Carlo algorithms for Bayesian variable selection in logistic mixed models. **Communications in Statistics: Simulation and Computation**, New York, v. 47, n. 8, p. 2234–2247, 2018.

PARRESOL, B. R. et al. Modeling forest site productivity using mapped geospatial attributes within a South Carolina Landscape, USA. **Forest Ecology and Management**, Amsterdam, v. 406, p. 196–207, 2017.

PEARL, R.; REED, L. J. On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. **Proceedings of the National Academy of Sciences**, New York, v. 6, n. 6, p. 275–288, 1920.

PENG, C.; ZHANG, L.; LIU, J. Developing and validating nonlinear height-diameter models for major tree species of ontario's boreal forests. **Northern Journal of Applied Forestry**, Bethesda, v. 18, n. 3, p. 87–94, 2001.

POHJANKUKKA, J. et al. Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery. **Scandinavian Journal of Forest Research**, Basingstoke,v. 33, n. 7, p. 681–694, 2018.

POORTER, L. et al. Can traits predict individual growth performance? A test in a hyperdiverse tropical forest. **New Phytologist**, London, v. 219, n. 1, p. 109–121, 2018.

POTAPENKO, O.; KUNAH, O. M.; FEDUSHKO, M. P. The effect of technological oil spill in soil within electrical generation substations, analysed by ecological regime in the context of relief properties. **Biosystems Diversity**, Dnipro, v. 27, n. 1, p. 43–50, 2019.

PRADO-JUNIOR, J. A. et al. Conservative species drive biomass productivity in tropical dry forests. **Journal of Ecology**, New York, v. 104, n. 3, p. 817–827, 2016.

PRODAN, M. **Forest Biometrics**. 1. ed., New York: PERGAMON, 1968.

QI, J. et al. A modified soil adjusted vegetation index. **Remote Sensing of Environment**, New York, v. 48, n. 2, p. 119–126, 1994.

QU, Y. et al. Remote sensing of leaf area index from LiDAR height percentile metrics and comparison with MODIS product in a selectively logged tropical forest area in Eastern Amazonia. **Remote Sensing**, Basel, v. 10, n. 6, 2018.

R CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018.

RIGUEIRO-RODRÍGUEZ, A.; MOSQUERA-LOSADA, M. R.; FERREIRO-DOMÍNGUEZ, N. Pasture and soil zinc evolution in forest and agriculture soils of Northwest Spain three years after fertilisation with sewage sludge. **Agriculture, Ecosystems and Environment**, Amsterdam, v. 150, p. 111–120, 2012.

ROUSE JR, J. W. ET AL. **Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation**. Greenbelt: NASA/GSFC, 1973.

RUIZ, L. Á. et al. An object-based approach for mapping forest structural types based on low-density LiDAR and multispectral imagery. **Geocarto International**, Abingdon, v. 33, n. 5, p. 443–457, 2018.

SAUERBECK, D. R. Plant element and soil properties governing uptake and availability of heavy metals derived from sewage sludge. **Water, Air, and Soil Pollution**, Dordrecht, v. 57–58, n. 1, p. 227–237, 1991.

SCOLFORO, J. R. et al. **Equções de volume, peso de materia seca e carbono para diferentes fisionomias da flora nativa**. 1. ed. Lavras: Editora UFLA, 2008.

SCOTTI, M. T. et al. Variable-selection approaches to generate QSAR models for a set of antichagasic semicarbazones and analogues. **Chemometrics and Intelligent Laboratory Systems**, New York, v. 154, p. 137–149, 2016.

SHARMA, R. P.; BREIDENBACH, J. Modeling height-diameter relationships for Norway spruce, Scots pine, and downy birch using Norwegian national forest inventory data. **Forest Science and Technology**, Abingdon, v. 11, n. 1, p. 44–53, 2015.

SHARMA, R. P.; VACEK, Z.; VACEK, S. Nonlinear mixed effect height-diameter model for mixed species forests in the central part of the Czech Republic. **Journal of Forest Science**, Prague, v. 62, n. 10, p. 470–484, 2016.

SHUNMUGAPRIYA, P.; KANMANI, S. A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid). **Swarm and Evolutionary Computation**, Amsterdam, v. 36, n. February 2016, p. 27–36, 2017.

STABEN, G.; LUCIEER, A.; SCARTH, P. Modelling LiDAR derived tree canopy height from Landsat TM, ETM+ and OLI satellite imagery—A machine learning approach. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 73, p. 666–681, 2018.

TÉO, S. J. et al. Relação hipsométrica geral com atributos biológicos para povoamentos de Pinus taeda L. **Cerne**, Lavras, v. 23, n. 4, p. 403–411, 2017.

UZOH, F. C. C. Height-Diameter Model for Managed Even-aged Stands of Ponderosa Pine for the Western United States Using Hierarchical Nonlinear Mixed-Effects Model. **Australian Journal of Basic and Applied Sciences**, Punjab, v. 11, p. 69–87, 2017.

VALLET, P.; PEROT, T. Tree diversity effect on dominant height in temperate forest. **Forest Ecology and Management**, Amsterdam, v. 381, p. 106–114, 2016.

VIEIRA, S. M. et al. **Metaheuristics for feature selection: application to sepsis outcome prediction**. Evolutionary Computation (CEC), 2012 IEEE Congress on. **Anais**...Brisbane: 2012

VIZCAÍNO-PALOMAR, N. et al. Adaptation and plasticity in aboveground allometry variation of four pine species along environmental gradients. **Ecology and Evolution**, Oxford, v. 6, n. 21, p. 7561–7573, 2016.

WANG, L.; WANG, Y.; CHANG, Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. **Methods**, San Diego, v. 111, p. 21–31, 2016.

WANG, S. et al. Predicting ship fuel consumption based on LASSO regression. **Transportation Research Part D: Transport and Environment**, New York, v. 65, n. October 2017, p. 817–824, 2018.

WARING, R.; LANDSBERG, J.; LINDER, S. Tamm Review: Insights gained from light use and leaf growth efficiency indices. **Forest Ecology and Management**, Amsterdam, v. 379, p. 232–242, 2016.

WILSON, E. H.; SADER, S. A. Detection of forest harvest type using multiple dates of Landsat TM imagery. **Remote Sensing of Environment**, New York, v. 80, n. 3, p. 385–396, 2002.

YUAN, Y. et al. Examination of the quantitative relationship between vegetation canopy height and LAI. **Advances in Meteorology**, Londres, v. 2013, 2013.

ZHANG, Z. et al. The tree height-related spatial variances of tree sap flux density and its scale-up to stand transpiration in a subtropical evergreen broadleaf forest. **Ecohydrology**, Malden, v. 11, n. 7, p. 1–12, 2018.

ZHAO, Q. et al. Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. **Forest Ecology and Management**, Amsterdam, v. 434, n. December 2018, p. 224–234, 2019.

ZIMMER, J.; ANZANELLO, M. J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Production**, Rio de Janeiro, v. 24, n. 1, p. 84–93, 2014.

ARTIGO 2 - MODELAGEM DA ALTURA INDIVIDUAL DE ÁRVORES VIA TÉCNICAS DE TREINAMENTO DE MÁQUINA E META-HEURÍSTICA

MODELING THE INDIVIDUAL TREE HEIGHT BY MACHINE LEARNING AND MEHEURISTIC TECHNIQUES

Artigo formatado conforme a NBR 6022 (ABNT, 2003) e adaptado as exigências do Manual de Normalização de Trabalhos Acadêmicos da UFLA.

RESUMO

A altura é uma variável comumente associada a erros de medição, mas dada sua importância na incorporação de modelos volumétricos e outros atributos biométricos, ela se torna importantíssima para resultados mais assertivos, e é estudada frequentemente para melhorar sua precisão. Para a obtenção deste atributo, normalmente modelos hipsométricos são utilizados, devido a seu baixo esforço e boas respostas, porém, nos últimos anos se viu necessário respostas mais precisas, por isso, variáveis ambientais veem sendo incorporadas em modelos de regressão e de inteligência. Assim, o objetivo do estudo foi modelar a altura individual do povoamento para a vegetação nativa presente na bacia do rio Grande – MG, com o uso de técnicas clássicas e de aprendizagem de máquina aplicadas a um grande conjunto de observações e variáveis. O banco de dados contém informações do povoamento, solo, morfométricas, edáficas, espectrais e geográficas. Foi aplicado um modelo clássico de regressão (Parabólico) com base na relação altura diâmetro, e diferentes abordagens com o algoritmo *Random Fores*t (RF) para todos os atributos: na sua forma pura com todas as variáveis (*RFall*); as mesmas entradas do modelo parabólico (*RFuni*); e um método híbrido com algoritmo genético para seleção de variáveis (*AGRFmulti*). A base de dados foi dividida em treino e validação, proporcionalmente por parcela e quantificada as métricas: a) coeficiente de determinação (R^2); b) bias e; c) raiz da média quadrática do erro em porcentagem (RMSE%). O modelo *AGRFmulti* foi superior as outras abordagens, com melhores métricas de avaliação, com valores de RMSE de 14,83%, bias de -0,01 m e R^2 de 84% para dados de treino, e para validação 22,79%, 0,74 m e 68%, de RMSE, bias e R^2, respectivamente. O método híbrido reduziu o corpo de variáveis independentes para 6, dentre as 189 possíveis, elas sendo por ordem de importância: DAP (diâmetro à altura do peito), Y (latitude), PLE (*Potential latent heat flux*), Fe_D (quantidade de ferro na camada de solo do horizonte D), BIO15 (*Precipitation Seasonality*) e Mg_D (quantidade de magnésio na camada de solo do horizonte D). Os resultados encontrados demonstram que o algoritmo genético foi uma ferramenta hábil e eficaz na seleção de variáveis, e sua abordagem com RF, concatenam em boas respostas na modelagem da altura individual em povoamentos florestais.

Palavras-chave: *Feature selection*. *Random forest*. Algoritmo genético.

ABSTRACT

Height is a variable commonly associated with measurement errors, but given its importance in incorporating volumetric models and other biometric attributes, it becomes very important for more assertive results, and is often studied to improve its accuracy. To obtain this attribute, usually hypsometric models are used due to their low effort and good responses, however, in recent years more precise answers have been necessary, so environmental variables have been incorporated into regression and intelligence models. Thus, the objective of the study was to model the individual stand height for native vegetation present in the Rio Grande basin - MG, using classical and machine learning techniques applied to a large set of observations and variables. The database contains stand, soil, morphometric, edaphic, spectral and geographic information. A classical (Parabolic) regression model was applied based on the height-diameter relationship, and different approaches with the Random Forest (RF) algorithm for all attributes: in its pure form with all variables (RFall); the same parabolic model entries (RFuni); and a hybrid method with genetic algorithm for variable selection (AGRFmulti). The database was divided into training and validation, proportionally by portion and quantified the metrics: a) coefficient of determination (R^2); b) bias and; c) root of the quadratic mean error percentage (RMSE%). The AGRFmulti model was superior to other approaches, with better evaluation metrics, with RMSE values of 14.83%, bias of -0.01 m and R^2 of 84% for training data, and for validation 22.79%, 0 , 74 m and 68%, of RMSE, bias and R^2, respectively. The hybrid method reduced the body of independent variables to 6 out of 189 possible, in order of importance: DAP (diameter at breast height), Y (latitude), PLE (Potential latent heat flux), Fe_D (amount of iron in horizon D soil layer), BIO15 (Precipitation Seasonality) and Mg_D (amount of magnesium in horizon D soil layer). The results show that the genetic algorithm was a skillful and effective tool in the selection of variables, and its approach with RF, concatenate in good responses in modeling individual height in forest stands.

Keywords: Feature selection. Random forest. Genetic algorithm.

1 INTRODUÇÃO

Na realização de inventários florestais, o emprego de relações hipsométricas constitui-se uma ferramenta fundamental para otimizar a coleta dos dados e assegurar sua precisão. A hipsometria estuda a relação direta entre o diâmetro das árvores e a sua altura, e por isso são tão úteis ao inventário florestal. A relação funcional entre essas variáveis é seguramente uma das mais estudadas na área de manejo florestal e inventário nos últimos anos. O ponto marcante dessa área do conhecimento é a possibilidade de se trabalhar com modelos específicos e pontuais, como os modelos tradicionais, bem como o uso de inteligência artificial, que trabalham de forma mais ampla os dados em diversas regiões. O dilema formado, entre tradicionais e inteligência artificial, é uma fonte rica de discussões no setor florestal. A prática de comparação entre diferentes abordagens e metodologias é recorrente nas pesquisas de inventário florestal. Contudo, um método bem difundido e utilizado é o *Random Forest* (RF), que vêm sendo aplicado nos últimos tempos com sucesso na modelagem florestal (HONG et al., 2018, REIS et al., 2019, SILVEIRA et al., 2019).

O RF é um método extremamente robusto e com ótima qualidade de predições, comumente utilizado em problemas de classificação (CARVALHO et al., 2017) e regressão, mesmo considerando um elevado número de dados e variáveis (CAO et al., 2018; JIN et al., 2018; ZHU; LIU, 2015). O método permite explorar um grande espaço de busca, retirando as informações mais relevantes na predição da variável resposta, mas ainda há possibilidade de melhoria por meio da seleção de variáveis. Essa abordagem é essencial quando se trabalha com grandes bancos de dados em macrorregiões, impactando no uso do método para as predições. A seleção de variáveis é uma temática desafiadora e atual, sendo aberta ao desenvolvimento de inúmeros métodos para potencializar o poder preditivo final, ou explicativo. Esse aspecto pode ser alcançado quando se trabalham métodos recursivos de redução de variáveis (ABDOH; ABO RIZKA; MAGHRABY, 2018), PCA (*principal component analysis*) (GEETHA et al., 2019), ou na pior das hipóteses, via tentativas manuais de seleção das melhores.

Porém, a maioria dos métodos de seleção são estáticos. Logo, uma alternativa que vêm sendo difundida para superar esta questão consiste na hibridação de métodos. Bader-El-Den e Gaber, (2012) evidenciaram o potencial da hibridação na melhoria da performance do RF, uma vez que, associado ao algoritmo genético para mudar dinamicamente as árvores na floresta, resultou no aumento da precisão do modelo. A excelência do GARF para seleção de características também é comprovada nos estudos de Cerrada et al., (2016), Crisman et al. (2016), Ma e Fan (2017), Paing e Choomchuay (2018).

Neste sentido, a qualidade das variáveis incorporadas ao modelo é fundamental para o alcance de um bom desempenho preditivo. No passado, devido ao número reduzido de variáveis disponíveis trabalhava-se apenas a sua transformação em escala, tendo efeitos positivos como os observados utilizando o logaritmo neperiano (HOLMGREN, 2004; SEGURA; KANNINEN, 2005). Abordagens de seleção de variáveis é uma temática desafiadora e atual, sendo aberta ao desenvolvimento de inúmeros métodos para potencializar o poder preditivo final, ou explicativo. Logo, no desenvolvimento de modelos de predição aplicadas ao manejo florestal, diversos pesquisadores empenham-se no sentido de incorporar novas variáveis, principalmente ambientais (SCOLFORO et al., 2016; SCOLFORO et al., 2017), variáveis em escala regional/global (SILVEIRA et al., 2019), e geográficas (GUGGER et al., 2018). A maior frequência envolve o uso de variáveis espectrais, oriundas de dados de sensoriamento remoto. Estudos consistentes demonstram que valores de reflectância do dossel da vegetação permitem relacionar valores espectrais a variável dependente altura, trazendo bons resultados, até mesmo quando comparados com métodos tradicionais (LARIBI et al., 2018; LEE et al., 2018; SCHLUND et al., 2019; URBAZAEV et al., 2018).

Assim, os objetivos do estudo foram avaliar o efeito da seleção de variáveis pelo algoritmo genético na melhoria da acurácia do método *Randon forest*; utilizar o modelo de regressão como critério comparativo entre os métodos; identificar as variáveis mais correlacionadas com a altura das árvores pelos métodos testados; explicar o gradiente de altura das árvores na bacia hidrográfica do rio grande.

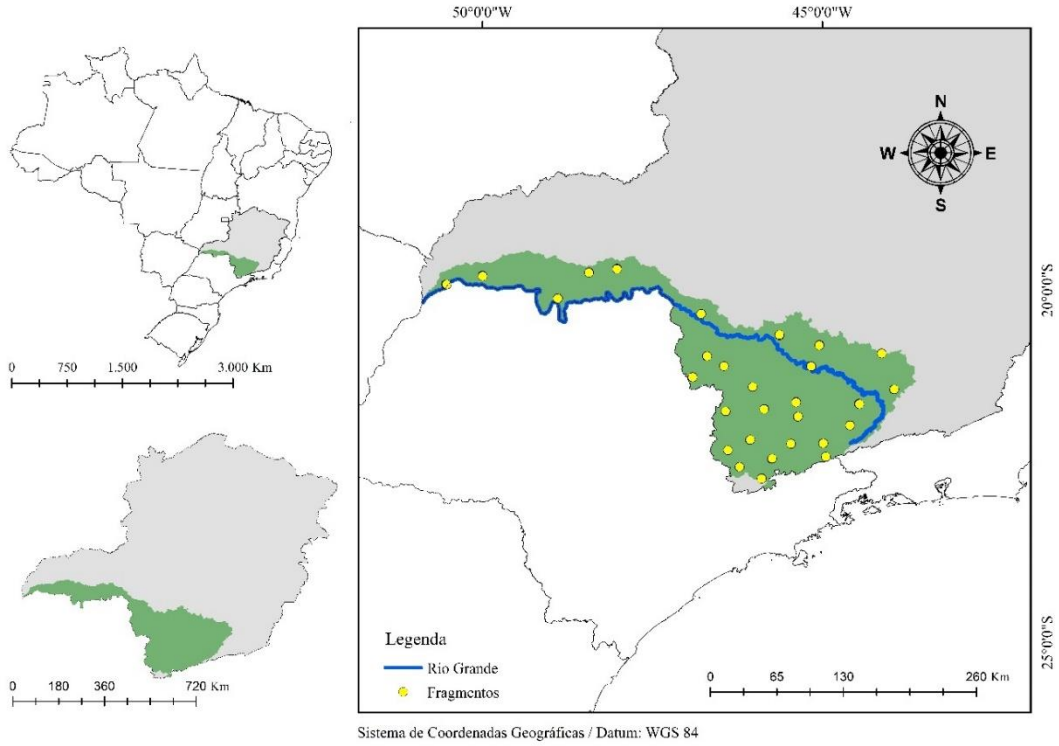
2 MATERIAL E MÉTODOS

2.1 Área de estudo

Os dados utilizados foram obtidos em uma área de estudo referente à Bacia Hidrográfica do rio Grande na porção do estado de Minas Gerais (Figura 1), com área de 86.050,29 km², ocupando aproximadamente 15% da área total do Estado de Minas Gerais. A Bacia do rio Grande apresenta classificação climática de Köppen do tipo Cwa (verão quente), Cwb (verão temperado) e Aw (verão quente e chuvoso) (ALVARES et al., 2013). O tipo do solo predominante são os Latossolos, com mosaicos de neossolo litólico, argissolo e cambissolo (CURI et al., 2008). De acordo com a disponibilidade das áreas em relação à composição da vegetação remanescente, o domínio Mata Atlântica recebe destaque. Essa cobertura vegetal

nativa da bacia hidrográfica representa aproximadamente 16% do seu território (SCOLFORO et al., 2008).

Figura 1 - Mapa de localização da bacia hidrográfica do rio Grande, com informação da fisiografia local e cobertura da terra.



Fonte: Do autor (2020).

Após a definição dessa macrorregião foi realizada uma amostragem dos fragmentos presentes nessa fisionomia. Nos fragmentos foram lançadas parcelas de 250 m^2 em conglomerado. Ao todo, Vinte e oito fragmentos foram inventariados, variando a intensidade amostral conforme área e local. As variáveis coletadas foram o diâmetro a altura do peito e a altura total, bem como a identificação botânica. As árvores medidas possuíam um DAP superior a 5 cm e compuseram a classe de variáveis populacionais, como área basal e densidade (N/ha).

2.2 Variáveis ambientais

A grande complexidade de ambientes presentes na bacia do rio Grande exige a busca por variáveis que contribuam pela melhoria das estimativas. Nesse sentido, o uso de variáveis climáticas e ambientais foram adotadas como estratégia de melhoria da acurácia dos modelos. O preparo de informações correlacionadas as informações ambientais, se deu em diferentes ambientes de trabalho, desde análise química e estrutural do solo, coletadas amostras por

parcelas e extrapoladas para hectare, à informações geográficas, topográficas, climáticas e sensoriais para as diferentes áreas amostrais da bacia.

Na obtenção de variáveis climáticas optou-se pelo uso dos dados meteorológicos globais do *WordlClim* versão 1.4 (worldclim.org). Essas derivam da compilação de séries históricas de dados climáticos do globo terrestre (1950-2000) em período mensal envolvendo a temperatura e precipitação interpolados. Assim foram obtidas 19 variáveis que representam a média, mínimo, máximo e variação da temperatura e precipitação, com resolução aproximada de 1km. Informações geográficas de latitude (Y) e longitude (X) foram coletadas a partir do centróide das parcelas (Tabela 1).

A aquisição dos valores espectrais do povoamento se deu em ambiente de geoprocessamento, com a obtenção dos valores através de imagens do satélite Landsat 8 OLI (30m de resolução) e MODIS (*Moderate Resolution Spectroradiometer*) (com resolução variável entre 250 a 1.000 m), adquiridos dentro do intervalo de tempo do inventário florestal (Tabela 1). Através de imagens do sensor Lansat OLI foram adquiridos índices de 12 cenas oriundas do *United States Geological Survey of Earth* (USGS/EROS) com as correções geométricas e radiométricas. Nos mosaico das imagens foram obtidos 7 índices de vegetação: NDVI (ROUSE et al., 1973); NDMI (WILSON; SADER, 2002); EVI (JUSTICE et al., 1998); SAVI (HUETE, 1988); mSAVI - (QI et al., 1994); NBR - (MILLER; THODE, 2007); NBR2 (MILLER; THODE, 2007). O modelo digital de elevação SRTM foi gerado com resolução espacial de 100 m, calculadas pela ferramenta *Terrain Analysis* do software SAGA GIS (v. 6.3.0). Através do sensor MODIS foram extraídas 12 variáveis morfométricas, as quais são relacionadas à temperatura da superfície da Terra (emis31, emis32, lstd, lstn), atividade fotossintética (fpar, lai), evapotranspiração (et, le, pet, ple), produtividade primária (gpp) e porcentagem de cobertura vegetal (treecover).

As variáveis extraídas das amostras de solo representam 130 atributos que se deram através de análise estrutural química e física, sendo: pH, potássio, fósforo, cálcio, magnésio, alumínio, acidez potencial (H + Al), soma das bases, CTC (capacidade de troca de cátion) efetiva, CTC a pH 7, saturação por bases, saturação por alumínio, matéria orgânica, fósforo remanescente, zinco, ferro, manganês, cobre, boro, enxofre, porcentual de argila, silte e areia, densidade do solo, teor de carbono e estoque de carbono no solo (Tabela 1). As amostras representam diferentes horizontes de profundidade do solo das parcelas (A - 0 a 10 cm, B - 10 a 20 cm, C - 20 a 40 cm, D - 40 a 60 cm e E - 60 a 100 cm).

Tabela 1 - Variáveis independentes utilizadas na modelagem da altura das árvores.

TIPO	PREDITORES	RESOLUÇÃO (m)/UNIDADE
Climática	BIO1 - Annual Mean Temperature, BIO2 - Mean Diurnal Range, BIO3 - Isothermality, BIO4 - Temperature Seasonality, BIO5 - Max Temperature of Warmest Month, BIO6 - Min Temperature of Coldest Month, BIO7 - Temperature Annual Range, BIO8 - Mean Temperature of Wettest Quarter, BIO9 - Mean Temperature of Driest Quarter, BIO10 - Mean Temperature of Warmest Quarter, BIO11 - Mean Temperature of Coldest Quarter, BIO12 - Annual Precipitation, BIO13 - Precipitation of Wettest Month, BIO14 - Precipitation of Driest Month, BIO15 - Precipitation Seasonality, BIO16 - Precipitation of Wettest Quarter, BIO17 - Precipitation of Driest Quarter, BIO18 - Precipitation of Warmest Quarter, BIO19 - Precipitation of Coldest Quarter	1000
Morfométrica	altitude - Altitude, hillshadin - Analytical hillshading, aspect - Aspect, cn_base_le - Channel network base level, conv_index - Convergence index, c_sec_curv - Cross sectional curvature, dif_insol - Diffuse insolation, direct_ins - Direct insolation, flow_accum - Flow accumulation, long_curv - Longitudinal curvature, ls_factor - LS factor, relative_s - Relative slope, valley_dep - Valley depth, vert_dist - Vertical distance, wet_index - Wetness index, slope_perc - Slope (%)	100
Espectral Landsat	evi - EVI, ndvi - Normalized Difference Vegetation Index, msavi - Modified Soil-adjusted Vegetation Index, nbr - Normalized Burn Ratio, nbr2 - Normalized Burn Ratio 2, ndmi - Normalized difference moisture index, savi - Soil-adjusted Vegetation Index	30
Espectral MODIS	emis31 - Emissivity bands 31, emis32 - Emissivity bands 32, et - Global evapotranspiration, fpar - Fraction of photosynthetically active radiation, le - Latent heat flux, lstd - Land Surface Temperature day, lstn - Land Surface Temperature night, pet - Potential global evapotranspiration, ple - Potential latent heat flux, treecover - Percent Tree Cover, gpp - Gross Primary Production, lai - Leaf Area Index	emis31 - 1000, emis32 - 1000, et - 500, fpar - 500, le - 500, lstd - 1000, lstn - 1000, pet - 500, ple - 500, treecover - 250, gpp - 500, lai - 500,
Geográfica	X - Longitude, Y - Latitude,	-
Solo*	pH - Acidez ou basicidade do solo, K - Potássio, P - Fósforo, Ca - Cálcio, Mg - Magnésio, Al - Alumínio, H+Al - Acidez potencial, SB - Soma de Bases, t - CTC efetiva, T - CTC a pH 7, V - Saturação por Bases, m - Saturação por Alumínio, M.O. - Matéria Orgânica, P-Rem - Fósforo Remanescente, Zn - Zinco, Fe - Ferro, Mn - Manganês, Cu - Cobre, B - Boro, S - Enxofre, Argila - Relação da análise física de Argila, Silte - Relação da análise física de Silte, Areia - Relação da análise física de Areia, Dens_solo - Densidade do solo, TeorC - Teor de Carbono, EstoqueC - Estoque de carbono	pH - categórico, K - mg/dm ³, P - mg/dm ³, Ca - cmol c/dm³, Mg - cmol c/dm³, Al - cmol c/dm³, H+Al - cmol c/dm³, SB - cmol c/dm³, t - cmol c/dm³, T - %, V - %, m - %, M.O. - dag kg-1, P-Rem - mg/L, Zn - cmol c/dm³, Fe - cmol c/dm³, Mn - cmol c/dm³, Cu - cmol c/dm³, B - cmol c/dm³, S - cmol c/dm³, Argila - dag kg-1, Silte - dag kg-1, Areia - dag kg-1, Dens_solo - g/cm ³, TeorC - %, EstoqueC - Mg/ha

* Variáveis distribuídas nos horizontes “A”, “B”, “C”, “D” e “E” do solo.

Fonte: Do autor (2020).

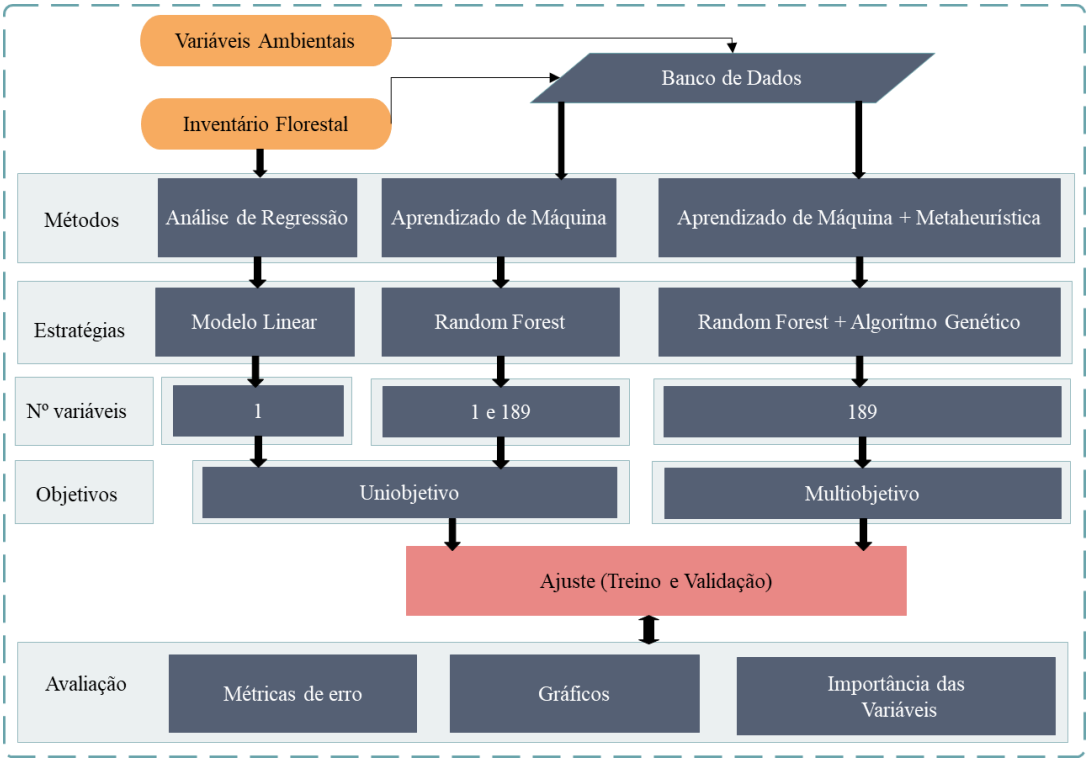
2.3 Padronização espacial da base de dados

A variável dependente a ser predita é a altura total das árvores, que foi relacionada com as demais extraídas para compor o banco de dados. Devido a continuidade espacial das variáveis, estas foram padronizadas em um conjunto de grids (100 x 100m), sendo obtido a média de seu valor para a centroide. Logo, para as áreas em estudo foram obtidas informações de 154 parcelas com valores médios de 56 variáveis independentes, do tipo climática, morfométrica, espectral Landsat e espectral MODIS. Soma-se a base de dados valores médios da estrutura química e física de solo das parcelas inventariadas. De modo complementar, variáveis populacionais foram incluídas como o diâmetro à altura do peito (DAP), e os valores populacionais de área basal (G) e número de indivíduos por hectare (N/ha). Ao término, o banco de dados ficou composto por 189 variáveis distribuídas em 5.608 árvores. Deste total, 80% dos indivíduos foram selecionados e fixados dentro de cada parcela para constituir o banco de dados de treino, com uma soma de 4427 árvores, os outros 20% constituíram a base validação, com 1181 árvores distribuídas de forma proporcional para cada parcela.

2.4 Modelagem matemática da altura individual das árvores

A relação diâmetro altura é fortemente usada na área florestal para obter a altura individual das árvores, e os modelos de regressão são consolidados na obtenção desta estimativa, porém, métodos de aprendizado de máquina veem sendo utilizados para reduzir custos operacionais e melhorar suas estimativas, assim como, a inclusão de variáveis ambientais que expliquem este atributo. Deste modo, adotou-se diferentes abordagens para a modelagem matemática da altura, desde regressão à algoritmos de aprendizagem máquina. A figura 2 traz um fluxograma dos métodos e estratégias adotados para estimar a altura individual das árvores.

Figura 2 - Fluxograma dos métodos e estratégias adotados para estimar a altura individual das árvores.



Fonte: Do autor (2020).

2.4.1 Analise de regressão

Existem na literatura uma série de modelos hipsométricos tradicionais, ou seja, que apresentam apenas a relação diâmetro (DAP) e altura. Contudo, a qualidade do ajuste depende de uma série de fatores, como idade, qualidade do sítio, espécie, dentre outros. Assim, previamente realizou-se uma série de testes para a seleção do modelo de regressão a ser adotado, chegando-se ao modelo parabólico. O modelo de regressão tradicional foi ajustado por parcela, e validado para a mesma parcela (Equação 1). Em que: H_i= altura total em metros; *DAP*= Diâmetro a altura do peito em centímetros; β_i= Parâmetros do modelo a ser estimado.

$$H_i = \beta_o + \beta_1 DAP + \beta_2 DAP^2 + \varepsilon_i \quad (1)$$

2.4.2 *Random Forest*

O *Random Forest* (RF) é um método de aprendizagem de máquina que utiliza uma combinação de árvores de decisão individuais para a formação de uma floresta na resolução de problemas (BREIMAN, 2001). O algoritmo apresenta uma série de particularidades que o torna

extremamente viável na resolução de problemas, como: a natureza aleatória de construção de cada árvore minimiza o sobreajuste; ocorre a geração de métricas de erro internas e de importância das variáveis preditoras; parametrização simples, e; se ajusta com grandes bancos de dados e muitos atributos. Como qualquer algoritmo ele exige uma parametrização inicial, a sua parametrização considerou um número de árvores (*ntrees*) em 500 unidades e variáveis (*mtry*) igual a 2. Utilizou-se o pacote *randomForest* (LIAW; WIENER, 2002) na linguagem de programação R (R CORE TEAM, 2019).

A aplicação do *Random Forest* considerou 3 estratégias distintas para estimar a altura das árvores, e diferente do modelo de regressão, que foi ajustado por parcela, os modelos de RF foram ajustados para a base toda, separadas em treino e validação. A primeira, prediz a altura em função da variável diâmetro à altura do peito (DAP) (RF_1). Para a segunda estratégia, pressupõe-se o seu uso puro, portanto, para 189 variáveis (RF_2). E por último, a introdução de um algoritmo híbrido para seleção de variáveis e predição do modelo de altura. Deste modo, utilizou-se o algoritmo genético (AG) na seleção de variáveis como entradas no modelo de *Random Forest* (GARF).

A implementação do algoritmo genético (AG) para seleção de recursos foi estabelecida após testes de parametrização. Os testes preliminares envolveram operadores de seleção (torneio), *crossover* (um ponto de corte) e mutação (bit aleatório), critério de parada (50 gerações), bem como o tamanho da população (100 indivíduos). A diversidade da população se manteve através do operador de mutação, com 10% de probabilidade de os indivíduos terem mutação, e 50% para troca aleatória dos genes. Os indivíduos da população foram dimensionados para um vetor fixo de 189 posições (genes). Os genes podendo assumir valores de 0 ou 1, 0 desativa, e 1 ativa a variável para formar o indivíduo que vai corresponder como entrada no modelo de RF.

Para a avaliação do GARF, buscou-se avaliar dois aspectos do algoritmo híbrido, o primeiro a minimização do erro médio quadrático (*out-of-bag* - OOB) da estimativa da altura, e depois, a redução do número de variáveis preditoras, portanto, a função de avaliação (*fitness*) possui uma natureza multiobjetiva. Assim, a função *fitness* consiste na soma de dois termos, o primeiro a razão entre o erro OOB das variáveis habilitadas pelo AG e o erro máximo possível (calculado através de testes preliminares). A segunda é dada pela razão entre número de variáveis habilitadas pelo AG (n) e pelo número total de variáveis testadas no experimento.

$$fitness = \left(\frac{erroOOB}{erroOOB_{\max}} + \frac{n}{NVT} \right) \quad (2)$$

2.5 Critérios de avaliação dos métodos

Para a avaliação das estimativas dos métodos testados utilizou-se estatísticas que refletissem a capacidade preditiva. Assim, considerando o conjunto de treinamento e validação calculou-se o coeficiente de determinação (R^2) (3), bias (%) (4) e a raiz do erro quadrado médio percentual (*Root Mean Square Error* – RMSE%) (5), sendo $i =$ número da instância; $n =$ número total de observações do conjunto de dados (treinamento 4.427 observações e validação 1.181); $Y_i =$ valor observado da altura (m); $Y_i =$ valor médio estimado de altura (m) da observação; $\overline{Y}_i =$ valor médio da altura observada (m). Para critério de comparação, foi adotado para o modelo de regressão o somatório das métricas. De forma complementar, adotou-se uma análise gráfica (dispersão dos erros e correlação) para avaliar tendências das estimativas.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{3}$$

$$Bias(\%) = \frac{\sum_{i=1}^{n}(Y_i - Y_i)}{n} \tag{4}$$

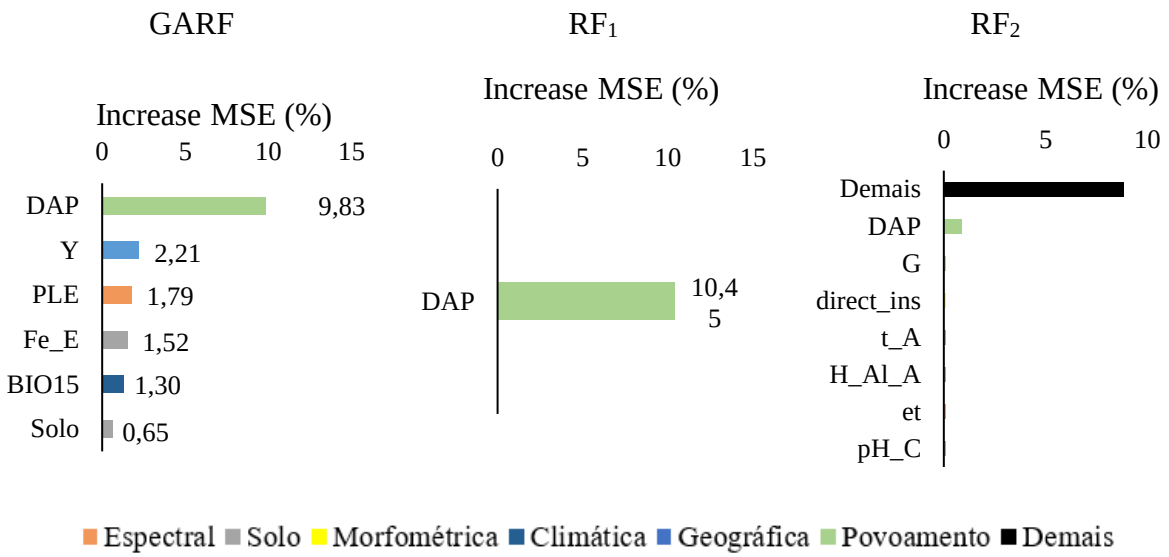
$$RMSE(\%) = \frac{100}{\overline{Y}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y)^2}{n}} \tag{5}$$

3 RESULTADOS

Uma análise exploratória da base de dados mostra valores heterogêneos da variável em estudo, com amplitude de 25 metros entre o mínimo e o máximo em altura individual. Além disso, dentro das parcelas amostrais o coeficiente de variação médio foi de 38,12%. Esses valores indicam uma grande dificuldade na obtenção de bons métodos de predição, já que usualmente em florestas tropicais esse padrão é recorrente. Contudo, apesar das dificuldades encontradas, os métodos testados foram capazes de obter valores compatíveis com a literatura. Inicialmente, avaliando os métodos via treinamento de máquina, observa-se que a variável diâmetro foi a mais importante, sendo mais potencializada nas condições de RF multiobjetivo e RF em função somente de DAP. O seu valor passou de 0,89 (RF_2) a 9,83% (GARF) graças

ao poder de seleção do algoritmo genético, e quando comparada somente em função do DAP (RF_1), o valor de importância foi de 10,45% (Figura 3). Destacando ainda uma troca na ordem de importância entre os métodos. O uso de uma metodologia que abordasse aspectos multiobjetivos, permitiu ao algoritmo GARF a redução do número de variáveis de 86 para 6, com queda substancial de 83%. Ao longo do processo de evolução do algoritmo, a minimização do erro apresentou uma redução percentual de 65%, com valores começando em 1,27, e terminando em 0,45 na última iteração. As variáveis selecionadas para a condição multiobjetivo foram: DAP (diâmetro à altura do peito), Y (latitude), PLE (*Potential latent heat flux*), Fe_D (quantidade de ferro na camada de solo do horizonte D), BIO15 (*Precipitation Seasonality*) e Mg_D (quantidade de magnésio na camada de solo do horizonte D). A variável geográfica latitude apresentou o segundo maior valor de importância, com 2,21% de erro. As demais variáveis selecionadas variando entre os grupos espectral, solo e climáticas com valores abaixo de 2.

Figura 3 - Valores médio de importância das variáveis explicativas da altura das árvores no treinamento dos métodos testados.



Em que: GARF: algoritmo genético + *random forest*; RF_1: *random forest* somente com DAP; RF_2: *random forest* com todas as variáveis.

Fonte: Do autor (2020).

A avaliação dos modelos de predição, como pode ser observado para as ambas bases de estudo (Tabela 2), apresenta uma ligeira vantagem para o modelo GARF diante dos demais, ele é seguido pelos modelos de regressão e RF_1. O pior desempenho se deu no modelo de RF_2,

como pode ser observado pelas estatísticas de precisão na tabela 2. Para a base de treino os valores de RMSE% apresentam os melhores resultados para o modelo GARF, com aumentos de 4,33%, 6,79% e 16,53%, respectivamente para regressão, RF_1 e RF_2. Já para a validação, os percentuais de aumento foram menores, com valores de 2,12%, 6,19% e 11,92%, respectivamente para regressão, RF_1 e RF_2.

Tabela 2 - Métricas de avaliação para as diferentes metodologias para dados de treinamento e validação.

Base	Método	RMSE (m)	RMSE (%)	Bias (m)	R^2 (%)	N
Treino	GARF	1,25	14,83	-0,0050	84	6
	Regressão	1,61	19,15	0,0001	74	2
	RF_1	1,82	21,61	0,0088	67	2
	RF_2	2,64	31,36	-0,0257	30	189
Validação	GARF	1,96	22,79	0,7428	68	6
	Regressão	2,14	24,91	0,4731	62	2
	RF_1	2,49	28,97	1,2310	49	2
	RF_2	2,98	34,71	1,8731	26	189

Em que: GARF: algoritmo genético + *random forest*; RF_1: *random forest* somente com DAP; RF_2: *random forest* com todas as variáveis.

Fonte: Do autor (2020).

Já para os valores de coeficiente de determinação, as melhores respostas também se deram para o modelo GARF, com explicação de 84% para o treino, e 68% para base de dados de validação. Para o bias, as melhores estatísticas de erro podem ser observadas no modelo de regressão, com 0,0001 m para o treino e 0,4731 m para a validação.

Os resultados demonstram que o uso do GARF é o mais apropriado para estimar a altura individual das árvores em povoamentos inequiâneos, devido ao seu método de seleção de variáveis. Também se pode observar pela tabela 2 que a regressão demonstrou um desempenho superior aos modelos de RFs (RF_1 e RF_2), isso significa que utilizar todas as variáveis ou apenas o DAP não é uma estratégia viável, o modelo de regressão supre a necessidade de utilizar um método tão robusto. Contudo, a aplicação do GARF se justifica, ele é capaz de selecionar e associar somente variáveis que mais influenciam na determinação da altura individual das árvores, além de diminuir consideravelmente os erros de estimativas. Veja que com a introdução e a seleção de variáveis ambientais no modelo GARF, ajudou a trazer uma explicação ecológica a previsão da altura, correlacionando a fatores como luminosidade, nutrientes disponíveis para as plantas, água disponível ao longo do ano e temperatura, ficando

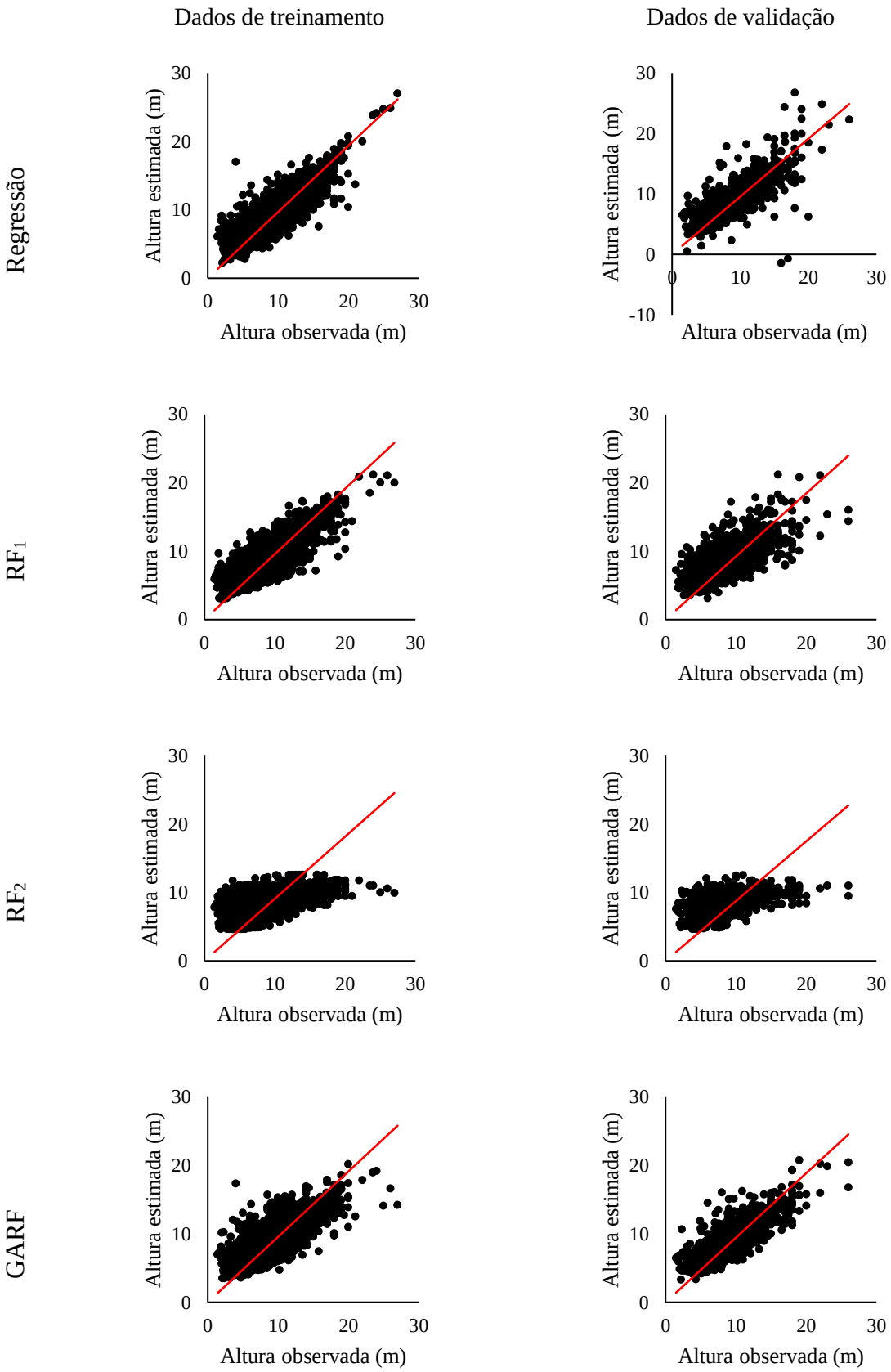
intrínseco que não somente variáveis empíricas altamente correlacionadas ajudam a explicar a altura.

Ao considerar-se a distribuição gráfica da altura estimada versus altura observada (Figura 4), e as diferenças entre as estatísticas de precisão (Tabela 2), observa-se o quanto a redução do número de variáveis foi eficiente na determinação da altura.

A linha reta entre o valor observado e o valor previsto é o fator determinante do padrão de avaliação dos modelos. A regressão, RF_1 e GARF, mostram uma variação uniforme do valor predito em relação a altura observada para o treino. Estes também apresentaram uma relação uniforme para a validação, apesar da regressão subestimar alguns indivíduos. Por outro lado, o RF_2 demonstra uma distribuição tendenciosa e achatada em relação a linha de tendência.

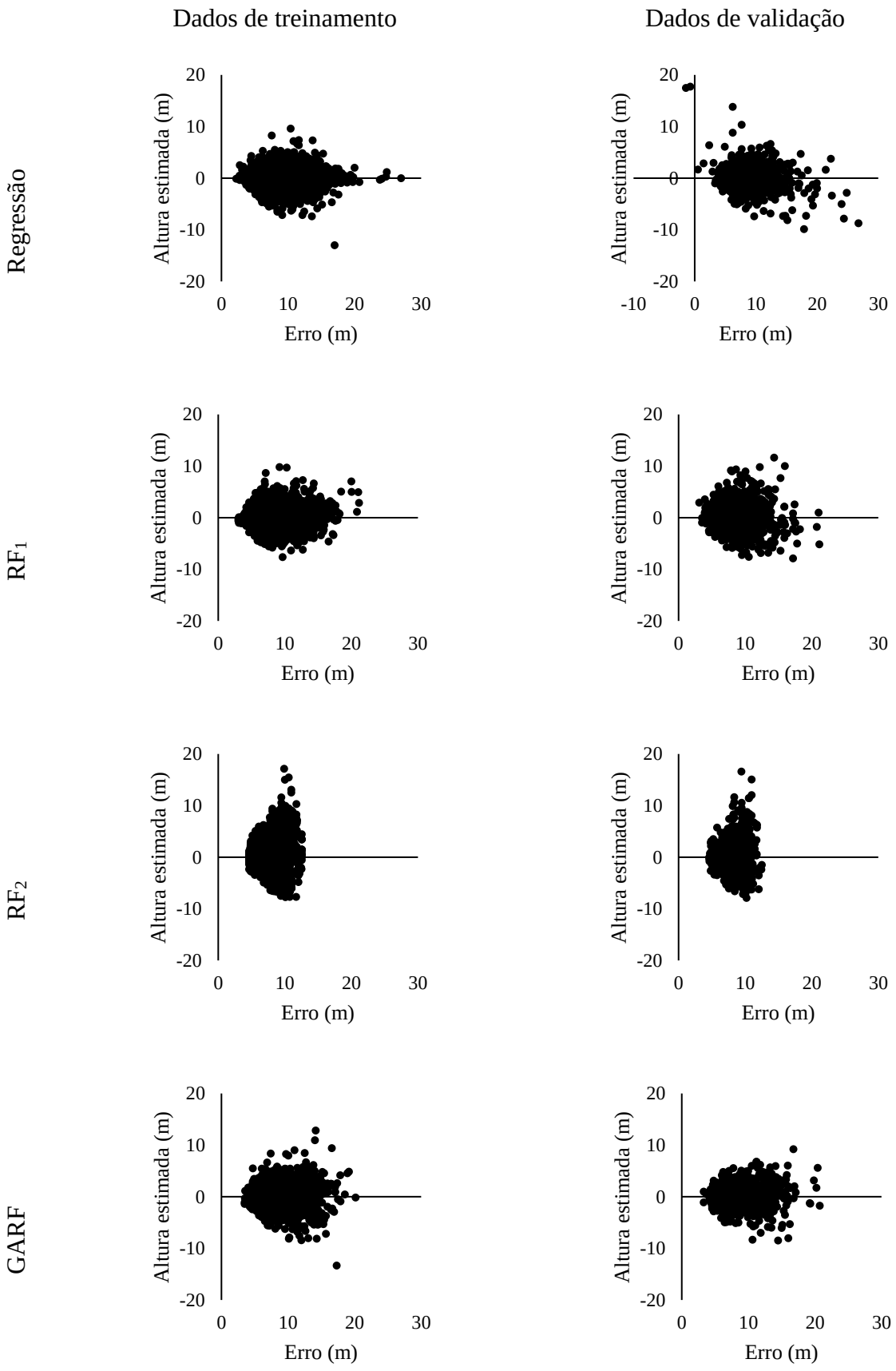
Ao analisar os gráficos de dispersão dos resíduos, observa-se para todos os modelos avaliados que os erros se concentraram no intervalo de 10 e -10 m, independente do banco de dados (Figura 5). O modelo GARF e RF_1 apresentaram a melhor dispersão dos resíduos, com valores de erro uniforme em relação à altura estimada, tanto para o treino, quanto para a validação. O mesmo comportamento pode ser observado para a regressão com dados de treinamento, mas para o conjunto de validação alguns indivíduos apresentaram valores subestimados, isso está associado a parcela em que foi ajustada, a regressão não conseguiu captar toda a variabilidade da altura intrínseca nela, e consequentemente prejudicou-se ao validar. O RF_2 foi ligeiramente o pior modelo, subestimando e superestimando em muito as estimativas da altura, e sendo ineficaz em estimar indivíduos com altura superiores a 12 metros. O uso do RF com todas as variáveis demonstrou não ser uma boa opção, pois o algoritmo não foi capaz de pesar a importância das variáveis ao relacionar a altura. Contudo, a implementação de um algoritmo genético permitiu ao método correlacionar variáveis que permitisse uma melhoria das estimativas.

Figura 4 - Gráfico de correlação da altura observada pela altura estimada pelos modelos testados.



Fonte: Do autor (2020).

Figura 5 - Gráficos de dispersão do erro nas estimativas de altura pelos métodos testados.



Fonte: Do autor (2020).

4 DISCUSSÃO

A altura é uma importante medida em povoamentos florestais, altamente correlacionado com aspectos biológicos, e está ligada diretamente ao estoque de carbono e outros importantes atributos. Os resultados mostraram que a seleção prévia de variáveis exerce um efeito direto na qualidade dos métodos, sendo um ponto preponderante na modelagem. O *Randon Forest* é um algoritmo robusto e confiável, mas foi influenciado negativamente quando introduzido todas as variáveis do banco de dados. Por outro lado, o modelo de regressão ajustado por parcela tem uma vantagem frente aos demais, apesar de adotar dados de florestas tropicais com alta diversidade.

O banco de dados apresenta um grande número de variáveis, com diferentes escalas, desde variável individual (DAP), a variáveis de macroescalas (climáticas, espectral), esta diferença compromete o resultado final. O RF é um algoritmo sensível a quantidade de amostras inseridas neles, porém uma quantidade de observações menor torna-o insensível a redução do número de variáveis, como pode ser observado no trabalho de Ma et al. (2017). Mas ele se beneficia muito de um tamanho de amostra maior (FASSNACHT et al., 2014; LI et al., 2016), o que pode ser observado também no trabalho em questão. A aplicação de um algoritmo de seleção multiobjetivo permitiu identificar quais recursos realmente se auto correlacionam para prever a altura das árvores, deste modo, reduzir o número de variáveis e melhorar as estimativas. Este número reduzido de recursos facilita o entendimento do algoritmo.

Por mais que existem métodos mais usuais para redução de variáveis, como a redução recursiva do valor de importância das variáveis para o RF, o algoritmo genético permite varrer um espaço de busca maior, associar variáveis que poderiam dar boas respostas e que por ventura seriam eliminadas pelo método recursivo. Abordagens híbridas são constantemente utilizadas com modelos de aprendizado de máquinas, apesar do maior tempo de processamento, mas na maioria dos trabalhos ela demonstra uma otimização dos resultados. Hong et al. (2018), aplicando algoritmo genético na seleção de variáveis, e *Randon Forest* na classificação de áreas susceptíveis a incêndios florestais, indicam a eficiência do método. Paul et al. (2017), com uma abordagem diferente, testaram 5 métodos diferentes de seleção de variáveis para classificar a previsão de resultados em câncer de esôfago com RF, com os melhores resultados para o AG. Os autores destacam a performance do algoritmo na seleção de recursos, onde suas abordagens

com RF potencializam a resposta. Estes resultados corroboram com a afirmação de que o GARF é um método promissor na redução dos erros e número de variáveis.

Em relação às variáveis que mais contribuíram para estimativas da altura individual, nota-se maior contribuição das variáveis diâmetro à altura do peito, que está altamente correlacionada com a altura, seguidas pelas variáveis geográfica (latitude), espectral MODIS (*potential latent heat flux*), solo (Fe e Mg) e climática (*precipitation seasonality*). A seleção da variável individual DAP já era esperada, pois ela está altamente correlacionada com a variável resposta e intrínseca a cada indivíduo. Modelos de altura diâmetro são usualmente objeto de estudos, dada a sua importância em modelos empíricos da área (BARBOSA et al., 2019; SCHRODER; FINGER, 2016; FERRAZ FILHO et al., 2018). A associação da variável DAP a outras variáveis elevou a capacidade preditiva do algoritmo, todas apresentando um processo biológico que pode ser ligado a estimativa da altura.

Observa-se para os padrões de deposição geográfica e altura média uma grande variabilidade para o gradiente latitudinal, a variável carrega consigo fatores bioclimáticos locais intrínsecos a ela, principalmente precipitação, solo e temperatura, que influenciam diretamente no crescimento das árvores. Os gradientes geográficos (latitude, longitude e altitude) comumente são identificadas como características importantes responsáveis por padrões biogeográficos que afetam a associação clima-crescimento (ALTMAN et al., 2017; GALVÁN; CAMARERO; GUTIÉRREZ, 2014; MÄKINEN et al., 2002; PRIMICIA et al., 2015). A seleção desta variável ajuda a explicar o comportamento da altura, associando a ela que determinadas localidades apresentam maior ou menor estatura dos indivíduos. Zhang et al. (2016) trabalharam com gradientes ambientais para detectar a altura global de dosséis de floresta, e observaram que as maiores alturas estão mais próximas a linha do Equador, e árvores de menor porte em latitude mais altas, o mesmo pode ser observado no presente estudo.

A sazonalidade da precipitação (BIO15) explica variações de precipitações ao longo do ano, deste modo, áreas que apresentam maior uniformidade de precipitação, tinham por média maiores valores de altura. Intervalos mais regulares de chuva atribui uma disponibilidade mais uniforme de água as árvores, o que favorece o crescimento dos indivíduos do povoamento. Ela vai de encontro aos estudos de Pompa-García e Jurado (2014) sobre crescimento de anéis de *Pinus cooperi*, onde apontaram que a sazonalidade da precipitação impactou diretamente em seu crescimento. Deb et al. (2017), ao estudarem a distribuição da Teca (*Tectona grandis*) na Ásia Tropical, também tiveram no conjunto final a variável, e indicaram que a precipitação

sazonal é um dos principais fatores na distribuição da espécie, isso insere a importância da seleção desta variável no modelo, ela está ligada diretamente a processos biológicos.

A inclusão de parâmetros, físicos, químicos e climáticos, normalmente são incluídos em modelos de regressão, no nosso caso Fe e Mg, estas atribuições desempenham um papel fundamental na qualidade de ajuste da altura. A seleção destas variáveis, que estão associadas a fatores ambientais do terreno, favorece a explicação do tamanho dos indivíduos. Ela corrobora com modelos para qualidade e produtividade de solo, onde também se observa a seleção de atributos do solo (BUEIS et al., 2016, 2017; BUEIS; TURRIÓN; BRAVO, 2019; GARTZIA-BENGOETXEA et al., 2009). Valores de Mg e Fe apresentaram valores inversamente proporcionais entres as parcelas amostradas, então pode-se associar que em determinadas áreas ocorreram espécies que exigiam mais ou menos de tais nutrientes. No modelo de RF, essas informações funcionaram como um filtro, ele associou que aquela área apresenta espécies de maior ou menor estatura. A inclusão de variáveis de solo também pode ser observada para componentes arbóreos e diversidade florística (HIGUCHI et al., 2012), fator preponderante para filtros biológicos na determinação da altura.

O uso de métricas de dados de sensoriamento remoto são exploradas continuamente para mapear e predizer atributos da floresta, isso se dá porque existe uma forte correlação entre a altura e a cobertura vegetal. Assim, dados ambientais obtidos por sensoriamento remoto contribuem de forma significativa na sua explicação. Trabalhos como de Wang et al. (2016) e Huang et al. (2017), em que mapearam a altura da floresta usando produtos percentuais de cobertura de árvores, utilizando produtos MODIS, reafirmam esse ponto, e, portanto, o porquê da seleção do atributo PLE (*potential latent heat flux*). O fluxo de calor latente refere-se à energia que é limitada pela evapotranspiração da água, é uma importante medida para a regulação do clima (BECK et al., 2011; BONAN, 2008; NAUDTS et al., 2016; THOM; RAMMER; SEIDL, 2017). Assim como para as outras variáveis, o uso do PLE funcionou como um filtro, associando espécies e áreas de maior fluxo de calor latente a indivíduos mais altos.

O algoritmo genético se mostrou uma ótima ferramenta na seleção de características que ajudasse a explicar altura individual das árvores, ajudando o modelo RF a transformar tais informações em ótimas estimativas da variável dependente. O uso das variáveis de forma separada não funcionaria, mas a inclusão e seleção delas em um modelo de regressão como o RF atribui a elas uma importância na determinação da altura individual das árvores, elas trabalham de forma conjunta, além de associar o porte dos indivíduos da floresta a processos ecológicos.

5 CONCLUSÃO

Ao analisar as diferentes abordagens, observa-se um desempenho superior para o modelo GARF. A escolha de um método de seleção de variáveis com função multiobjetiva, se mostrou assertiva, o GARF foi capaz de identificar correlações expressivas entre os atributos selecionados e concatenar em respostas preditivas superiores aos métodos tradicionais e de aprendizado de máquina na sua forma pura. O GARF foi capaz de associar variáveis que apresentassem ligações ente si e conectá-las a processos biológicos, tal seleção ajudou a explicar a altura individual das árvores por fatores ecológicos. As variáveis selecionadas pelo algoritmo foram: DAP (diâmetro à altura do peito), Y (latitude), PLE (*Potential latent heat flux*), Fe_D (quantidade de ferro na camada de solo do horizonte D), BIO15 (*Precipitation Seasonality*) e Mg_D (quantidade de magnésio na camada de solo do horizonte D). A inclusão de variáveis ambientais como entradas no modelo RF trouxe estimativas mais precisas, então futuramente a aplicação de atributos ambientais como entrada no modelo RF pode tornar o método mais preciso para estimativa de atributos relacionadas à floresta em remanescentes de florestas nativas.

REFERÊNCIAS

ABDOH, S. F.; ABO RIZKA, M.; MAGHRABY, F. A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. **IEEE Access**, Piscataway, v. 6, p. 59475–59485, 2018.

ALTMAN, J. et al. Environmental factors exert strong control over the climate-growth relationships of Picea abies in Central Europe. **Science of the Total Environment**, New York, v. 609, p. 506–516, 2017.

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

BADER-EL-DEN, M.; GABER, M. GARF: Towards Self-optimised Random Forests. **International Conference on Neural Information Processing**. Springer, Berlin, Heidelberg, p. 506–515, 2012.

BARBOSA, R. I. et al. Allometric models to estimate tree height in northern amazonian ecotone forests. **Acta Amazonica**, Manaus, v. 49, n. 2, p. 81–90, 2019.

BECK, P. S. A. et al. The impacts and implications of an intensifying fire regime on Alaskan boreal forest composition and albedo. **Global Change Biology**, Oxford, v. 17, n. 9, p. 2853–2866, 2011.

BONAN, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. **Science**, Washington, v. 320, n. 5882, p. 1444–1449, 2008.

BREIMAN, L. Random Forests. **Machine Learning**, Boston, v. 45, p. 5–32, 2001.

BUEIS, T. et al. Relationship between environmental parameters and Pinus sylvestris L. site index in forest plantations in northern Spain acidic plateau. **IForest**, Potenza, v. 9, n. JUNE2016, p. 394–401, 2016.

BUEIS, T. et al. Site factors as predictors for Pinus halepensis Mill. productivity in Spanish plantations. **Annals of Forest Science**, Paris, v. 74, n. 1, 2017.

BUEIS, T.; TURRIÓN, M. B.; BRAVO, F. Stand and environmental data from Pinus halepensis Mill. and Pinus sylvestris L. plantations in Spain. **Annals of Forest Science**, Paris, v. 76, n. 2, 2019.

CAO, L. et al. Integrating airborne LiDAR and optical data to estimate forest aboveground biomass in arid and semi-arid regions of China. **Remote Sensing**, Basel, v. 10, n. 4, 2018.

CARVALHO, M. C. et al. Modelagem do nicho ecológicos de espécies arbóreas em uma área tropical brasileira. **Cerne**, Lavras, v. 23, n. 2, p. 229–240, 2017.

CERRADA, M. et al. Fault diagnosis in spur gears based on genetic algorithm and random forest. **Mechanical Systems and Signal Processing**, New York, v. 70–71, p. 87–103, 2016.

COSTA, E. A.; SCHRODER, T.; FINGER, C. A. G. Relação altura-diâmetro para araucaria angustifolia (Bertol.) kuntze no sul do Brasil. **Cerne**, Lavras, v. 22, n. 4, p. 493–500, 2016.

CRISMAN, T. J. et al. Identification of an efficient gene expression panel for glioblastoma classification. **PLoS ONE**, San Francisco, v. 11, n. 11, p. 1–19, 2016.

CURI, N. et al. **Zoneamento ecológico-econômico do Estado de Minas Gerais: componentes geofísicos e biótico**. 1. ed. Lavras: Editora UFLA, 2008.

DEB, J. C. et al. Climatic-Induced Shifts in the Distribution of Teak (Tectona grandis) in Tropical Asia: Implications for Forest Management and Planning. **Environmental Management**, New York, v. 60, n. 3, p. 422–435, 2017.

FASSNACHT, F. E. et al. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. **Remote Sensing of Environment**, New York, v. 154, n. 1, p. 102–114, 2014.

FERRAZ FILHO, A. C. et al. Height-diameter models for eucalyptus sp. plantations in Brazil. **Cerne**, Lavras, v. 24, n. 1, p. 9–17, 2018.

GALVÁN, J. D.; CAMARERO, J. J.; GUTIÉRREZ, E. Seeing the trees for the forest: Drivers of individual growth responses to climate in Pinus uncinata mountain forests. **Journal of Ecology**, New York, v. 102, n. 5, p. 1244–1257, 2014.

GARTZIA-BENGOETXEA, N. et al. Potential indicators of soil quality in temperate forest ecosystems: A case study in the Basque Country. **Annals of Forest Science**, Paris, v. 66, n. 3, 2009.

GEETHA, R. et al. Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. **Journal of Medical Systems**, Dordrecht, v. 43, n. 9, 2019.

GÓMEZ-GARCÍA, E. et al. Height-diameter models for maritime pine in portugal: A comparison of basic, generalized and mixed-effects models. **IForest**, Viterbo, v. 9, n. Feb 2016, p. 72–78, 2015.

GUGGER, P. F. et al. Applying landscape genomic tools to forest management and restoration of Hawaiian koa (Acacia koa) in a changing environment. **Evolutionary Applications**, Oxford, v. 11, n. 2, p. 231–242, 2018.

HOLMGREN, J. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. **Scandinavian Journal of Forest Research**, Stockholm, v. 19, n. 6, p. 543–553, 2004.

HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, New York, v. 630, p. 1044–1056, 2018.

HUANG, H. et al. Mapping vegetation heights in China using slope correction ICESat data, SRTM, MODIS-derived and climate data. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 129, p. 189–199, 2017.

HUETE, A. . A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, New York, v. 25, n. 3, p. 295–309, ago. 1988.

JIN, S. et al. The transferability of Random Forest in canopy height estimation from multi-source remote sensing data. **Remote Sensing**, Basel, v. 10, n. 8, p. 1–21, 2018.

JUSTICE, C. O. et al. The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. **IEEE Transactions on Geoscience and Remote Sensing**, New York, v. 36, n. 4, p. 1228–1249, 1998.

LARIBI, A. et al. A Machine Learning Approach for Radar Based Height Estimation. In: **2018 21st International Conference on Intelligent Transportation Systems (ITSC).**

Phuket: IEEE, 2018. p. 2364-2370.

LEE, J. et al. Machine learning approaches for estimating forest stand height using plot-based observations and Airborne LiDAR data. **Forests**, Basel, v. 9, n. 5, 2018.

LI, M. et al. A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 49, p. 87–98, 2016.

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

MA, L. et al. Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. **ISPRS International Journal of Geo-Information**, Basel, v. 6, n. 2, 2017.

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

MÄKINEN, H. et al. Radial growth variation of Norway spruce (Picea abies (L.) Karst.) across latitudinal and altitudinal gradients in central and northern Europe. **Forest Ecology and Management**, Amsterdam, v. 171, n. 3, p. 243–259, 2002.

MILLER, J. D.; THODE, A. E. Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR). **Remote Sensing of Environment**, New York, v. 109, n. 1, p. 66–80, 2007.

NAUDTS, K. et al. Mitigate Climate Warming. **Science**, New York, v. 351, n. 6273, p. 597–601, 2016.

PAING, M. P.; CHOOMCHUAY, S. Improved Random Forest (RF) classifier for imbalanced classification of lung nodules. In: **2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)**. Phuket: IEEE, 2018. p. 1-4.

PAUL, D. et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. **Computerized Medical Imaging and Graphics**, Amsterdam, v. 60, p. 42–49, 2017.

PENG, C.; ZHANG, L.; LIU, J. Developing and validating nonlinear height-diameter models for major tree species of ontario's boreal forests. **Northern Journal of Applied Forestry**, Bethesda, v. 18, n. 3, p. 87–94, 2001.

POMPA-GARCÍA, M.; JURADO, E. Seasonal precipitation reconstruction and teleconnections with ENSO based on tree ring analysis of *Pinus cooperi*. **Theoretical and Applied Climatology**, Wien, v. 117, n. 3–4, p. 495–500, 2014.

PRIMICIA, I. et al. Age, competition, disturbance and elevation effects on tree and stand growth response of primary Picea abies forest to climate. **Forest Ecology and Management**, Amsterdam, v. 354, p. 77–86, 2015.

QI, J. et al. A modified soil adjusted vegetation index. **Remote Sensing of Environment**, New York, v. 48, n. 2, p. 119–126, 1994.

R CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018.

REIS, A. A. et al. Volume estimation in a Eucalyptus plantation using multi-source remote sensing and digital terrain data: a case study in Minas Gerais State, Brazil. **International Journal of Remote Sensing**, Basingstoke, v. 40, n. 7, p. 2683–2702, 2019.

ROUSE JR, J. W. ET AL. **Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation**. Greenbelt: NASA/GSFC, 1973.

SCHLUND, M. et al. Canopy height estimation with TanDEM-X in temperate and boreal forests. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 82, n. November 2018, p. 101904, 2019.

SCOLFORO, H. F. et al. Modeling dominant height growth of eucalyptus plantations with parameters conditioned to climatic variations. **Forest Ecology and Management**, Amsterdam, v. 380, p. 182–195, 2016.

SCOLFORO, H. F. et al. Incorporating rainfall data to better plan eucalyptus clones deployment in eastern Brazil. **Forest Ecology and Management**, Amsterdam, v. 391, p. 145–153, 2017.

SCOLFORO, J. R. et al. **Equções de volume, peso de materia seca e carbono para diferentes fisionomias da flora nativa**. Lavras: Editora UFLA, 2008.

SEGURA, M.; KANNINEN, M. Allometric models for tree volume and total aboveground biomass in a tropical humid forest in Costa Rica. **Biotropica**, Washington, v. 37, n. 1, p. 2–8, 2005.

SILVEIRA, E. M. O. et al. Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 78, p. 175–188, 2019.

THOM, D.; RAMMER, W.; SEIDL, R. The impact of future forest dynamics on climate: interactive effects of changing vegetation and disturbance regimes. **Ecological Monographs**, Durhan, v. 87, n. 4, p. 665–684, 2017.

URBAZAEV, M. et al. Potential of multi-temporal ALOS-2 PALSAR-2 ScanSAR data for vegetation height estimation in tropical forests of Mexico. **Remote Sensing**, Basel, v. 10, n. 8, 2018.

WANG, Y. et al. A combined GLAS and MODIS estimation of the global distribution of mean forest canopy height. **Remote Sensing of Environment**, New York, v. 174, p. 24–43, 2016.

WILSON, E. H.; SADER, S. A. Detection of forest harvest type using multiple dates of Landsat TM imagery. **Remote Sensing of Environment**, New York, v. 80, n. 3, p. 385–396, 2002.

ZHANG, J. et al. Regional and historical factors supplement current climate in shaping global forest canopy height. **Journal of Ecology**, Oxford, v. 104, n. 2, p. 469–478, 2016.

ZHU, X.; LIU, D. Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 102, p. 222–231, 2015.