MÔNICA CANAAN CARVALHO

INTELIGÊNCIA COMPUTACIONAL NA MODELAGEM FLORESTAL: TEOR DE CARBONO E DISTRIBUIÇÃO GEOGRÁFICA DE ESPÉCIES

LAVRAS – MG
2019

MÔNICA CANAAN CARVALHO

INTELIGÊNCIA COMPUTACIONAL NA MODELAGEM FLORESTAL: TEOR DE CARBONO E DISTRIBUIÇÃO GEOGRÁFICA DE ESPÉCIES

Tese 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 Doutor.

Prof. Dr. Lucas Rezende Gomide
Orientador

LAVRAS – MG
2019

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

Carvalho, Mônica Canaan.
Inteligência computacional na modelagem florestal: teor de Carbono e distribuição geográfica de espécies / Mônica Canaan Carvalho. – 2019.
148 p. : il.

Orientador: Lucas Rezende Gomide.
Tese (Doutorado) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. Aprendizagem de máquina. 2. *Random Forest*. 3. Florestas nativas. I. Gomide, Lucas Rezende. II. Título.

MÔNICA CANAAN CARVALHO

INTELIGÊNCIA COMPUTACIONAL NA MODELAGEM FLORESTAL: TEOR DE CARBONO E DISTRIBUIÇÃO GEOGRÁFICA DE ESPÉCIES

COMPUTER INTELLIGENCE IN FOREST MODELING: CARBON STOCK AND GEOGRAPHICAL SPECIES DISTRIBUTION

Tese 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 Doutor.

APROVADA em 15 de Março de 2019.

Prof. Dra. Mayra Luiza Marques Silva	DEFLO/UFSJ
Dra. Marcela de Castro Nunes Santos Terra	UFLA
Prof. Dr. Marcelo Ribeiro Viola	UFLA
Prof. Dr. Fausto Weimar Acerbi Júnior	UFLA

Prof. Dr. Lucas Rezende Gomide
Orientador

LAVRAS – MG
2019

Ao meu tríade, pais e irmã.

DEDICO.

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que deram contribuições técnico-científicas para a elaboração desta tese:

Lucas Rezende Gomide

Fausto Weimar Acerbi Junior

Eduarda Martiniano Silveira

David Tng

Aliny Aparecida dos Reis

Rafael Menali

Cassio Augusto Ussi Monti

Luciano Cavalcante de Jesus França

Isáira Leite e Lopes

Rubens Manoel dos Santos

Sérgio Henrique Godinho

Agradeço ao suporte prestado em todos estes anos de pesquisa pela Universidade Federal de Lavras (UFLA), Laboratório de Estudos e Projetos em Manejo Florestal (LEMAF) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Agradeço àqueles que na convivência diária no Laboratório de Estudos e Projetos em Manejo Florestal (LEMAF) trouxeram alegria e inspiraçaõ para continuar a jornada.

Por fim, agradeço aos familiares, amigos, namorado e cachorros, pelo amor, apoio e confiança incondicionais nestes 4 longos anos.

RESUMO GERAL

A modelagem de florestas, seja de suas variáveis dendrométricas ou de sua distribuição geográfica, é uma prática consolidada na Engenharia Florestal, na qual tradicionalmente são empregados modelos da estatística clássica. Entretanto, o progresso obtido pela ciência da computação nas últimas décadas tem possibilitado novos desafios e soluções, os quais podem limitar a utilização dos modelos estatísticos tradicionais. É no contexto de relações não-lineares pouco conhecidas que os algoritmos de aprendizagem de máquina ganham sua utilidade no setor florestal. Dentre esses algoritmos, recebe destaque o *random forest* (RF), por sua robustez, facilidade de parametrização e métricas internas. Apesar do grande potencial deste algoritmo, o mesmo carece de mais estudos para consolidar o uso da técnica. Nessa tese, o algoritmo foi aplicado em três situações diferentes para florestas nativas, abordando problemas de classificação e regressão, além de dados heterogêneos oriúndos de diversas fontes. O primeiro artigo (1) teve como objetivo avaliar três métodos de aprendizagem de máquina (árvore de decisão - J48, RF e redes neurais artificias), na modelagem da distribuição potencial de dez espécies arbóreas mais abundantes em uma sub-bacia do rio São Francisco, em Minas Gerais. Conclui-se que o algoritmo RF apresentou-se como o mais robusto para a modelagem da distribuição potencial de espécies arbóreas. Diante destes resultados obtidos pelo algoritmo, no segundo artigo (2) procurou-se modelar a distribuição potencial de *Eremanthus erythropappus*, considerando cenários de mudanças climáticas. A hipótese testada está associada aos efeitos no nicho ecológico da espécie no futuro (2050 e 2070). Os resultados indicam uma boa acurácia do método empregado, destacando as cadeias montanhosas do Espinhaço, Canastra e Mantiqueira como os três principais refúgios ecológicos da espécie. Constatou-se, caso confirmado os cenários de mudanças no clima, uma redução drástica na área potencial de desenvolvimento da espécie, no que tange sua interação com o clima local. O último artigo (3) concentrou-se na aplicação do RF para problemas de regressão, envolvendo a predição da variável quantitativa teor de Carbono acima do solo. O objetivo foi testar a combinação de métodos e estratégias na seleção de variáveis, empregando o RF. Os resultados alcançados indicam o algoritmo RF como um método robusto, pouco afetado pela inclusão de um grande número de variáveis correlacionadas. Mesmo com a pequena melhora nos erros do algoritmo, o uso das técnicas de seleção de variáveis se justifica visto que diminui consideravelmente o número de variáveis utilizadas. O algoritmo genético multiobjetivo obteve menor conjunto de dados selecionados bem como menor erro. Perante os resultados encontrados, atribui-se ao RF grande potencial para explorar relações ainda pouco conhecidas na Engenharia Florestal, sejam estas relações problemas classificatórios ou de regressão.

Palavras-chave: Aprendizagem de máquina. *Random forest*. Florestas nativas.

GENERAL ABSTRACT

Forest modeling, whether for its dendrometric variables or geographical distribution, is a consolidated practice in Forest Engineering in which traditional statistical models are typically employed. However, the progress made by computer science in recent decades has made new challenges and solutions possible, which may limit the use of traditional statistical models. Machine learning algorithms gain their use in the forestry sector within the context of little-known non-linear relations derived from large databases. Among these algorithms, the random forest (RF) is prominent due to its robustness, ease of parameterization, and internal metrics. Despite its great potential, this algorithm demands further studies to consolidate its use. In this dissertation, we applied the algorithm in three different situations for native forests, addressing classification and regression issues, as well as heterogeneous data from different sources. The first article (1) aimed to evaluate three machine learning methods (decision tree - J48, RF, and artificial neural networks) for the potential distribution modeling of the ten most abundant tree species in a subbasin of the Sao Francisco River, in Minas Gerais, Brazil. In conclusion, the RF algorithm presented the most robust tree species potential distribution model. With the results obtained by the algorithm, we wrote the second article (2) seeking to model the potential distribution of *Eremanthus erythropappus*, considering climatic change scenarios. The hypothesis tested is associated with future effects (2050 and 2070) on the ecological niche of the species. The results indicate a good accuracy of the method used, highlighting the Espinhaço, Canastra, and Mantiqueira mountain ranges as the three primary ecological refuges of the species. We verified a drastic reduction in the potential area of species development regarding its interaction with the local climate if the climate changes scenarios be come real. The last article (3) focused on the application of RF on regression problems involving the prediction of the quantitative variable of carbon content above the soil. The objective was to test the combination of methods and strategies in the selection of variables using the random forest. The results obtained indicate that the RF algorithm is a robust method, little affected by the inclusion of many correlated variables. Even with the slight improvement in algorithm errors, the use of variable selection techniques is justified since it considerably reduces the number of variables used. The multi-objective genetic algorithm obtained a smaller set of selected data and lower error. Given the results found, the RF has great potential to explore relationships still little known in Forest Engineering, whether they are classification or regression issues.

Keywords: Machine learningRandom Forest. Native Forests.

LISTA DE FIGURAS

PRIMEIRA PARTE

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

ARTIGO 2

ARTIGO 3

LISTA DE TABELAS

PRIMEIRA PARTE

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

ARTIGO 2

ARTIGO 3

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

A modelagem no setor florestal, bem como em outras áreas do conhecimento, é imprescindível para o manejo e planejamento das atividades, uma vez que conhecer as variáveis de uma floresta e suas relações por completo é financeiramente e temporalmente inconcebível. Desde os primórdios da Engenharia Florestal, a utilização da estatística e seus modelos paramétricos estão sempre presentes em estudos de crescimento e produção, classificação de sítio, estimativas de variáveis dendrométricas, etc. No entanto, dados biológicos, como os provenientes dos inventários florestais, comumente apresentam relações complexas, com dados faltantes e com a ocorrência de *outliers*. Muitas bases de dados de origem biológica dificilmente assumem a forma fixada por parâmetros pelos modelos matemáticos da estatística paramétrica, e/ ou até mesmo apresentam diferentes tipos de dados (diferentes fontes, formatos, resoluções, etc.), impossibilitando o emprego de modelos da estatística clássica.

Com o avanço dos processadores computacionais, algoritmos foram desenvolvidos a fim de reproduzir a habilidade do cérebro humano: o aprendizado. Esses algoritmos, intitulados algoritmos de Aprendizagem de Máquina (AM), extraem conhecimento de exemplos e extrapolam esse conhecimento para novos dados. AM se baseia no princípio do raciocínio indutivo, ou seja, que informações de um conjunto de exemplos fornecem informação para generalização do todo. De acordo com Witten e Frank (2005) não existe uma linha que separa a estatística clássica dos métodos de AM, no entanto estes últimos estão mais intimamente ligados à área da computação, enquanto os primeiros relacionados à matemática.

Conceitualmente, esses algoritmos possuem características vantajosas no uso de *big data* para a mineração dos dados. Uma dessas características é a capacidade de lidar com dados que contenham imperfeições ou ruídos. Além disso, os algoritmos tentam minimizar o impacto de *outliers* e aumentar a capacidade de generalização. São modelos flexíveis que trabalham com dados contínuos e/ou categóricos. A maioria dos algoritmos pode ser empregada para problemas de regressão e classificação (WITTEN; FRANK, 2005). Levando-se em consideração o fato exposto, tais métodos têm retratado robustos resultados em diversas aplicações (ELITH et al., 2006; GAO et al., 2018; PANDIT; TSUYUKI; DUBE, 2018).

Além da gama de opções em relação aos métodos de modelagem, a área de modelagem preditiva da vegetação, principalmente quando se trabalha em meso/larga escala, ganha um novo desafio: a grande quantidade de dados que descrevem o meio ambiente e o comportamento da vegetação. Esses dois fatos citados, variedade de métodos e dados, impõem à ciência a necessidade de se obter respostas quanto à sua utilização.

Diversas pesquisas vêm buscando estimar os diferentes resultados obtidos entre os métodos de modelagem preditiva da vegetação para a escolha da melhor técnica (ELITH et al., 2006; GAO et al., 2018; GARZÓN et al., 2006; WERE et al., 2015), bem como procurado determinar e compreender a influência de variáveis ambientais e espectrais no comportamento da vegetação (LATIFI; NOTHDURFT; KOCH, 2010; LU et al., 2014; REIS et al., 2018; SILVEIRA et al., 2019). No entanto, os resultados obtidos por essas pesquisas não apontam em uma única direção e nem compreendem a totalidade de combinações entre diferentes tipos de vegetação/variáveis dendrométricas, técnicas de modelagem e variáveis preditoras. Resultados esses que incitam novas aplicações na área de modelagem preditiva da vegetação.

Diante disso, objetiva-se nesta tese empregar algoritmos da área de aprendizagem de máquina, em destaque o algoritmo *random forest*, em problemas envolvendo dados de florestas nativas em macroescala. Propõe-se noartigo 1 avaliar três métodos de aprendizagem de máquina (árvore de decisão - J48, *Random forest* e redes neurais artificias), na modelagem da distribuição de dez espécies arbóreas mais abundantes em uma sub-bacia do rio São Francisco em Minas Gerais. Noartigo 2 tem-se como objetivo modelar a distribuição de *Eremanthus erythropappus*, popularmente conhecida como Candeia, para o cenário atual e futuro, considerando as mudanças climáticas, utilizando o algoritmo *random forest* e variáveis ambientais. Já no artigo 3 objetiva-se testar diferentes métodos de seleção de variáveis no desempenho do algoritmo *random forest* para a modelagem do estoque de Carbono acima do solo da vegetação nativa na bacia do Rio Grande, Minas Gerais, utilizando grande quantidade de dados de diferentes fontes, resoluções e formatos.

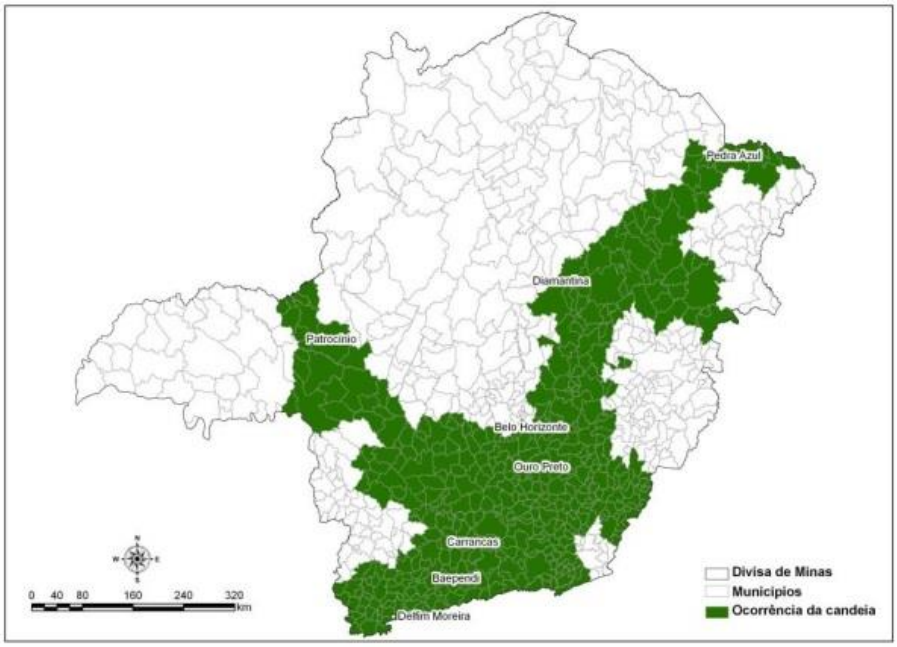
2 REVISÃO DE LITERATURA

2.1 A árvore Candeia e suas relações ecológicas econômicas

A Candeia pertence à família Asteraceae (Compositae) e ao gênero *Eremanthus*. As espécies de maior ocorrência no estado de Minas Gerais e, portanto, abordadas nesta pesquisa são *Eremanthus erythropappus* (DC.) MacLeish e *Eremanthus incanus* (Less.) Less. São espécies endêmicas de ecótonos, distribuídas nas áreas de transição entre as matas semideciduais e os campos abertos (cerrado) ou também os campos de altitude (SCOLFORO et al., 2004). A Candeia *Eremanthus erythropappus* está distribuída em toda a parte do sudeste do Planalto Central, compreendendo os estados de Goiás, Distrito Federal, Minas Gerais, Espírito Santo, São Paulo e Rio de Janeiro. Já a Candeia *Eremanthus incanus* ocorre no nordeste (Bahia) e sudeste (Minas Gerais, nos domínios Cerrado, Caatinga e Mata Atlântica (LOEUILLE, 2017; MACLEISH, 1987; SCOLFORO; OLIVEIRA; DAVIDE, 2012).

Em relação à caracterização do ambiente onde a Candeia é naturalmente encontrada em Minas Gerais (FIGURA 1), atemperatura média anual varia entre 18°C e 20°C e temperatura do mês mais quente entre 22°C e 30°C. A média pluviométrica anual varia entre 1.400 e 1.550 mm e ocorre em altitudes entre 900 a 1.700m. Desenvolvem-se em solos pouco férteis, com acidez elevada e textura média a arenosa. As classes de solos compreendidas por candeais nativos em Minas Gerais são: Cambissolo álico (Ca), Cambissolo distrófico (Cd), Latossolo Vermelho escuro distrófico (LEd), Latossolo Vermelho escuro álico (LEa), Latossolo Vermelho Amarelo distrófico (LVd), Latossolo Roxo distrófico (LRd), solo Litólico álico (Ra) e Podzólico Vermelho Amarelo distrófico (PVd) (SCOLFORO; OLIVEIRA; DAVIDE, 2012).

Figura 1 - Área hipotética de ocorrência natural da Candeia em Minas Gerais.



Fonte: Altoé (2012).

A Candeia apresenta sementes abundantes e pequenas, alta taxa de regeneração natural em condições adequadas (exigentes de luz) e dispersão anemocórica. São consideradas precursoras na colonização de campos abertos, clareiras e áreas desmatadas, pertencentes ao grupo ecológico das pioneiras, secundárias iniciais. Essas características lhe conferem grande capacidade de disseminação, ocupando grandes áreas homogêneas (povoamentos com alta porcentagem de cadeia) denominadas candeais.

Da Candeia são extraídos produtos de alto valor comercial, como o óleo essencial e o α-bisabolol (principalmente da espécie *E. erythropappus*) utilizados pelas indústrias cosmética e farmacêutica (SILVA et al., 2008). Pela alta durabilidade de sua madeira, a Candeia é muito utilizada como moirões para cerca (OLIVEIRA et al., 2010). Ademais, a Candeia desenvolve-se em solos

pouco férteis, rasos e, predominantemente, em áreas de campos de altitude, locais pouco propícios para empreendimentos agrícolas, destacando-se como uma espécie em potencial para fins de recuperação ambiental (SCOLFORO; OLIVEIRA; DAVIDE, 2012). Segundo Ribeiro et al. (2017), as espécies *E. erythroppapus* e *E. incanus*, por se desenvolverem em campos de altitude, onde os solos são distróficos e ricos em metais, desenvolveram adaptações para essas condições extremas (acúmulo de metais como Manganês e Ferro). Esse esforço é evolucionariamente compensado pelo decréscimo do ataque às folhas por herbivoria.

Com relação à produtividade da Candeia e de acordo com inventário realizado em candeais nativos (*E. erythropappus* e *E. incanus*) em Minas Gerais (Delfim Moreira, Ouro Preto e Aiuruoca), Scolforo, Oliveira e Davide (2012) obtiveram valores médios do número de indivíduos, volume com casca, massa de óleo e número de moirões por hectare, que podem ser visualizados na Tabela 1. O fator de empilhamento médio encontrado pelos autores para conversão de volume em metro estéreo para volume em metro cúbico foi 2,46.

Tabela 1 - Valores médios por hectare do número de indivíduos, volume total com casca (Vcc), massa de óleo em Kg extraída do volume total com casca (m^3cc) da Candeia e número de moirões (N/há).

Região	Delfim Moreira	Aiuruoca	Ouro Preto
Número de indivíduos (N/ha)	875,47	936,12	1.281,15
Vcc total (m^3/ha)	35,32	38,33	77,46
Massa de óleo kg (m^3cc/ha)	363,7	393,05	819,2
Número de moirões (N/ha)	2.637,48	2.825,93	4.796,52

Fonte: Scolforo, Oliveira e Davide (2012).

2.2 Estudo da Interação planta ambiente através da análise de gradientes

Desde meados do século 20, a ciência tem buscado entender e replicar as relações existentes entre a vegetação e o ambiente por meio da abordagem de análise de gradientes. Como definido por Whittaker (1967), a análise de gradiente tem como objetivo descrever e entender a distribuição da vegetação em resposta a um ou mais fatores ambientais e/ou gradientes temporais. A partir dessa época, a ciência tem comprovado a influência de diferentes gradientes ambientais na distribuição e composição de ecossistemas vegetais (FRANKLIN, 1995).

Os primeiros gradientes a serem estudados estão relacionados com o terreno, em nível local, através da utilização de transectos com amostragem da vegetação ao longo de um único gradiente, como elevação do terreno (WHITTAKER, 1967). Diversas pesquisas comprovam a influência do gradiente altitudinal na composição e produção de florestas nativas (LIEBERMAN et al., 1996; RANA; SINGH; SINGH, 1989; VÁZQUEZ GARCÍA; GIVNISH, 1998). Vázquez García e Givnish (1998) encontraram forte relação entre o gradiente altitudinal e a composição e diversidade em uma Floresta Tropical do México. Os autores verificaram o declínio do número de espécies e diversidade com o aumento das cotas de altitude. Em Rana, Singh e Singh (1989), os autores apuraram alta produção primária relacionada ao gradiente altitudinal, fato relacionado com o aumento da umidade dado o incremento em altitude. Esse gradiente é tão explícito na produção primária das florestas, que deve ser incorporado em modelos para estimativa da biomassa acima do solo também em larga escala (SCOLFORO et al., 2015).

Outros gradientes ambientais também foram analisados, como precipitação (GWITIRA et al., 2014; PALMER; STADEN, 1992), temperatura (BRZEZIECKI; KIENAST; WILDI, 1993; GWITIRA et al., 2014), solos

(LOWELL, 1991), geologia (FRANK; GOETZ, 1990; PAYNE; STOCKWELL; DAVEY, 1994), radiação (BROWN, 1994; MACKEY, 1994), etc., revelando tendências claras de respostas da vegetação a tais fatores. Esses estudos ganharam força com o advento das áreas de Sensoriamento Remoto e Sistemas de Informações Geográficas (SIG's), que possibilitaram a disponibilidade de gradientes disponíveis e métodos de análise. Graças ao desenvolvimento dessas áreas foi possível tornar a análise de gradientes, antes empregadas em escala local, a um mapeamento preditivo da vegetação, aplicado a estudos em meso/larga escala. Esse tipo de mapeamento consiste em utilizar esses gradientes para explicar e predizer a composição, diversidade e biomassa das florestas (FRANKLIN, 1995).

Dentre os diversos aspectos da vegetação que podem ser preditos por gradientes ambientais, agora também descritos por variáveis do sensoriamento remoto, este trabalho destaca a modelagem da distribuição potencial de espécies e a modelagem de variáveis dendrométricas de um povoamento florestal, ambos de suma importância para o planejamento e manejo florestal sustentável.

2.3 Mudanças climáticas

As mudanças no clima, comprovadas cientificamente pelo Painel Intergovernamental sobre Mudanças Climáticas (*Intergovernmental Panel on Climate Change* - IPCC), constituem um novo paradigma para o mapeamento preditivo da vegetação. Em 2014, o IPCC lançou seu quinto relatório (AR5), no qual confirmam-se as hipóteses dos relatórios anteriores (INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE - IPCC, 2014). Em 2007, no AR4, já havia 90% de certeza de que mais da metade do aumento observado na temperatura média da superfície entre os anos de 1951 a 2010 teve como causa o aumento das concentrações de gases do efeito estufa, devido às

atividades antropogênicas aliadas a forçantes climáticas. Os indicadores de mudanças climáticas mostrados por esses relatórios foram o aumento da temperatura global, aumento do nível do mar e a redução da cobertura de gelo.

Nesse relatório (AR5) foram criadas quatro projeções diferentes, chamados "Caminhos Representativos de Concentração" (*Representative Concentration Pathways* – RCP), representados pelos valores de sua respectiva forçante radiativa (FR). A forçante radiativa é oriúnda de um agente climático, e é definida como a diferença em irradiância líquida na tropopausa, entre um estado de referência e um estado perturbado devido a ação do agente climático (FORSTER et al., 2007). Uma forçante radiativa positiva significa que um agente tende a aquecer o planeta, ao passo que valores negativos indicam uma tendência de resfriamento. As estimativas das FR's consideram o histórico evolutivo de diversos fatores, como emissão de gases, concentração de gases de efeito estufa, e informações de tipo de cobertura terrestre. A escala de projeções vai de 2.6 W/m² (cenário otimista) a 8.5 W/m² (cenário pessimista), sendo o cenário atual de 2.2 W/m².

Os RCPs mais utilizados em trabalhos de averiguação dos diversos impactos das mudanças climáticas na vegetação são o 4.5 (moderado) e 8.5 W/m²(pessimista), também empregados no artigo 2 desta tese. O RCP4.5 pressupõe que a forçante radiativa estabiliza pouco depois do ano de 2100, sem ultrapassar o nível de radiação em longo prazo de 4,5 W/m^2. Essa projeção é consistente com a estabilização da demanda energética mundial, programas de reflorestamento fortes e políticas climáticas rigorosas. Além disso, sugere uma estabilização das emissões de metano associadas a um leve aumento das emissões de CO_2 até 2040, atingindo o valor alvo de 650 ppm de CO_2 equivalente na segunda metade do século XXI. O RCP8.5 sugere um crescimento contínuo da população associado a um desenvolvimento tecnológico lento, resultando em acentuadas emissões de dióxido de carbono.

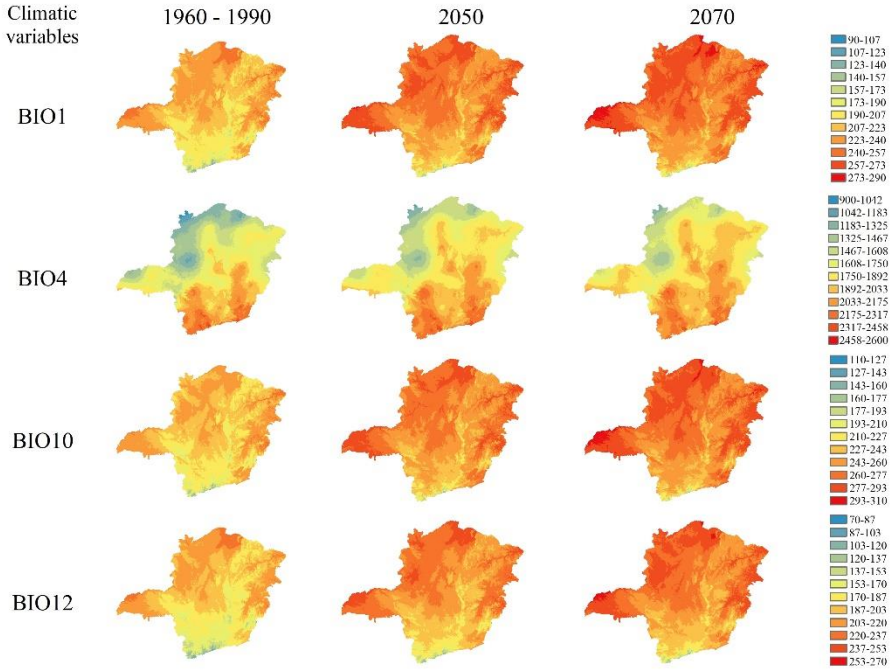
Esse cenário é considerado o mais pessimista para o século XXI em termos de emissões de gases do efeito estufa, consistente com nenhuma mudança política para reduzir as emissões e forte dependência de combustíveis fósseis e com aumento na temperatura de até 4°C (CHOU et al., 2014; MARENGO et al., 2014; SILVEIRA et al., 2016).

Existem diversos modelos matemáticos utilizados para estimar e espacializar o impacto das forçantes climáticas na alteração do clima na superfície da Terra, denominados *General Circulation Models* (GCM).. Dentre os modelos disponíveis, destacam-se pelo uso em diversas pesquisas realizadas na América do Sul os modelos *Hadley Global Environment Model 2 - Earth System (*HadGEM2-ES*)* e *Model for Interdisciplinary Research on Climate version 5 (*MIROC5*)* (WATANABE et al., 2010). A grande empregabilidade destes modelos é devida à regionalização e disponibilização de suas estimativas pelo Instituto de Pesquisas Espaciais (INPE) do Brasil.

Diante da instabilidade dos cenários de mudanças no clima, quer seja em relação aos RCP's ou aos GCM's, é uma prática comum realizar as projeções baseadas na média dos cenários para cada período. Tal prática, além de incorporar as incertezas em relação às mudanças climáticas, facilita a apresentação e interpretação dos resultados (WANG et al., 2012; ZHANG et al., 2015). As Figuras 2 e 3 apresentam as médias de variáveis de temperatura e precipitação, respectivamente, para os três períodos temporais trabalhados nesta pesquisa (atual, 2050 e 2070) de acordo com os modelos HadGEM2-ES e MIROC5 baseados nos RCP's 4.5 e 8.5 para o estado de Minas Gerais. É possível visualizar um aumento de temperatura crescente para os perídos de 2050 e 2070, tanto para a média anual (BIO01) quanto para as estações de verão e inverno (BIO10 e BIO11, respectivamente). A sazonalidade da temperatura (BIO04) não apresentou grandes mudanças em relação aos períodos futuros. A precipitação média anual apresentou comportamento oposto nas regiões do

estado, com aumento nas regiões leste e centro-sul e diminuição na região nordeste. A precipitação média do verão (BIO16) demonstrou aumento principalmente nas regiões montanhosas do estado de Minas Gerais, abrangendo as regiões da Serra do Espinhaço, Serra da Mantiqueira e Serra da Canastra. A precipitação média da estação seca (BIO17) não exibiu grandes variações para os diferentes cenários.

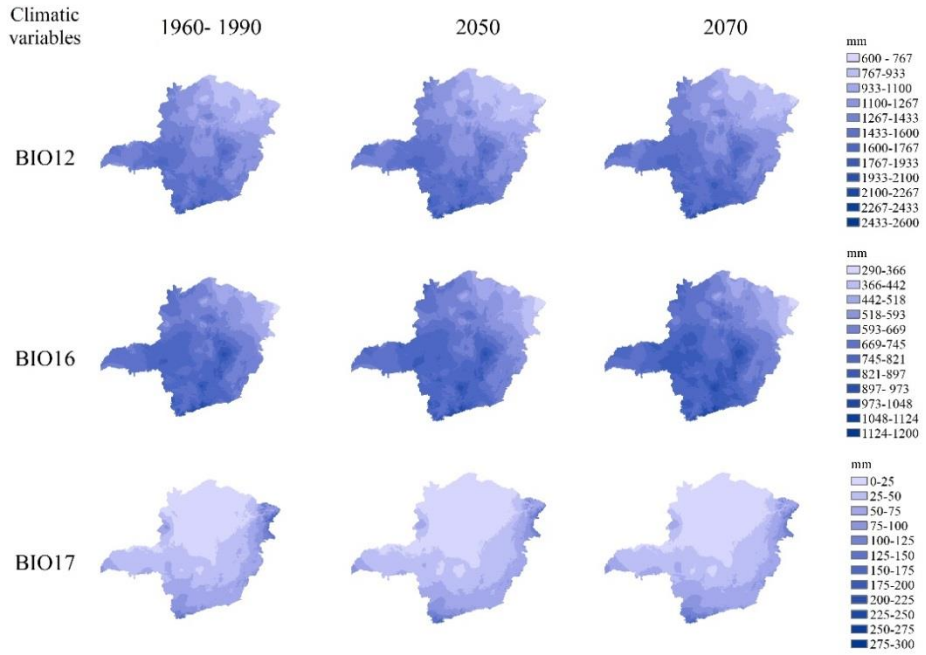
Figura 2 - Médias de variáveis bioclimáticas relacionadas à temperatura para os três períodos temporais trabalhados nesta pesquisa (atual, 2050 e 2070) de acordo com os modelos HadGEM2-ES e MIROC5 baseados nos RCP's 4.5 e 8.5 para o estado de Minas Gerais.



Legenda: Dados oriúndos da base de dados do WorldClim versão 1.0 (HIJMANS et al., 2005). Em que: BIO01 – temperatura media anual (°C*10); BIO04 – sazonalidade da temperatura (°C*1.000); BIO10 – temperatura media da estação quente (°C*10); BIO11 – temperatura media da estação fria (°C*10);

Fonte: Da autora (2018).

Figura 3 - Médias de variáveis bioclimáticas relacionadas à precipitação (mm) para os três períodos temporais trabalhados nesta pesquisa (atual, 2050 e 2070) de acordo com os modelos HadGEM2-ES e MIROC5 baseados nos RCP's 4.5 e 8.5 para o estado de Minas Gerais.



Legenda: Dados oriúndos da base de dados do WorldClim versão 1.0 (HIJMANS et al., 2005). Em que: BIO12 – precipitação media anual (mm); BIO16 – precipitação media da estação úmida (mm); BIO17 – precipitação media da estação seca (mm).
Fonte: Da autora (2018).

2.4 Modelagem da distribuição potencial de espécies

Seja para fins de conservação, restauração ou produção, a questão de como o reino vegetal e animal estão distribuídos na Terra tem sido fonte de estudos para diversos pesquisadores. É certo que fatores climáticos, físicos e biológicos são responsáveis, em diferentes escalas, pela distribuição das espécies no planeta. A partir dessa informação, pesquisadores têm utilizado esses fatores como variáveis de entrada em modelos matemáticos, para a predição de locais que satisfaçam às necessidades da espécie. Esse tipo de modelagem é

denominado modelagem preditiva de habitat ou modelagem da distribuição potencial de espécies (GIANNINI et al., 2012).

De acordo com Hutchinson (1957) toda espécie possui um hiper-espaço n-dimensional próprio que possibilita sua reprodução e crescimento, denominado nicho. Esse espaço é limitado por todos os fatores ambientais que atuam sobre o organismo, como por exemplo radiação solar, temperatura, precipitação, etc. Em outras palavras, o nicho representa o espaço abrangido pela faixa de variação dos fatores ambientais o qual a espécie é capaz de sobreviver e se reproduzir (GIANNINI et al., 2012). Com base nesse conceito, é possível entender e explicar como as espécies se distribuem geograficamente na superfície terrestre, a partir da análise das condições ambientais de seus locais de ocorrência.

O processo de modelagem da distribuição potencial de espécies tem como objetivo encontrar relações não aleatórias, entre as variáveis ambientais, relevantes para a espécie, e seus dados de ocorrência. A modelagem preditiva de habitat combina dados de ocorrência de espécies (coordenadas geográficas) com variáveis ambientais e ecológicas (como por exemplo: temperatura, precipitação, altitude, tipo de solo, índices de vegetação, etc.) para realizar a predição de ambientes adequados, onde, em teoria, uma população possa se manter viável. O resultado da modelagem é, então, projetado em um mapa, indicando as regiões com distribuição potencial da espécie (ANDERSON; LEW; PETERSON, 2003).

De acordo com Soberón e Peterson (2005), a distribuição espacial de uma espécie está relacionada a quatro fatores principais:

a) Fatores abióticos – condições ambientais que limitam a capacidade de sobrevivência e reprodução da espécie em determinada região (por exemplo: altitude, inclinação do terreno, fertilidade do solo, pluviosidade, temperatura, etc.);

b) Fatores bióticos – Conjunto de interações com outras espécies que influenciam na sobrevivência da espécie em estudo, como competição, parasitismo, predação, mutualismo, etc;
c) Fatores de acessibilidade – Relacionados à capacidade de dispersão, que refletem quais locais são acessíveis para indivíduos de uma determinada espécie (importante para distinguir distribuição atual e distribuição potencial);
d) Fatores evolucionários – Relacionados com a capacidade de adaptação às novas condições (plasticidade da espécie).

Utilizando apenas os fatores abióticos como condicionadores de um habitat viável, obtêm-se locais satisfatórios para a espécie (potencial) e não exatamente locais ocupados pela espécie (real) (SÓBERON; PETERSON, 2005). A principal razão para a utilização desses fatores em detrimento dos demais é a dificuldade de se obter variáveis que representem condições bióticas, cuja interpretação é complexa. Por isso, o uso do termo potencial, indicando áreas com aptidão para a ocorrência da espécie em questão.

Nas últimas duas décadas, graças aos avanços em SR e SIG, esse tipo de modelagem tem ganhado grande interesse no meio científico. Sua utilidade vai além do entendimento das relações entre a vegetação e o ambiente, empregado para estimar as consequências das mudanças climáticas na distribuição de diferentes ecossistemas e espécies florestais (CASALEGNO et al., 2011; IVERSON; MCKENZIE, 2013; WANG et al., 2016), bem como para indicação de espécies para reflorestamento para fins de produção e conservação (CARVALHO et al., 2017; COELHO; TAVARES; GOMIDE, 2016).

2.5 Modelagem de variáveis dendrométricas em meso-larga escala

O crescimento e produção de uma espécie florestal, de acordo com Husch, Miller e Beers (1982), é definido em função da espécie, idade, densidade e sítio. Este último é definido como a totalidade dos fatores ambientais (fatores climáticos, edáficos e bióticos) que direta ou indiretamente influenciam a sobrevivência e crescimento das florestas. Partindo dessa premissa, diversas pesquisas que têm como objetivo comum estimar o crescimento e produção de determinada variável dendrométrica, vêm incorporando dados ambientais na modelagem dessas variáveis (DUBE; MUTANGA, 2016; REIS et al., 2018).

Conhecer os atributos de produção de uma floresta, como área basal, volume, biomassa e teor de Carbono, sempre foi a questão central dentro do manejo florestal. Modelos da estatística clássica constituem o arcabouço utilizado para se estimar essas variáveis para o povoamento florestal como um todo, apresentando modelos lineares e não lineares consolidados para estimativas de parâmetros quantitativos das florestas. Entretanto, tal metodologia é de difícil utilização quando se deseja obter estimativas para estudos em escala regional e global. Dificuldade esta que aumenta na mesma medida da complexidade da base de dados a ser modelada.

Com o progresso na área da computação e sistemas de informações geográficas (SIG's), uma sucessão de técnicas e sensores surgiram nas pesquisas sobre a modelagem de variáveis dendrométricas em larga escala. Além da utilização de medições em campo como entrada para os modelos, passou-se a incorporar variáveis com características do ambiente, aumentando o poder explicativo dos modelos e permitindo a espacialização das estimativas na região estudada.

Para florestas plantadas, como as de *Eucalyptus* sp., essas novas técnicas já têm sido bastante empregadas, alcançando melhoria na acurácia das

estimativas (BOISVENUE et al., 2016; FAYAD et al., 2016; MORENO; NEUMANN; HASENAUER, 2016). Reis et al. (2018) utilizaram dados multiespectrais e atributos do terreno para estimar o volume de *Eucalyptus* sp. no nordeste de Minas Gerais. Dube e Mutanga (2016) verificaram que a integração de conjuntos de dados multiespectrais com variáveis ambientais fornece um conjunto de ferramentas robusto necessário para a recuperação precisa e confiável de biomassa acima do solo da floresta e estoques de Carbono em áreas densamente florestadas. Os resultados encontrados por esses autores corroboram o potencial da utilização de dados coletados em imagens de sensoriamento remoto e Sistema de informação geográfica para a estimativa da produtividade de florestas plantadas.

Essa mesma metodologia tem sido empregada para florestas nativas, porém obtendo acurácias inferiores dos modelos devido à heterogeneidade da vegetação. As principais variáveis estimadas para esses ecossitemas florestais consistem na produção primária (biomassa) e teor de Carbono, a quais representam a chave para o entendimento do ciclo do Carbono na Terra e para a implementação de políticas públicas para redução do dióxido de Carbono na atmosfera, visando à contenção das mudanças no clima (BACCINI et al., 2008). Silveira et al. (2019a), através da integração de dados espectrais e variáveis do terreno, estimou a biomassa da vegetação nativa para a bacia do Rio Doce, Brasil. Os autores encontraram grande variação de biomassa, entre 25,2 a 238 $Mg.ha^{-1}$, com erro médio das estimativas (RMSE) e coeficiente de determinação (R^2) de 20,08 $Mg.ha^{-1}$ e 0,86, respectivamente. Scolforo et al. (2015) mapearam o teor de Carbono presente na vegetação nativa de Minas Gerais por meio de modelo linear múltiplo, utilizando as variáveis latitude, longitude, altitude e bioma. As variáveis altitude e latitude foram correlacionadas significativamente com os dados obtidos em campo. Os autores encontraram valores de teor de Carbono maiores nas regiões Centro-Sul e Leste, as quais apresentam

predominância do bioma Mata Atlântica, com média ponderada de 48,6 e 48,4 $Mg.ha^{-1}$, respectivamente. O modelo linear múltiplo, denominado modelo geográfico, obteve um erro padrão residual de 16,20 $Mg.ha^{-1}$e coeficiente de determinação de 0,53.

2.6 Desafios e tendências para o mapeamento preditivo da vegetação

O mapeamento preditivo da vegetação é um ramo da ciência que utiliza das relações entre a vegetação e o ambiente para mapear e predizer características desses ecossistemas naturais. No final do século 20, a prática comum para esses estudos de modelagem envolviam poucos dados provenientes de SIG e algumas técnicas da estatística clássica, como as regressões lineares (FRANKLIN, 1995). Com o avanço das técnicas computacionais, por meio de melhorias em hardwares e softwares, e desenvolvimento de novos sensores, uma grande quantidade de dados e métodos para extrair conhecimento desses dados tornaram-se acessíveis.

No entanto essa grande evolução na disponibilização dos dados a serem modelados e da existência de diferentes técnicas de modelagem, carecem de estudos minusciosos sobre suas influências e desempenhos. Como é o caso da utilização de dados espectrais para o mapeamento da distribuição, composição e produção da vegetação (LU et al., 2014). Atualmente existe uma ampla variedade de satélites, radares, e lasers, demandando grande esforço científico para o entendimento e aplicação dessas variáveis e suas modificações no mapeamento preditivo da vegetação. Sensores ópticos como IKONOS, Quickbird, Worldview, ZY-3, sistema SPOT, Sentinel, Landsat e MODIS, com resoluções espaciais que variam de menos de um metro a centenas de metros, têm sido utilizados com sucesso como variáveis para estimativas de características dendrométricas das florestas (REIS et al., 2018; SILVEIRA et al.,

2019). No entanto os resultados obtidos são diversos, não há consenso sobre quais variáveis e sensores espectrais sejam mais indicados para determinado objetivo, como por exemplo estimativa da biomassa e estoque de Carbono.

A diversidade de variáveis ambientais somadas à disponibilidade de variáveis espectrais, caracteriza o típico problema resolvido pela mineração de dados: a seleção de variáveis mais importantes para a modelagem. Essa prática tem se tornado obrigatória em estudos em larga escala que utilizam múltiplas fontes de dados, como é o caso do mapeamento preditivo da vegetação. Além de diminuir consideravelmente a dimensionalidade do problema, a seleção de variáveis elimina parte da multicolinearidade dos dados, problema que pode levar à instabilidade das predições (LATIFI; NOTHDURFT; KOCH, 2010).

Uma técnica usualmente empregada consiste na avaliação dos erros obtidos com a remoção recursiva das variáveis na modelagem, na qual seus resultados podem ser interpretados como quanto maior o erro após a remoção de determinada variável, maior a importância e contribuição desta no modelo. Após o cômputo da importância relativa da variável, as variáveis com menor importância são retiradas recursivamente, e o conjunto remanescente de variáveis é avaliado quanto ao desempenho do modelo. O conjunto de variáveis que obtiver menor erro é, então, selecionado para a modelagem (REIS et al., 2018; WERE et al., 2015).

Silveira et al. (2019) empregaram essa metodologia para mapear a biomassa acima do solo em uma região tropical do Brasil por meio de dados espectrais, climáticos e de terreno, conseguindo uma diminuição significativa no conjunto de dados: de 44 variáveis disponíveis, 6 foram selecionadas para a análise baseada em objeto e apenas 4 para os dados oriúndos dos pixels. Entretanto Pandit, Tsuyuki e Dube (2018), ao aplicar a metodologia de remoção recursiva das variáveis, não obtiveram nenhum ganho de acurácia no modelo. O

comportamento observado pelos autores foi o de aumento do erro e diminuição do coeficiente de determinação com a eliminação de variáveis.

Embora o método de remoção recursiva das variáveis venha apresentando satisfatórios resultados na redução da dimensão dos dados, ele não explora todas as combinações possíveis existentes entre as variáveis preditoras. Nesse sentido, o algoritmo genético exibe a habilidade de testar diferentes combinações de dados, independentemente da sua medida de importância. Esse método é baseado na teoria evolucionária, sendo programado para realizar buscas por todo o espaço factível de possíveis soluções, utilizando de operações como cruzamento, mutação e seleção para otimizar essa busca (LATIFI; NOTHDURFT; KOCH, 2010).

Garcia-Gutierrez et al. (2014) testaram esse algoritmo evolucionário com outras técnicas usuais de seleção de variáveis para estimar atributos de uma Floresta plantada utilizando dados de LiDAR e regressão linear múltipla, alcançando melhores resultados com as variáveis selecionadas pelo algoritmo genético. Latifi, Nothdurft e Koch (2010) também encontraram resultados superiores do algoritmo genético quando comparado ao método de seleção stepwise para a modelagem da biomassa e volume de uma Floresta, a partir de dados de sensoriamento remoto. Mesmo com sua potencialidade, o método tem sido pouco aplicado aos problemas de mapeamento da vegetação que envolvem grande quantidade de dados de sensores ópticos e de SIG, nos quais a quantidade de variáveis correlacionadas é consideravelmente alta.

Outra questão que exige melhor compreendimento na área de mapeamento da vegetação é quanto à escolha do método de modelagem. Por tradição, técnicas estatísticas representadas pelos métodos paramétricos, como os modelos lineares múltiplos, foram amplamente empregadas nesses estudos (FRANKIN, 1995). No entanto é notório o crescimento da utilização de novas técnicas, especialmente dos algoritmos de aprendizagem de máquina, da área da

inteligência computacional. Essas técnicas são consideradas métodos não paramétricos de predição, ou seja, não fazem nenhuma pré-suposição sobre o comportamento dos dados, o que permite a investigação de diferentes variáveis com atuação pouco conhecida. Já as técnicas paramétricas assumem que a relação entre a variáveis dependente e independentes possui uma estrutura de modelo explícita que pode ser especificada por parâmetros (ZHANG et al., 2018).

As técnicas de aprendizagem de máquina (AM) são capazes de aprender a partir de exemplos, extraindo conhecimento dos dados previamente observados e gerar predições com base em novos dados. O tipo de inferência lógica utilizada pelos algoritmos de AM é a indução, ou seja, o raciocínio originado em um conceito específico é generalizado para o restante dos dados (MITCHELL, 1997). Dentre os diversos algoritmos de aprendizagem de máquina, destaca-se o uso dos algoritmos rede neural artificial (*artificial neural network*), maquina de vetor de suporte (*support vector machine*) e floresta aleatória (*random forest*) tanto para a modelagem da distribuição potencial de espécies, quanto para o mapeamento preditivo da vegetação.

Elith et al. (2006), comparando 16 métodos de modelagem da distribuição potencial de espécies, obtiveram maior eficácia dos modelos baseados nos métodos de aprendizagem de máquina em relação aos modelos bem estabelecidos, como modelos aditivos generalizados (GAM´s) e envelopes climáticos (BIOCLIM). Segurado e Araújo (2004) também encontraram resultados que confirmam o potencial dos métodos de aprendizado de máquina na modelagem da distribuição de espécies, ao comparar sete métodos em 44 espécies de anfíbios.

Os resultados obtidos em estudos que integram variáveis ambientais e espectrais para o mapeamento de variáveis dendrométricas também reforçam o potencial das técnicas de aprendizadem de máquina. Gao et al. (2018) testaram

as técnicas descritas acima e regressão linear múltipla para a estimativa da biomassa acima do solo em florestas tropicais, utilizando a integração de dados provenientes de imagens Landsat e ALOS PALSAR. Os autores confirmam melhores resultados de tais técnicas frente à regressão linear, porém concluem que o modelo da estatística clássica ainda é um método com desempenho satisfatório para a modelagem da biomassa. O estudo de Zhang et al. (2018) também confirma a superioridade desses métodos não paramétricos diante da regressão linear múltipla para a quantificação da biomassa da vegetação costal da Flórida.

Dentre esses métodos paramétricos não há unanimidade quanto ao melhor desempenho, tanto para a modelagem da distribuição potencial de espécies quanto para a quantificação de variáveis dendrométricas em larga escala (ELITH et al., 2006; GAO et al., 2018; GARZÓN et al., 2006). No entanto merece destaque o algoritmo *random forest* (RF), pois além de apresentar grande robustez, possui métricas internas que propiciam mais informações sobre as relações dos dados dentro do modelo, e parametrização simples quando comparado à rede neural artificial e máquina de vetor de suporte (CLUTER et al., 2007).

2.7 *Random forest*

Em 2001, o matemático Leo Breiman (1928 – 2005), conhecido como um dos maiores responsáveis pela confluência entre a estatística e a ciência da computação, publicou na revista *Machine Learning* seu mais famoso trabalho, denominado *Random forests* (RF) (BREIMAN, 2001). Antes desse algoritmo, foi autor dos métodos *Classification and Regression Trees* (CART) (BREIMAN et al., 1984) e *Bagging* (BREIMAN, 1996), imprescindíveis para a construção do RF e, portanto, brevemente descritos neste texto.

O algoritmo RF consiste na coleção de classificadores do tipo árvores (CART) $\{h(x, \Theta_k), k = 1, \ldots\}$, na quak k é igual ao número de árvores da floresta, Θ_k são vetores aleatórios das classes com distribuição idêntica, x o vetor de entrada contendo todos os atributos e h o classificador do tipo árvore. A partir dos dados(x, Θ_k), cada árvore lançará seu voto para cada instância do conjunto total de dados, a classe mais popular em votos é a resposta dada pelo RF para cada dado classificado (BREIMAN, 2001). As árvores de classificação ou regressão (CART) realizam uma categorização hierárquica do conjunto de dados, obtendo regras similares a uma chave de classificação. O algoritmo CART particiona um conjunto de dados heterogêneo (raiz) em classes homogêneas (folhas), gerando regras de classificação com base em atributos (nós). Em cada divisão da árvore, *m* atributos são aleatoriamente selecionados para direcionar o crescimento da árvore. Destes *m* atributos, aquele que melhor dividir a base de dados será utilizado para realizar a divisão naquele nó (Split). Para problemas de regressão o critério de repartição é a soma do quadrado dos resíduos (SQR) de cada folha (1) e para problemas de classificação é utilizado o critério de Gini (2), em que: $\overline{y_d}$ = média dos valores de y para o nó direito; $\overline{y_e}$= média dos valores de y para o nó esquerdo; n_d = número de valores no nó direito; n_e = número de valores no nó esquerdo; p_{kd} = proporção da classe k no nó direito; p_{ke} = proporção da classe k no nó esquerdo. De maneira geral, o valor de *m* deve ser menor do que o número total de atributos, para que possam ser geradas árvores distintas.

$$SQR = \sum_{direito}(y_i - \overline{y_d})^2 + (y_i - \overline{y_e})^2 \quad (1)$$

$$Gini = n_d \sum_{k=1} p_{kd}\,(1 - p_{kd}) + n_e \sum_{k=1} p_{ke}\,(1 - p_{ke}) \quad (2)$$

A propriedade em combinar as predições das árvores, ou de outro tipo de classificadores, é denominada na ciência da computação como *Ensemble* e foi

empregada inicialmente no algoritmo *Bagging*. É importante ressaltar, que também foi desse algoritmo (BREIMAN, 1996) que surgiu a ideia de utilizar o *bootstrap aggregating* (abreviado como *bagging*) para a geração dos vetores aleatórios com mesma distribuição, que irão compor os dados para o crescimento de cada árvore. Esse procedimento garantirá a boa acurácia quando também aplicado à aleatoriedade dos atributos em cada nó das árvores, pois diminui a variância dos dados. Além disso, permitirá a estimação do erro de generalização do ensemble de árvores, por meio das amostras *out-of-bag* (OOB).

No processo de geração dos vetores aleatórios por meio da amostragem com reposição (*bootstraping*), cada instância y_n aparece em cerca de 63% dos vetores aleatórios gerados na floresta, ou seja, 63% das árvores obtiveram a instância y_n em seu aprendizado. O restante, aproximadamente 37%, é utilizado então para medir o erro de generalização do algoritmo, bem como para estimar as métricas de importância de cada atributo, correlação e força das árvores (CLUTER et al., 2007). Para a medida da importância da variável (atributo), os valores da variável de interesse na amostra OOB são permutados e então apresentados ao classificador já treinado. A diferença do erro de classificação entre os dados OOB originais e os dados permutados, em razão do erro padrão, é a medida de importância da variável. Representa, em porcentagem, o aumento no erro de generalização em relação à média do índice de Gini (classificação) ou erro quadrático (regressão). As medidas de força e correlação indicam quão acuradas são cada árvore e qual a dependência entre elas. Para maior detalhamento matemático favor consultar a obra citada (BREIMAN, 2001).

Além das características citadas acima, RF é mais rápido que os métodos *Adaboost* e *Bagging*, e tão acurado quanto o primeiro. Outra grande vantagem é a fácil parametrização. O principal parâmetro é o número de atributos sorteados em cada nó, visto que o autor comprovou que o aumento do número de árvores

não precede maior *overfitting* do modelo. RF também apresenta desempenho satisfatório em dados com poucas amostras (BREIMAN, 2001).

Na área da ecologia, o algoritmo tem ganhado aplicações cada vez em maiores magnitudes e importância. Cluter (2007), ao comparar o RF com outros métodos estatísticos na modelagem de espécies raras e invasivas, listou as principais vantagens encontradas no método (1) alta acurácia na classificação; (2) novo método para estimativa da importância de cada variável; (3) habilidade em modelar relações complexas entre atributos; (4) método flexível na análise dos dados, podendo ser empregado em problemas de classificação, regressão, análise de sobrevivência e aprendizado não supervisionado; (5) algoritmo para atribuição de valores faltantes. É apontado como um dos métodos mais robustos para o mapeamento preditivo da vegetação em estudos comparativos, superando inclusive outras técnicas de aprendizagem de máquina (Redes Neurais Artificiais, Algoritmos Genéticos, Máqunas de Vetores de Suporte, etc.) e métodos de regressão (regressão Logística, modelos mistos, modelos lineares generalizados, etc.) (CLUTER et al., 2007; FUKUDA et al., 2013; GARZON et al., 2006; LORENA et al., 2011).

É também a técnica mais adotada em ações governamentais quando se trata de mudanças climáticas e a ocorrência/produtividade de espécies arbóreas, é empregada pelo Serviço Florestal Norte Americano (USDA) (IVERSON; MCKENZIE, 2013) União Européia (CASALEGNO et al., 2011) e pelo Ministério da Ciência e Tecnologia do Brasil (BRASIL, 2016).

3 CONSIDERAÇÕES FINAIS

O mapeamento preditivo da vegetação constitui uma importante ferramenta para o manejo florestal, permitindo ações em larga escala por meio da replicação das relações entre a vegetação e as variáveis preditoras. No entanto a área carece de novos estudos que verifiquem as potencialidades de utilização da grande quantidade de dados disponíveis, oriundos de sistema de informação geográfica e sensoriamento remoto, bem como de pesquisas que investiguem a utilização de métodos para extrair conhecimento desses dados, permitindo a modelagem e o mapeamento de forma robusta dos aspectos da vegetação.

REFERÊNCIAS

ALTOÉ, T. F. **Sustentabilidade de plantações de Candeia (*Eremanthus erythropappus* (DC.) MacLeish) na produção e qualidade de óleo essencial**. 2012. 153 p. Dissertação (Mestrado em Engenharia Florestal)-Universidade Federal de Lavras, Lavras, 2012.

ANDERSON, R. P.; LEW, D.; PETERSON, A. T. Evaluating predictive modeling of species' distributions: criteria for selecting optimal models. **Ecological Modelling**, Amsterdam, v. 162, p. 211-232, 2003.

BACCINI, A. et al. A first map of tropical Africa's above-ground biomass derived from satellite imagery. **Environmental Research Letters**, Bristol, v. 3, n. 4, 2008. Disponível em: <https://iopscience.iop.org/article/10.1088/1748-9326/3/4/045011/meta>. Acesso em: 10 dez. 2018.

BOISVENUE, C. et al. Integration of Landsat time series and field plots for forest productivity estimates in decision support models. **Forest Ecology and Management**, Amsterdam, v. 376, p. 284-297, 2016.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. Secretaria de Políticas e Programas de Pesquisa e Desenvolvimento. Coordenação-Geral de Mudanças Globais de Clima. **Modelagem climática e vulnerabilidades Setoriais à mudança do clima no Brasil**. Brasília, DF, 2016. 590 p.

BREIMAN, L. Bagging predictors. **Machine Learning**, Boston, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, L. Random forest. **Machine Learning**, Boston, v. 45, p. 5-32, 2001.

BREIMAN, L. et al. **Classification and regression trees**. College Park: Chapman & Hall/CRC, 1984.

BROWN, D. G. Predicting vegetation types at treeline using topography and biophysical disturbance variables. **Journal of Vegetation Science**, Knivsta, v. 5, p. 641-656, 1994.

BRZEZIECKI, B.; KIENAST, F.; WILDI, O. A simulated map of the potential natural forest vegetation of Switzerland. **Journal of Vegetation Science**, Knivsta, v. 4, n. 4, p. 499-508, 1993.

CARVALHO, M. C. et al. Modeling ecological niche of tree species in brazilian tropical area. **Cerne**, Lavras, v. 23, n. 2, p. 229-240, jun. 2017.

CASALEGNO, S. et al. Modelling and mapping the suitability of European forest formations at 1-km resolution. **European Journal of Forest Research**, Georgetown, v. 130, n. 6, p. 971-981, 2011.

CHOU, S. C. et al. Assessment of climate change over South America under RCP 4.5 and 8.5 Downscaling Scenarios. **American Journal of Climate Change**, Wuhan, v. 3, p. 512-525, 2014.

CLUTER, R. D. et al. Random forest for classification in ecology. **Ecology**, Durham, n. 11, p. 2783-2792, 2007.

COELHO, G.; TAVARES, L.; GOMIDE, L. Modelagem preditiva de distribuição de espécies pioneiras no Estado de Minas Gerais. **Pesquisa Agropecuária Brasileira**, Brasília, DF, v. 51, n. 3, p. 207-214, mar. 2016.

DUBE, T.; MUTANGA, O. The impact of integrating WorldView-2 sensor and environmental variables in estimating plantation forest species aboveground biomass and carbon stocks in uMgeni Catchment, South Africa. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 119, p. 415-425, Sept. 2016.

ELITH, J. et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, Copenhagen, v. 29, p. 129-151, 2006.

FAYAD, I. et al. Aboveground biomass mapping in French Guiana by combining remote sensing, forest inventories and environmental data. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 52, p. 502-514, 2016.

FORSTER, P. et al. Changes in atmospheric constituents and in radiative forcing. In: SOLOMON, S. et al. (Ed.). **Climate change 2007:** the physical science basis: contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, 2007. p. 129-234.

FRANK, D. W.; GOETZ, S. Modeling vegetation pattern using digital terrain data. **Landscape Ecology**, Dordrecht, v. 4, n. 1, p. 69-80, 1990.

FRANKLIN, J. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. **Progress in Physical Geography**, London, v. 19, n. 4, p. 474-499, 1995.

FUKUDA, S. et al. Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.) using a broad range of species distribution models. **Enviromental Modelling & Software**, New York, n. 47, p. 1-6, 2013.

GAO, Y. et al. Comparative analysis of modeling algorithms for forest aboveground biomass estimation in a subtropical region. **Remote Sensing**, Basel, v. 10, n. 4, p. 627, 2018.

GARCIA-GUTIERREZ, J. et al. Evolutionary feature selection to estimate forest stand variables using LiDAR. **International Journal of Applied Earth Observation and Geoinformation**, Enschede, v. 26, p. 119-131, 2014.

GARZÓN, M. B. et al. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. **Ecological Modelling**, Amsterdam, n. 197, p. 383-393, 2006.

GIANNINI, T. C. et al. Desafios atuais na modelagem preditiva de distribuição de espécies. **Rodriguésia**, Rio de Janeiro, v. 63, n. 3, p. 733-749, 2012.

GWITIRA, I. et al. Precipitation of the warmest quarter and temperature of the warmest month are key to understanding the effect of climate change on plant species diversity in Southern African Savanna. **African Journal of Ecology**, Oxford, v. 52, n. 2, p. 209-216, 2014.

HIJMANS, R. et al. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, Chichester, v. 25, n. 15, p. 1965-1978, 2005.

HUSCH, B.; MILLER, C. I.; BEERS, T. W. **Forest mensuration**. 3rd ed. New York: Wiley, 1982. 402 p.

HUTCHINSON, G. E. Concluding remarks. **Cold Spring Harbour Symposium on Quantitative Biology**, New York, n. 22, p. 415-427, 1957.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. **Climate change 2014:** synthesis report: contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Geneva, 2014. 151 p.

IVERSON, L.; MCKENZIE, D. Tree-species range shifts in a Changing climate: detecting, modeling, assisting. **Landscape Ecology in Review**, Dordrecht, v. 28, p. 879-889, 2013.

LATIFI, H.; NOTHDURFT, A.; KOCH, B. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. **Forestry**, Oxford, v. 83, n. 4, p. 395-407, 2010.

LIEBERMAN, D. et al. Tropical forest structure and composition on a large-scale altitudinal gradient in Costa Rica. **The Journal of Ecology**, Oxford, v. 84, n. 2, p. 137, 1996.

LOEUILLE, B. ***Eremanthus*** **in lista de espécies da flora do Brasil**. Disponível em: <http://floradobrasil.jbrj.gov.br/2010/FB005312>. Acesso em: 25 jan. 2017.

LORENA, A. C. et al. Comparing machine learning classifiers in potential distribution modelling. **Expert Systems with Applications**, New York, n. 38, p. 5268-5275, 2011.

LOWELL, K. Utilizing discriminant function analysis with a geographical information system to model ecological succession spatially. **International Journal of Geographical Information Systems**, London, v. 5, n. 2, p. 175-191, 1991.

LU, D. et al. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. **International Journal of Digital Earth**, London, v. 9, n. 1, p. 63-105, 2014.

MACKEY, B. G. Predicting the potential distribution of rain-forest structural characteristics. **Journal of Vegetation Science**, Knivsta, v. 5, p. 43-54, 1994.

MACLEISH, N. F. F. Revision of *Eremanthus* (Compositae: Vernonieae). **Annals of the Missouri Botanical Garden**, Saint Louis, v. 47, n. 2, p. 265-290, 1987.

MARENGO, J. A. et al. **Climate Change in Central and South America:** recent trends, future projections, and impacts on regional agriculture. Copenhagen: CGIAR Research Program on Climate Change, Agriculture and Food Security, 2014. (CCAFS Working Paper, 73).

MITCHELL, T. M. **Machine learning**. Boston: WCB/McGraw-Hill, 1997.

MORENO, A.; NEUMANN, M.; HASENAUER, H. Optimal resolution for linking remotely sensed and forest inventory data in Europe. **Remote Sensing of Environment**, New York, v. 183, p. 109-119, 2016.

OLIVEIRA, A. D. O. et al. Economic analysis of sustainable management os Candeia. **Cerne**, Lavras, v. 16, n. 3, p. 335-345, 2010.

PALMER, A. R.; STADEN, J. M. Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. **Journal of Vegetation Science**, Knivsta, v. 3, n. 2, p. 261-266, 1992.

PANDIT, S.; TSUYUKI, S.; DUBE, T. Estimating above-ground biomass in sub-tropical buffer zone community forests, Nepal, Using Sentinel 2 Data. **Remote Sensing**, Basel, v. 10, n. 4, p. 601, 2018.

PAYNE, K.; STOCKWELL, D.; DAVEY, S. A methodology for improving the accuracy of vegetation mapping using GIS, remote sensing and genetic algorithms. In: REGIONAL CONFERENCE OF THE INTERNATIONAL UNION OF GEOGRAPHERS: 'ENVIRONMENT AND THE QUALITY OF LIFE IN CENTRAL EUROPE: PROBLEMS OF TRANSITION', 1994, Prague. **Proceedings…** Prague, 1994.

RANA, B. S.; SINGH, S. P.; SINGH, R. P. Biomass and net primary productivity in Central Himalayan forests along an altitudinal gradient. **Forest Ecology and Management**, Amsterdam, v. 27, n. 3/4, p. 199-218, 1989.

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, 2018. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01431161.2018.1530808>. Acesso em: 10 dez. 2018.

RIBEIRO, S. P. Plant defense against leaf herbivory based on metal accumulation: examples from tropical high altitude ecosystem. **Plant Species Biology**, Hoboken, v. 32, n. 2, p. 147-155, Apr. 2017.

SCOLFORO, H. F. et al. Spatial distribution of aboveground carbon stock of the arboreal vegetation in Brazilian biomes of Savanna, Atlantic Forest and Semi-Arid Woodland. **Plos One**, San Francisco, v. 10, n. 6, 2015. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128781>. Acesso em: 10 dez. 2018.

SCOLFORO, J. R. S. et al. Estimativas de volume, peso seco, peso de óleo e quantidade de moirões para a Candeia (*Eremanthus erythropappus* (DC.) MacLeish). **Cerne**, Lavras, v. 10, n. 1, p. 87-102, 2004.

SCOLFORO, J. R. S.; OLIVEIRA, A. D. de; DAVIDE, A. C. **Manejo sustentável da candeia:** o caminhar de uma nova experiência florestal em Minas Gerais. Lavras: Ed. UFLA, 2012. 329 p.

SEGURADO, P.; ARAÚJO, M. B. An evaluation of methods for modelling species distributions. **Journal of Biogeography**, Oxford, n. 31, p. 1555-1568, 2004.

SILVA, M. A. et al. Análise da distribuição espacial da Candeia (*Eremanthus erythropappus* (DC.) MacLeish) sujeita ao sistema de manejo porta-sementes. **Cerne**, Lavras, v. 14, n. 4, p. 311-316, 2008.

SILVEIRA, C. S. et al. Mudanças climáticas na bacia do rio São Francisco: uma análise para precipitação e temperatura. **Revista Brasileira de Recursos Hídricos**, Porto Alegre, v. 21, n. 2, p. 416-428, 2016.

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, June 2019.

SOBERÓN, J.; PETERSON, A. T. Interpretation of models of fundamental ecological niches and species distributional areas. **Biodiversity Informatics**, Lawrence, v. 2, p. 1-10, 2005.

VÁZQUEZ GARCÍA, J. A.; GIVNISH, T. J. Altitudinal gradients in tropical forest composition, structure, and diversity in the Sierra de Manantlan. **Journal of Ecology**, Oxford, v. 86, n. 6, p. 999-1020, 1998.

WANG, T. et al. Climatic niche models and their consensus projections for future climates for four major forest tree species in the Asia-Pacific region. **Forest Ecology and Management**, Amsterdam, v. 360, n. 15, p. 357-366, 2016.

WANG, T. et al. Projecting future distributions of ecosystem climate niches: uncertainties and management applications. **Forest Ecology and Management**, Amsterdam, v. 279, n. 1, p. 128-140, 2012.

WATANABE, M. et al. Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. **Journal of Climate**, Boston, v. 23, p. 6312-6335, 2010.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, London, v. 52, p. 394-403, 2015.

WHITTAKER, R. H. Gradient analysis of vegetation. **Biological Reviews**, Cambridge, v. 42, p. 207-264, 1967.

WITTEN, I. H.; FRANK, E. **Data mining:** practical machine learning tools and techniques. 2nd ed. San Francisco: M. Kaufmann, 2005. 560 p.

ZHANG, C. et al. Quantification of sawgrass marsh aboveground biomass in the coastal Everglades using object-based ensemble analysis and Landsat data. **Remote Sensing of Environment**, New York, v. 204, p. 366-379, 2018.

ZHANG, L. et al. Consensus forecasting of species distributions: the effects of niche model performance and niche properties. **PLoS One**, San Francisco, v. 10, n. 3, 2015. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120056>. Acesso em: 10 dez. 2018.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 - ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA MODELAGEM DA DISTRIBUIÇÃO POTENCIAL DE HABITATS DE ESPÉCIES ARBOREAS

Mônica Canaan Carvalho[1]; Luciano Cavalcante de Jesus França[2]; Isáira Leite e Lopes[2]; Laís Almeida Araújo[3]; Rubens Manoel dos Santos[4]; Lucas Rezende Gomide[4]

[1]Professora, Instituto Federal do Sudeste de Minas Gerais (IFSULDESTEMG), Barbacena, MG – E-mail: monicacanaan@gmail.com.br; [2]Doutorando (a) em Engenharia Florestal, Universidade Federal de Lavras – UFLA. E-mail: lucianodejesus@florestal.eng.br; isairaleite2010@gmail.com; [3]Mestranda em Engenharia Florestal, Universidade Federal de Lavras – UFLA. E-mail:la_sal@hotmail.com; [4]Professor, Universidade Federal de Lavras (UFLA), Lavras, MG. E-mail: lucasgomide@dcf.ufla.br; rubensmanoel@dcf.ufla.br

Status de publicação: Artigo aceito na revista **Nativa** em 07 de janeiro de 2019.

RESUMO. O estudo teve como objetivo avaliar três métodos de aprendizagem de máquina (árvore de decisão-J48, *Random forest* e redes neurais artificias), na modelagem da distribuição de dez espécies arbóreas mais abundantes em uma sub-bacia do rio São Francisco (MG). Utilizaram-se dados provenientes do Inventário Florestal de Minas, com total de 77 fragmentos amostrados e 2.234 parcelas, nas quais foram computadas a presença/ausência de cada espécie. Empregaram-se 12 variáveis ambientais categóricas procedentes do Zoneamento Ecológico Econômico de Minas Gerais (ZEE/MG), além de variáveis relacionadas ao balanço hídrico do solo (evapotranspiração atual e potencial, aridez e índice *alpha*). A parametrização dos três algoritmos para as dez espécies selecionadas foi feita com o auxílio do algoritmo *cvparameter* do *software* WEKA. Os resultados mostram que os algoritmos testados apresentaram desempenhos estatisticamente iguais em 60% das espécies arbóreas. Os algoritmos *Random forest* e *multilayer perceptron* foram estatisticamente iguais para a espécie *Eugenia dysenterica*, sendo superiores ao algoritmo J48. Contudo, o algoritmo *Random forest* foi superior aos demais para as três espécies do gênero *Qualea*. Conclui-se que o algoritmo *Random forest* apresentou-se como o mais robusto para a modelagem da distribuição potencial de habitat de espécies arbóreas.

Palavras-Chave: Inteligência artificial, árvore de decisão, *Random forest*, Redes Neurais Artificiais.

MACHINE LEARNING ALGORITHMS FOR MODELING THE POTENTIAL DISTRIBUTIONHABITAT OF TREE SPECIES

ABSTRACT: The aim of the present study was to evaluate three methods of machine learning (decision tree-J48, Random Forest and artificial neural networks) to model the potential habitat distribution of the ten most abundant tree species of the São Francisco river watershed. The presence/absence tree species data were from 77 fragments sampled with 2,234 plots. We used 12 categorical environmental variables from the Economic Ecological Zoning of Minas Gerais (ZEE/MG), as well as variables related to soil water balance (current and potential evapotranspiration, aridity and alpha index). The parameterization of the three algorithms was done with cvparameter algorithm of the WEKA software. The results showed the applied algorithms were statistically similar for 60% of the tree species. The Random Forest and multilayer perceptron algorithms were statistically similar considering the *Eugenia dysenterica* and superior to J48 algorithm. However, the Random Forest algorithm was superior to the other for the three species of *Qualea genera*. The conclusion is the Random Forest was the most robust model for the potential distribution habitat of tree species.

Key-Words: artificial intelligence, decision trees, Random Forest, artificial neural networks.

1. INTRODUÇÃO

As alterações no ambiente resultantes da ação do homem têm colocado em risco a distribuição de espécies arbóreas no planeta. A fragmentação de habitat, mudanças no uso da terra e as mudanças climáticas ameaçam a existência e perpetuação delas. Por outro lado, a necessidade crescente de proteção e restauração dos ecossistemas florestais demandam novas tecnologias capazes de entender as relações entre as características do meio ambiente e a ocorrência de espécies (GIANNINI et al., 2012; EHRLÉN; MORRIS, 2015; MORENO-FERNÁNDEZ et al., 2016).

Atualmente, estudos vêm sendo desenvolvidos para prever áreas de desmatamento (SOUZA; MARCO JR, 2014), ambientes favoráveis à invasão de plantas exóticas (CANESSA et al., 2018) e impactos das mudanças climáticas sobre a distribuição de espécies ameaçadas de extinção (QIN et al., 2017) por meio da Modelagem da Distribuição de Espécies (MDE). Esta metodologia é relevante para conservação, ecologia e manejo florestal (HENDERSON et al., 2014; MATEO et al., 2018), visto que direciona tomadas de decisão e implementação de medidas de gestão, de modo a auxiliar na seleção de áreas para conservação ou proteção de espécies, assegurando que estas não sejam enquadradas em categoria de extinção (COSTA et al., 2018).

O ponto inicial da MDE é o uso de coordenadas geográficas precisas dos dados de ocorrência/ausência das espécies, em conjunto com o uso de variáveis climáticas e ambientais, como precipitação, temperatura, relevo, dentre outras. Após o uso de diversos métodos é possível realizar predições espaciais do habitat mais adequado para uma determinada espécie em análise (CHAKRABORTY; JOSHI; SACHDEVA, 2016).

Devido à complexidade do processo de modelagem, no que tange previsões confiáveis, diferentes abordagens de algoritmos e métodos têm sido

aplicadas, como exemplo, os métodos estatísticos (modelos lineares generalizados – GLM e modelos aditivos generalizados – GAM) e de aprendizagem de máquina (Redes Neurais Artificiais – RNA, suport vector machine– SVM, árvores de decisão – CART, Random Forest – RF e entropia máxima – MAXENT), conforme observado nos trabalhos de Paglia et al. (2012); Merowet al. (2014); García-Callejas; Araújo (2016). Pesquisas comprovam que o desempenho dos algoritmos varia de acordo com os dados referentes às espécies e sua distribuição espacial (ROBERTSON et al., 2003; CARVALHO et al., 2017). Não há um consenso de qual o melhor método, já que existem variações dessa natureza, o que decorre em uma lacuna sobre qual a melhor técnica de modelagem e qual algoritmo possui desempenho superior.

Neste sentido, este estudo tem como objetivo avaliar três métodos de aprendizagem de máquina (Árvore de Decisão, *Random Forest* e Redes Neurais Artificiais) na modelagem da distribuição de dez espécies arbóreas mais abundantes em uma sub-bacia hidrográfica do rio São Francisco. Concomitante a este objetivo, pretende-se entender quais são os fatores ambientais que estão mais correlacionados com a distribuição de cada espécie.

2. MATERIAL E MÉTODOS

2.1. Área de estudo e dados de ocorrência

A área amostrada compreende a bacia hidrográfica do rio das Velhas, sub-bacia do rio São Francisco, localizada no estado de Minas Gerais, Brasil. Abrange as regiões do Alto e Médio São Francisco, com área de drenagem de aproximadamente 14.155 km² (Figura 1).

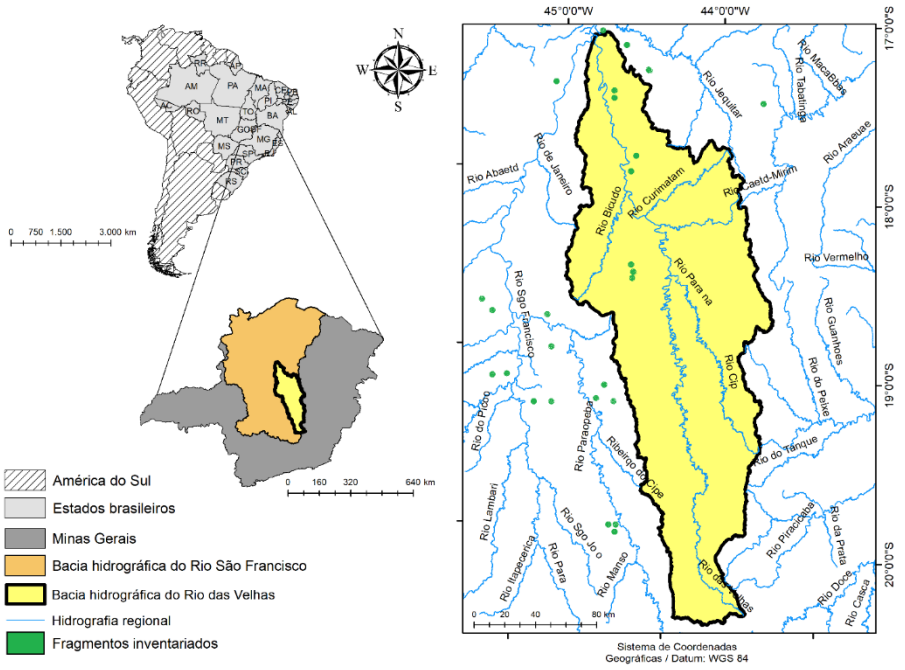


Figura 1. Mapa de localização da área de estudo (Bacia do rio das Velhas), no contexto da América do Sul, Brasil, Minas Gerais e na bacia hidrográfica do Rio São Francisco.

Figure 1. Location map of the study area (Velhas river watershed), in the context of South America, Brazil, Minas Gerais and the São Francisco River Basin.

O parâmetro de seleção das espécies arbóreas a serem modeladas baseou-se no critério das 10 espécies com maior abundância total dentro da bacia. Os dados empregados foram provenientes do Inventário Florestal de Minas Gerais (SCOLFORO et al., 2006), realizado entre os anos de 2006 - 2008. Nesse sentido, um conjunto de 77 remanescentes florestais foram selecionados, totalizando 2.234 parcelas.

Devido à resolução espacial dos dados empregados para as variáveis ambientais, optou-se por trabalhar os dados de ocorrência em nível de fragmento ao invés de parcela, visto que a suficiência amostral foi atingida. Assim, foram extraídos os valores das variáveis ambientais no ponto centroide de cada fragmento, bem como atribuído o valor de presença (1) ou ausência (0). Desta forma, o conjunto de treinamento dos algoritmos foi constituído por 77 observações por espécie, nas quais estão disponíveis os valores das variáveis independentes e a ocorrência da espécie, por meio de valores binários (0 – ausência / 1 – presença). Trabalhou-se com classes de presença/ausência desbalanceadas, sendo o número de presença variável entre 30 e 47 de acordo com cada espécie.

2.2. Variáveis Ambientais

No total um conjunto inicial de 39 variáveis indepentes foi compilado. 20 variáveis ambientais foram selecionadas provenientes da base de dados do World Clim (HIJMANS et al., 2005), com resolução inicial de 1 km. Além destas variáveis, empregaram-se ainda 12 variáveis ambientais categóricas procedentes do Zoneamento Ecológico Econômico de Minas Gerais (ZEE/MG), com resolução espacial variando entre 30 a 270 metros (SCOLFORO; CARVALHO; OLIVEIRA, 2008). Utilizaram-se também variáveis relacionadas ao balanço hídrico do solo (evapotranspiração atual e potencial, aridez e índice alpha) oriundas da base de dados CGIAR-CSI (TRABUCCO; ZOMER, 2010)

com resolução espacial original de 1 km. Os valores de latitude e longitude também entraram como variáveis ambientais devido à seu poder de síntese de condições ambientais. Para a classificação dos solos foi utilizado o mapa de solos de Minas Gerais (SEMAD, 2010). Todas as variáveis, quando necessário, foram transformadas em formato raster com resolução de 270 m e projetadas para o sistema de coordenadas South America Albers Equal Area Conic. Essas variáveis assumem valores numéricos ou categóricos.

De posse do conjunto de dados de treinamento contendo as 39 variáveis ambientais, aplicou-se então o algoritmo de seleção de atributos Correlation-based feature selection (CFS). A seleção foi realizada utilizando o método de validação cruzada com a formação de 10 subconjuntos. Foi adotado o critério de seleção como sendo as 4 variáveis ambientais mais escolhidas pelo algoritmo nos dez subconjuntos

2.3. Processo de modelagem

As etapas do processo de modelagem variam de acordo com o objetivo do trabalho, algoritmos, base de dados e software utilizado. Assim, foi desenvolvido um fluxograma (Figura 2) adaptado de Garzón et al. (2006), para representar e ordenar as etapas da modelagem utilizada nesta pesquisa.

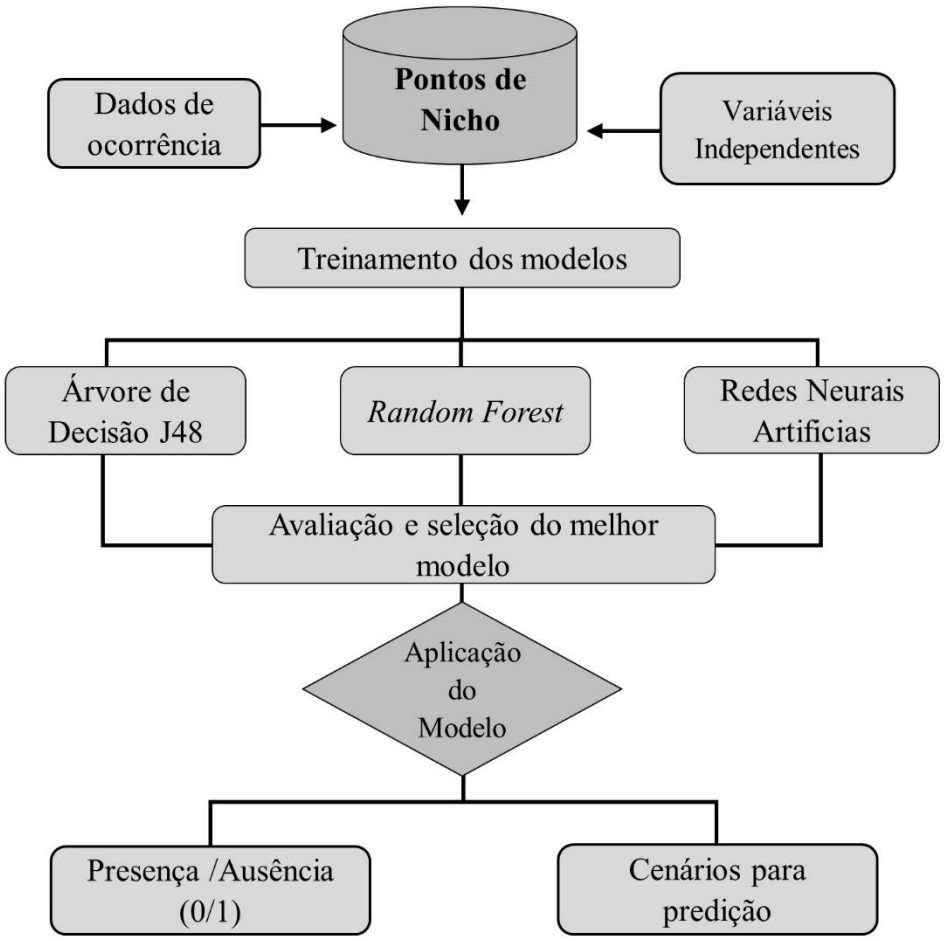


Figura 2. Fluxograma das etapas metodológicas para modelagem de distribuição de espécies florestais.

Figure 2. Flowchart of the methodological steps for modeling the distribution of forest species.

O software utilizado neste artigo para o treinamento e aplicação dos algoritmos foi o WEKA – Waikato Environment for Knowledge Analysis (GARNER, 1995). Neste estudo optou-se por testar Árvore de Decisão

(QUINLAN, 1993), devido à sua simplicidade e legibilidade; *Random Forest* (BREIMAN, 2001), pelos bons resultados apresentados em outras pesquisas (INZA et al., 2009; BHERING et al., 2016); e Redes Neurais Artificiais pela robustez e habilidade em lidar com dados muito complexos.

A parametrização dos três algoritmos para as dez espécies selecionadas foi feita com o auxílio do algoritmo *cvparameter* implementado no software WEKA. É um meta-classificador que testa vários valores (pré-definidos) para diferentes parâmetros de cada algoritmo. Para o algoritmo J48 (Árvore de Decisão) foram avaliados os parâmetros *seed* (1 a 10), *numFolds* (1 a 10) e *confidence Factor* (0,1 a 0,9). No algoritmo *Random Forest* foram testados os parâmetros *numTrees* (1 a 15) e *seed* (1 a 10). Para o algoritmo *Multilayer Perceptron* (Redes Neurais Artificiais) testou-se os parâmetros *hidden Layers* (0 a 2), *learning Rate* (0,1 a 0,9) e *momentum* (0,1 a 0,9). Para cada espécie foi estabelecido o número de 10 repetições por algoritmo, sendo que a avaliação foi feita por validação cruzada, com 10 repetições. As configurações que apresentaram maior AUC (area under the curve) em cada algoritmo foram selecionadas para nova aplicação na base de dados de treinamento.

Após a parametrização e escolha das configurações otimizadas para cada algoritmo por espécie, os modelos foram aplicados novamente no conjunto de treinamento. Foi realizado um experimento, utilizando 10 iterações e validação cruzada com 10 sub-amostragens, em que comparou-se os valores da métrica área abaixo da curva ROC (AUC- *area under the curve ROC*) obtidos por cada modelo (JIMÉNEZ-VALVERDE, 2011). Utilizou-se o teste estatístico T-pareado a 95% de confiança ou probabilidade, já que se trata da mesma base de dados.

2.4. Seleção dos atributos principais

A fim de diminuir o número de atributos, complexidade dos modelos e determinar quais variáveis ambientais são mais representativas na distribuição de determinada espécie, foi aplicado um algoritmo de seleção (CfsSubsetEval) implementado no WEKA. Este algoritmo primeiramente calcula uma matriz de correlação entre as variáveis ambientais e ocorrência, além de uma matriz de correlação entre as variáveis ambientais. Em seguida calcula o mérito (score) para cada subconjunto formado utilizando a equação 1. Nesta equação, o numerador pode ser interpretado como o poder preditivo do subconjunto de atributos e o denominador como o grau de redundância existente entre os atributos.

Neste sentido, o correlation-based feature selection (CFS) começa com um conjunto vazio de atributos e utiliza a heurística best-first-search como algoritmo de busca, na qual o critério de parada é 5 subconjuntos consecutivos que não melhoram o mérito calculado pelo algoritmo.

$$\text{Mérito(S)} = \frac{k \times \overline{r_{ac}}}{\sqrt{k + k(k-1)\overline{r_{aa}}}} \qquad \text{(Equação 1)}$$

em que: k = número de atributos; $\overline{r_{ac}}$ = média de correlação entre atributo-classe; $\overline{r_{aa}}$ = média de correlação entre atributo-atributo.

3. RESULTADOS

Após a análise dos dados do inventário, chegou-se ao resultado de que as dez espécies mais abundantes na bacia do São Francisco, são: *Anadenanthera colubrina* (Vell.) Brenan, *Eugenia dysenterica* DC., *Hymenaea stignocarpa* Mart. Ex Hayne, *Lafoensia vandelliana* Cham. &Schltdl., *Magonia pubescens* S. St.-Hil., Pouteriaramiflora (Mart.) Radlk., *Qualea grandiflora* Mart., *Qualea multiflora* Mart., *Qualea parviflora* Mart. e *Terminalia fagifolia* Mart. Os

resultados da seleção de variáveis por espécie são apresentados na Tabela 1, e os valores da métrica de avaliação AUC para as diferentes técnicas por espécie são apresentados na Tabela 2.

Tabela 1. Porcentagem de seleção dos atributos utilizando o algoritmo CFS com validação cruzada.

Table 1. Percentage of attribute selection using the CFS algorithm with cross-validation.

N	Variáveis Ambientais	Espécies arbóreas									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	Altitude	100	100	100	100	100	100	100	100	100	0
2	Aridez	40	0	0	0	0	0	0	0	0	0
3	Declividade	0	0	0	0	0	10	0	0	0	0
4	Clima (Classe Thornthwaite)	0	0	0	0	0	0	0	0	0	0
5	Erodibilidade	0	0	0	0	0	0	0	0	0	0
6	Evapotranspiração atual	20	50	0	0	90	0	0	0	10	0
7	Evapotranspiração potencial	10	0	0	0	0	0	0	0	0	10
8	Fitofisionomias	10	100	20	0	0	0	50	90	40	10
9	Grau de conservação da vegetação	0	0	0	0	0	0	0	10	0	0
10	Grau de erosão	0	0	0	0	0	0	0	0	0	0
11	Grau de exposição do solo	50	0	0	0	0	0	0	0	0	0
12	Índice alpha	0	0	0	0	100	0	0	0	0	0
13	Intensidade da chuva	0	0	10	0	0	0	0	0	0	0
14	Isotermalidade	40	0	0	0	0	10	0	0	0	40
15	Lâmina explotável	0	0	0	0	0	0	0	0	0	0
16	Latitude	0	0	0	0	0	0	0	10	0	100
17	Longitude	80	0	0	0	0	0	0	20	0	30
18	Variação média diurna da temperatura	0	0	0	0	0	0	0	0	0	10
19	Precipitação anual	10	0	0	0	0	10	0	50	0	0
20	Precipitação do mês seco	0	0	0	40	0	10	0	0	0	10
21	Precipitação do mês úmido	20	60	0	0	0	0	20	0	0	0
22	Precipitação do trimestre frio	60	10	20	10	0	0	10	0	0	80
23	Precipitação do trimestre quente	0	10	0	0	30	0	0	0	0	20
24	Precipitação do trimestre seco	50	90	20	70	10	0	60	40	50	0

25	Precipitação do trimestre úmido	60	10	0	0	0	10	0	0	0	0
26	Qualidade da água	0	0	30	0	80	0	0	0	0	0
27	Rendimento específico	0	0	0	0	0	0	0	0	0	0
28	Sazonalidade da precipitação	0	100	100	0	10	0	40	40	30	10
29	Sazonalidade da temperatura	0	0	0	0	0	0	0	0	0	100
30	Taxa de decomposição da matéria orgânica	0	0	0	0	30	0	0	0	0	0
31	Temperatura anual média	0	0	0	0	0	10	0	0	0	10
32	Temperatura máxima do mês mais quente	0	0	30	0	50	0	10	60	10	0
33	Temperatura média do trimestre frio	0	90	80	0	10	0	70	0	80	0
34	Temperatura média do trimestre quente	90	0	0	0	0	0	0	0	0	0
35	Temperatura média do trimestre seco	0	0	0	0	0	0	0	0	0	20
36	Temperatura média do trimestre úmido	60	0	0	0	0	10	0	0	0	0
37	Temperatura mínima do mês mais frio	40	10	0	0	30	50	0	50	10	100
38	Tipo de solo	90	0	0	50	20	100	0	10	0	90
39	Variação anual da temperatura	0	10	0	90	20	0	20	0	20	80

Em que: (1) *Anadenanthera colubrina* (Vell.) Brenan*;* (2) *Eugenia dysenterica DC.*; (3) *Hymenaea stignocarpa* Mart. ex Hayne ;(4) *Lafoensia vandelliana* Cham. &Schltdl.; (5) *Magonia pubescens* S. St.-Hil.; (6) *Pouteria ramiflora*(Mart.) Radlk.; (7) *Qualea grandiflora* Mart.; (8) *Qualea multiflora* Mart.; (9) *Qualea parviflora* Mart.; (10) *Terminalia fagifolia* Mart.

Do conjunto total de 39 variáveis, 34 variáveis foram selecionadas pela metodologia CFS ao menos uma vez para alguma espécie, e 5 variáveis não foramselecionadas nenhuma vez para nenhuma das espécies, sendo elas: Clima (classes *Thornthwaite*), rendimento específico, lâmina explotável (relacionados a disponibilidade de água superficial e subterrânea), grau de erosão e erodibilidade. A altitude foi a variável que mais se destacou entre as demais, sendo selecionada em 100% das repetições para 9 espécies (Tabela 1).

Entre as dez espécies arbóreas modeladas, em seis delas os algoritmos apresentaram desempenhos (AUC) estatisticamente iguais. Para a espécie *Eugenia dysenterica DC.* os algoritmos *Random forest* e *multilayer perceptron* apresentaram desempenho superior ao J48, porém estatisticamente iguais entre si. Para as espécies *Qualea grandiflora Mart., Qualea multiflora Mart.* E *Qualea parviflora Mart.* o modelo *Random forest* obteve uma diferença significativa frente aos modelos testados (Tabela 2).

Tabela 2. Resultado do teste T-pareado (0,05) entre os valores de AUC obtidos pelos três modelos nas dez espécies arbóreas.

Table 2. Results of the T-paired test (0.05) between the AUC values obtained by the three models in the ten tree species.

Espécies	***Area Under the Curve* (AUC)**		
	J48	***Random forest***	***Multilayer Perceptron***
Anadenanthera colubrina	0,73	0,73	0,80
Eugenia dysenterica	0,72	0,96*	0,89*
Hymenaea stignocarpa	0,90	0,96	0,85
Lafoensia vandelliana	0,80	0,75	0,75
Magonia pubescens	0,70	0,78	0,75
Pouteria ramiflora	0,68	0,75	0,78
Qualea grandiflora	0,85	0,97*	0,94
Qualea multiflora	0,70	0,84*	0,78
Qualea parviflora	0,82	0,94*	0,87
Terminalia fagifolia	0,69	0,70	0,79

*Diferença estatístisca de acordo com o teste T-pareado a 0,05 de significância.

Após a aplicação do teste T-pareado, os algoritmos que obtiveram melhor desempenho (maior AUC) por espécie, foram então aplicados para modelar a distribuição destas espécies arbóreas na sub-bacia do rio das Velhas. Os resultados desta aplicação por espécie podem ser visualizados na Figura 3.

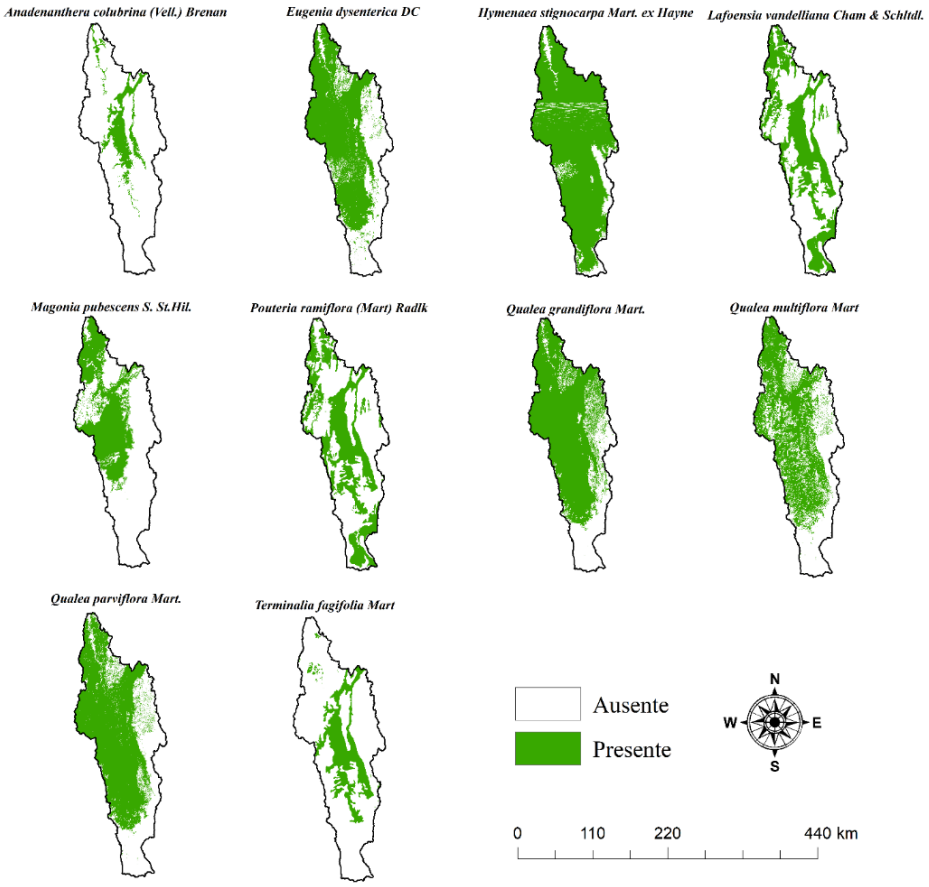


Figura 3. Mapa da distribuição potencial das espécies arbóreas florestais.

Figure 3. Map of potential distribution of forest tree species.

4. DISCUSSÕES

A distribuição das espécies resulta de uma série de fatores, como as características ambientais. A variável altitude, por exemplo, correlacionou-se com 9 das 10 espécies analisadas, seguido das variáveis precipitação dos trimestres seco, sazonalidade da precipitação e tipo de solo (8, 7 e 6 espécies, respectivamente). Um comportamento semelhante foi evidenciado por Callegaro et al. (2018), em que a maior parte das espécies mais abundantes foram

influenciadas pela altitude. Chakraborty; Joshi; Sachadeva, (2016), ao avaliarem a distribuição de quatros espécies florestais, em relação aos impactos das mudanças climáticas, encontraram que a variável elevação tem forte relevância por aumentar significativamente a acurácia do modelo de distribuição. Também constataram que a tipologia de solos, como variável preditora da distribuição de espécies, influencia na melhoria do desempenho do modelo (WAN; WANG; YU, 2017). Em relação à precipitação, Amissah et al. (2014), identificaram este fator como o principal na distribuição de espécies arbóreas.

Os resultados observados indicam que o desempenho de cada algoritmo está intrinsecamente relacionado ao tipo de distribuição da espécie modelada, como também pode ser constatado em estudos anteriores por Segurado e Araújo (2006); Elith et al. (2006); Pearson et al. (2006). Apesar de alguns modelos serem estatisticamente superiores frente aos demais para determinadas espécies, não há consenso de um método superior para todas as circunstâncias.

Lorena et al. (2011), comparando 9 classificadores de aprendizado de máquina na modelagem da distribuição de 35 espécies vegetais da família Bignoniaceae, obteve resultados satisfatórios com o algoritmo *Random Forest*. O método apresentou melhor desempenho em 29 espécies, dentre as 35 testadas. Nesta pesquisa o desempenho desse algoritmo também foi notável, sendo superior em 6 das 10 espécies modeladas. Em ambas as pesquisas o modelo apresentou desempenho estável (baixa variação entre os valores de AUC). Carvalho et al. (2017) compararam o *Random Forest* e as Redes Neurais Artificiais para modelar o nicho ecológico de quatro espécies florestais. O método *Random Forest* apresentou melhor desempenho para a modelagem de distribuição de todas as espécies.

5. CONCLUSÕES

Os resultados obtidos nesta pesquisa, assim como em outros estudos, demonstram que o desempenho dos modelos está intrinsecamente relacionado à espécie modelada. No entanto, dentre os algoritmos testados, o *Random Forest* surge como uma opção robusta para a modelagem da distribuição de espécies. Dentre as variáveis testadas, altitude, precipitação dos trimestres seco, sazonalidade da precipitação e tipo de solo destacaram-se como as variáveis que mais influencial a distribuição das espécies arbóreas dentro da área de estudo.

Diante das ações antropogênicas sobre os ecossistemas florestais naturais e, seus impactos negativos sobre os serviços ecológicos prestados, a utilização da modelagem da distribuição de espécies pode auxiliar em estratégias para proteção conservação e restauração da biodiversidade.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AMISSAH, L.; MOHREN, G. M.; BONGERS, F.; HAWTHORNE, W. D.; POORTER, L. Rainfall and temperature affect tree species distribution in Ghana. **Journal of Tropical Ecology**, v. 30, n. 5, p. 435-446, 2014. DOI: http://dx.doi.org/10.1017/S026646741400025X

BHERING, S. B.; CHAGAS, C. S.; JUNIOR, W. C.; PEREIRA.; FILHO, B. C.; PINHEIRO, H. S. K. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. **Pesq. Agropec. Bras**., v.51, n.9, p. 1359-1370, 2016. DOI: http://dx.doi.org/10.1590/S0100-204X2016000900035

CALLEGARO, R. N.; ARAUJO, M. M.; LONGHI, S. J.; ANDRZEJEWSKI, C. Influência de fatores ambientais sobre espécies vegetais em florestas estacionais para uso potencial em restauração. **Nativa**, v. 6, n. 1, p. 91-99, 2018. DOI: http://dx.doi.org/10.31413/nativa.v6i1.4728

CANESSA, R. SALDAÑA, A.; RÍOS, R.S.; GIANOLI, E. Functional trait variation predicts distribution of alien plant species across the light gradient in a temperate rainforest. **Perspectives in Plant Ecology, Evolution and Systematics**, v. 32, p. 49 - 55, 2018. DOI: https://doi.org/10.1016/j.ppees.2018.04.002

CARVALHO, M. C.; GOMIDE, L.R.; SANTOS, R.M.; SCOLFORO, J.S.; CARVALHO, L.M.T. ; MELLO, J.M. Modeling ecological niche of tree species in brazilian tropical area. **CERNE**, v. 23, n. 2, p.229-240, jun. 2017.DOI:https://doi.org/10.1590/01047760201723022308

CHAKRABORTY, A.; JOSHI, P. K.; SACHDEVA, K. Predicting distribution of major forest tree species to potential impacts of climate change in the central Himalayan region. **Ecological Engineering**, v. 97, p. 593-609, 2016. DOI: http://dx.doi.org/10.1016/j.ecoleng.2016.10.006

EHRLÉN, J.; MORRIS, W. F. Predicting changes in the distribution and abundance of species under environmental change. **Ecology Letters**, v. 18, n. 3, p. 303-314, 2015.DOI: https://doi.org/10.1111/ele.12410

ELITH, J.; GRAHAM, C.H.; ANDERSON, R.P.; DUDÍK, M.; FERRIER, S.; GUISAN, A.; HIJMANS, R.; HUETTMANN, F.; LEATHWICK, J.R.; LEHMANN, A.; LI, J.; LOHMANN, L. G.; LOISELLE, B. A.; MANION, G.; MORITZ, C.; NAKAMURA, M.; NAKAZAWA, Y.; OVERTON, J. McC.; PETERSON, T.; PHILLIPS, S. J.; RICHARDSON, K.; SCACHETTI-PEREIRA, R.; SCHAPIRE, R.; SOBERÓN, J.; WILLIAMNS, S.; WISZ, M. S.; ZIMMERMANN, N. E. Novel methods to improve prediction of species distributions from occurrence data. **Ecography**, n. 29, p. 129-151, 2006. DOI: https://doi.org/10.1111/j.2006.0906-7590.04596.x

GARCÍA-CALLEJAS, D.; ARAÚJO, M. B. The effects of model and data complexity on predictions from species distributions models. **Ecological modelling**, v. 326, p. 4-12, 2016. DOI: https://doi.org/10.1016/j.ecolmodel.2015.06.002

GARNER, S. R. Weka: the Waikato environment for knowledge analysis. In: Proc. of the New Zealand Computer Science Research Studants Conference, p. 57-64, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.3371>.

GARZÓN, M. B. ; BLAZEK, R.; NETELER, M.; DIOS, R.S.; OLLERO, H.S.; FURLANELLO, C. Predicting habitat suitability with machine learning models:

The potential area of *Pinus sylvestris* L. um the Iberian Peninsula. **Ecological Modelling**, n. 197, p. 383-393, 2006.DOI:https://doi.org/10.1016/j.ecolmodel.2006.03.015

GIANNINI, T. C. SIQUEIRA, M.F.; ACOSTA, A.L.; BARRETO, F.C.C.; SARAIVA, A.M.; ALVES-DOS-SANTOS, I. Desafios atuais na modelagem preditiva de distribuição de espécies. **Rodriguésia**, v. 63, n. 3, p. 733-749, 2012. Disponível em:http://rodriguesia-seer.jbrj.gov.br/index.php/rodriguesia/article/view/339

HALL, M. A. Correlation-based Feature Subset Selection for Machine Learning. Department of Computer Science, University of Waikato.Hamilton, New Zealand, 1999. Disponível em: https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

HENDERSON, E. B.; OHMANN, J. L.; GREGORY, M. J.; ROBERTS, H.M.; ZALD, H. Species distribution modelling for plant communities: stacked single species or multivariate modellingapproaches? **Applied vegetation science**, v. 17, n. 3, p. 516-527, 2014. DOI : https://doi.org/10.1111/avsc.12085

HIJMANS, R.J. ; HIJMANS, R.J.; CAMERON, S.E.; PARRA, J.L.; JONES, P.G.; JARVIS, A. Very highresolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, n. 25, p. 1965-1978, 2005. DOI: https://doi.org/10.1002/joc.1276

INZA, I.; CALVO, B.; ARMANANZAS, R.; BENGOETXEA, E.; LARRANAGA, P.; LOZANO, J. A. Machine Learnign: An Indispensable Tool in Bioinformatics. **Bioinformatics Methods in Clinical Research**, p.25-48, 2009. DOI: https://doi.org/10.1007/978-1-60327-194-3_2

JIMÉNEZ-VALVERDE, A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. **Global Ecology and Biogeography**, v. 21, n.4, p. 498 – 507, 2011. DOI: https://doi.org/10.1111/j.1466-8238.2011.00683.x

LORENA, A. C.; JACINTHO, L. F. O.; SIQUEIRA, M. F.; GIOVANNI, R.; LOHMANN, L. G.; CARVALHO, A. P. L. F.; YAMAMOTO, M. Comparing machine learning classifiers in potential distribution modelling. **Expert Systems with Applications**, n. 38, p. 5268 – 5275, 2011. DOI: https://doi.org/10.1016/j.eswa.2010.10.031

MATEO, R. G. ; GASTÓN, A.; AROCA-FERNANDEZ, M.J.; SAURA, S.; GARCÍA-VIÑAS, J. I. Optimization of forest sampling strategies for woody plant species distribution modelling at the landscape scale. **Forest Ecology and Management**, v. 410, p. 104-113, 2018. DOI: https://doi.org/10.1016/j.foreco.2017.12.046

MEROW, C.; SMITH, M.J.; EDWARDS, T. C.; GUISAN, A. McMAHON, S.M.; NORMAND, S.; THUILLER, W.; WUEST, R.O.; ZIMMERMANN, N.E.; ELITH, J.What do we gain from simplicity versus complexity in species distribution models?. **Ecography**, v. 37, n. 12, p. 1267-1281, 2014. DOI: https://doi.org/10.1111/ecog.00845

MORENO-FERNÁNDEZ, D.; Space–time modeling of changes in the abundance and distribution of tree species. **Forest Ecology and Management**, v. 372, p. 206-216, 2016. DOI: https://doi.org/10.1016/j.foreco.2016.04.024

PAGLIA, A. P.; REZENDE, D. T.; KOCH, I.; KORTZ, A. R.; DONATTI, C. Modelos de distribuição de espécies em estratégias para a conservação da biodiversidade e para adaptação baseada em ecossistemas frente a mudanças climáticas. **Natureza & Conservação**, v. 10, n. 2, p. 231-234, 2012.DOI: http://dx.doi.org/10.4322/natcon.2012.031

QIN, A.; LIU, B.; GUO, Q.; BUSSMANN, R.W.; MA, F.; JIAN, Z.; XU, G.; PEI, S. Maxent modeling for predicting impacts of climate change on the potential distribution of *Thuja sutchuenensis* Franch., an extremely endangered conifer from southwestern China. **Global Ecology and Conservation**, v. 10, p. 139-146, 2017.DOI: https://doi.org/10.1016/j.gecco.2017.02.004

QUINLAN, J.R. C4.5: programs for Machine Learning. Elsevier: 302p., 1993. ROBERTSON, M. P.; PETER, C. I.; VILLET, M.; RIPLEY, B.S. Comparing models for predicting species' potential distributions: a case study using correlative and mechanism predictive modelling techniques. **Ecological Modelling**, n. 164, p. 153-167, 2003.DOI: https://doi.org/10.1016/S0304-3800(03)00028-0

SCOLFORO, J. R. S.; CARVALHO, L. M. T.; OLIVEIRA, A. D. Zoneamento ecológico-econômico do Estado de Minas Gerais: componentes geofísico e biótico. Lavras: Editora UFLA, 161 p., 2008.

SEGURADO, P.; ARAÚJO, M. B. An evaluation methods for modeling species distributions. **Journal of Biogeography**, n. 31, v. 10, p. 1555-1568, 2004.DOI: https://doi.org/10.1111/j.1365-2699.2004.01076.x

SOUZA, R. A.; MARCO J. R. P. The use of species distribution models to predict the spatial distribution of deforestation in the western Brazilian Amazon. **Ecological Modelling**, v. 291, p. 250-259, 2014. DOI: https://doi.org/10.1016/j.ecolmodel.2014.07.007

TRABUCCO, A.; ZOMER, R. J. Global Soil Water Balance Geospatial Database. CGIAR Consortium for Spatial Information. 2010. Published online, available from the CGIAR-CSI GeoPortal at: http://www.cgiar-csi.org

WAN, J. Z.; WANG, C. J.; YU, F. H. Modeling impacts of human footprint and soil variability on the potential distribution of invasive plant species in different biomes. **Acta Oecologica**, 85, 141-149, 2017. DOI: https://doi.org/10.1016/j.actao.2017.10.008

ARTIGO 2 - POTENTIAL AND FUTURE GEOGRAPHICAL DISTRIBUTION OF *Eremanthus erythropappus* (DC.) MACLEISH: A TREE THREATENED BY CLIMATE CHANGE.

Mônica Canaan Carvalho[1]*; Lucas Rezende Gomide[1]; Fausto Weimar Acerbi Junior[1]; David Yue Phin Tng[2]

[1]Department of Forestry Sciences, Federal University of Lavras, Av. Doutor Sylvio Menicucci, 1001, Aquenta Sol, 37200-000, Lavras, Minas Gerais, Brazil
[2]Institute of Biology, Federal University of Bahia, R. Barão Jeremoabo, Ondina, 40170-115 Salvador, Bahia, Brazil.

Status de publicação: Artigo submetido na revista **Floram** em 07 de janeiro de 2019.

Abstract

Eremanthus erythropappus is a commercially-important tree which has a long history of exploitation in the Brazilian State of Minas Gerais. The knowledge of the potential and future potential geographical distribution of *E. erythropappus* is therefore critical for the sustainability of the species. Thus, the aim of this study was to estimate and map current and future ecological niche for *E. erythropappus* in Minas Gerais. We used the Random Forests algorithm to model the ecological niche for current and future climates scenarios. Our predictions indicate Espinhaço, Mantiqueira, and Canastra mountain ranges as core areas of distribution and forecast drastic reductions in potential areas under all climate scenarios. Based on our results, we highlight that the continual harvesting of naturally-occurring *E. erythropappus* populations will not be able to supply the market demand. Silviculture practices would likely serve as be an economically viable and ecological sustainable alternative to harvesting natural populations.

Additional keywords: Candeia trees; Random Forests; Habitat suitability.

Introduction

Eremanthus erythropappus (DC.) MacLeish, (commonly-known as Candeia), is a commercially-important tree which occupy transition areas between the Brazilian Atlantic Rain Forest and Savannah Biomes. The bulk of the naturally-occurring populations of the species in South America are within the state of Minas Gerais, Brazil, particularly in upland areas (>900m) (Silva *et al.*, 2008; Clark *et al.*, 2011).

Eremanthus erythropappus trees are light-demanding ecotone specialists that are often found in open grasslands and edge of forests, and rarely under the forest canopies (Oliveira Filho & Fluminhan Filho, 1999; Araújo *et al.* 2017). Habitats with a high dominance of *E. erythropappus* (≥ 70% frequency) trees are locally known as "Candeais", and these habitats often represent a transitional vegetation zone between closed forest and open savannah habitats (Oliveira Filho & Fluminhan Filho, 1999; Araújo *et al.* 2017).

This tree species has a long history of exploitation, either for use as fence posts, firewood, and on a larger economic scale, for essential oil extraction (Clark *et al.*, 2011; Donadelli, 2012; Scolforo H *et al.*, 2016). In the case of essential oil extraction, the principal component targeted is alpha-Bisabolol (Scolforo *et al.*, 2016), which is a key ingredient for skincare and global cosmetics industries (Vieira *et al.*, 2012; Scolforo *et al.*, 2016). Despite the availability of synthetic substitutes of alpha-Bisabolol (Clark *et al.* 2011) and also the lengthy period of tree growth which spans at least 12-15 years (Pérez *et al.*, 2004; Mori *et al.*, 2009), naturally-occurring populations of *E. erythropappus* are still important sources of alpha-Bisabolol for the global market (GFA, 2006), and comprises 80% of the alpha-Bisabolol exported (Clark *et al.*, 2011).

Before regulations on *E. erythropappus* extraction, the species was often exploited in an unsustainable fashion. In 2004, regulations were placed on *E.*

erythropappus exploitation in the Minas Gerais state, and in 2007, further regulations were implemented on the quantity of *E. erythropappus* that may be harvested and types of exploitation management systems (Scolforo *et al*., 2012).

Nevertheless, alpha-Bisabolol is still being purchased in the European market without any requirement of traceability, and to date only two companies, Atina and Citróleo, possess the Forest Stewardship Council certification regarding suitable management of *E. erythropappus* populations (Donadelli, 2012). Also, the increasing global demand for natural alpha-Bisabolol has not been matched by adequate or sufficient mechanisms of regulation and governance, and thus illegal extraction and unsustainable harvesting practices of the species is still rampant (Clark *et al.*, 2011; Donadelli, 2012).

Due to these factors, *E. erythropappus* is increasingly being threatened by human activities. Additionally, there is the risk, as yet unexplored, of climate changes negatively affecting the future distribution and survival of the species. It is well-known that each species has its climatic niche, and species that are within their climatic niche will have a better chance of remaining healthy, productive and able to maximize their ecological and/or economic value under changing climates (Wang *et al.*, 2016). Species growing outside their climatic niche are therefore likely to have limited productivity and to be more vulnerable to disturbances (Hamann & Wang, 2006; Kurz *et al.*, 2008).

The knowledge of the potential and future geographical distribution of *E. erythropappus* is therefore critical for the sustainability of the native populations, as well as to inform silvicultural practices to ensure productivity of *E. erythropappus* plantations. The aims of our study are therefore to: (1) model and map the current ecological niche of *E. erythropappus* in the state of Minas Gerais, and identify the core areas of *E. erythropappus* occurrence; (2) predict and assess the changes in the potential distribution of *E. erythropappus* under scenarios of climate change by the years 2050 and 2070, and; (3) access the

adequacy of conservation units for protecting the current and future populations of the species.

Material and methods

Study region

The study region is the Brazilian State of Minas Gerais, with a land area of over 586,000 km², equivalent to the area of countries such as France and Spain (Figure 1). Minas Gerais harbors the largest occurrence of *E. erythropappus* populations, and also the highest number of extractive industries of alpha-Bisabolol (Donadelli, 2012). The region also encompasses a wide altitudinal range (between 40 and 2,600 meters) and five climate classes according to the climate classification system of Koppen (Martins *et al.*, 2018). This variability gives rise to the occurrence of different vegetation biomes (Savannah, rainforest and semi-arid woodland) and the diverse phytogeography within the region. Savannah covers 57% of the area of state (west/northwest), with warmer and wetter climate during summer, and a pronounced dry period. Rainforest, locally-known as Atlantic forest, dominates the south and east regions occupying over 41% of state, and features wetter and milder climates, particularly in the south. Finally, semi-arid woodland comprises only 2% of the state, predominantly in the north, and is characterized by semi-arid and sub-humid climates, with higher temperatures and lower rainfall throughout the year. Oxisols represent the predominant soil type in the state, with areas of Cambisols and Quartzarenic/Litholic Neosols (Curi *et al.*, 2008).

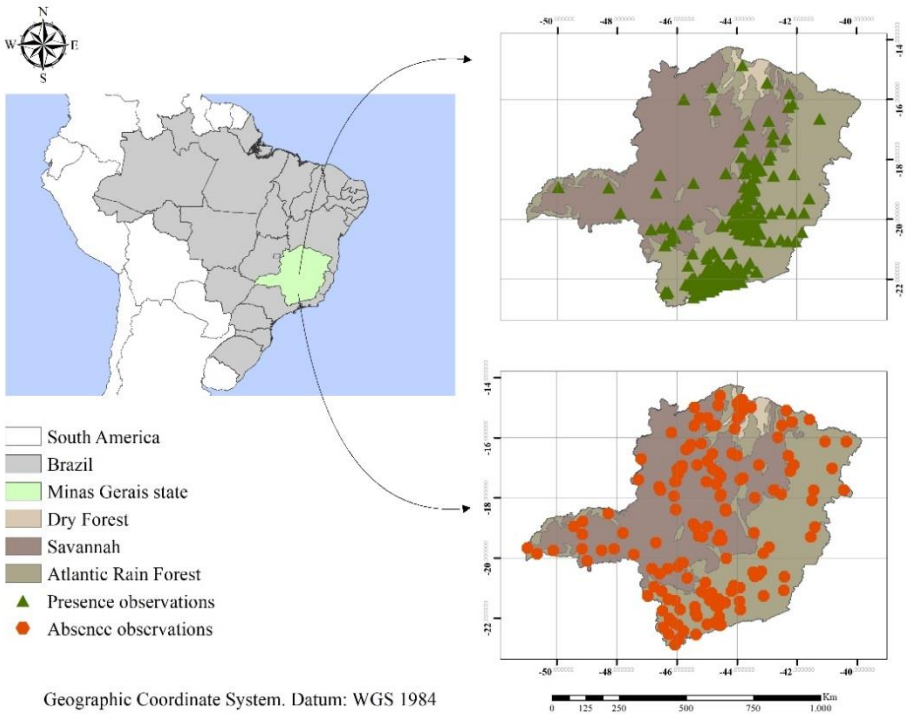


Figure 1. Location of study area and *Eremanthus erythropappus* presence and absence observations in the State of Minas Gerais, Brazil.

Species distribution modeling

We used an ecological niche modeling (ENM) approach to model the potential distribution of *E. erythropappus* under climate change scenarios. This approach has been widely used for tree species, either to verify the vulnerability and impacts of climate change in order to maintain the species sustainability (Pouteau *et al.*, 2012; Turchetto-Zolet *et al.*, 2016), and also as a tool for guiding recovery plans and indicating potential plantation areas (Coelho *et al.*, 2016; Carvalho *et al.*, 2017). The ecological niche models are based on Hutchinson's fundamental niche and its corresponding abiotic component known as the ecological niche (Sóberon & Peterson, 2005). The modeling process allows us to

use abiotic conditions observed in the known geographic distribution of our target species (realized niche) to predict potentially adequate environments where the species is theoretically viable. By projecting our model results onto a map, we can then visualize regions with potential distribution of the species at the present or in the future.

Several techniques are available for ENMs, among them the Random Forests algorithm (RF), which stands out for its performance and simple parametrization (Cluter *et al.*, 2007; Wang *et al.*, 2016). The algorithm is considered the best-performing statistical approach for mapping vegetation and for ecological niche model building (Garzon *et al.*, 2006; Lorena *et al.*, 2011; Wang *et al.*, 2012), and is being adopted by various countries for biodiversity modeling and for informing government politics on global warming (Wang *et al.*, 2016; Prasad *et al.*, 2016).

Occurrence data

Presence and absence observations for *E. erythropappus* in Minas Gerais state were obtained from several database providers. Presence data was obtained from 324 sustainable management plans of native *E. erythropappus* populations. These data were collected at environmental agencies within the state. To complement this dataset, we obtained an additional 372 and 294 observations respectively of presence data points from the online databases speciesLink (August 2017, http://www.splink.org.br) and Global Biodiversity Information Facility (GBIF) (June 2017, www.gbif.org). We also used field data from 197 fragments of native vegetation, where *E. erythropappus* were present in 36 and absent in 161 fragments. In total therefore, we compiled 1,187 points (1,026 presences observations and 161 absences). These points were rasterized at the spatial resolution of 0.008333 arc min (approximately 1 km), which was the same resolution we used to rasterize climatic variables. Many of these observations were within 1 km of each other, belonging to the same raster pixel,

and therefore we engaged in a process of data "cleaning". To accomplish this, each pixel was assigned a presence/absence status of *E. erythropappus*, resulting in 452 pixels with presence and 160 pixels with absence observation status (Figure 1). We then used this final 612 presence and absence observations to extract environmental variables for input into the subsequent modeling process.

Environmental data

The historical and future climate data were obtained from the WorldClim online database (Hijmans *et al.*, 2005) using the finest spatial resolution of 30 arc-seconds (~1 km) available. For modeling, we used seven climatic variables relating to temperature and rainfall and corresponding to the current period (1960 – 1990) (Carnaval & Moritz, 2008):BIO1 - Annual Mean Temperature; BIO4 - Temperature Seasonality; BIO10 - Mean Temperature of Warmest Quarter; BIO11 - Mean Temperature of Coldest Quarter; BIO12 - Annual Precipitation; BIO16 - Precipitation of Wettest Quarter; and BIO17 - Precipitation of Driest Quarter. According to Bucklin *et al.* (2015) climate predictors alone are an effective and efficient approach for initial assessments of environmental suitability, whereas additional predictors have relatively minor effects on the accuracy of predictions. Nevertheless, we incorporated an additional three topographical and soil features as predictive variables (Table 1), as these variables have been documented as environmental drivers of *E. erythropappus* distribution within the region (Pérez *et al.*, 2004; Scolforo *et al.*, 2012). Soil texture with original scale of 1:500,000 was categorized into three classes (fine, medium and coarse). Information on slope was obtained from the Shuttle Radar Topography Mission (SRTM) with 90m of resolution, and was classified into four classes (flat or gently sloping, sloping, moderately or strongly sloping, and steep or very steep). These variables were derived from the Ecological Economic Zoning of Minas Gerais (Curi *et al.*, 2008) and rasterized at 1 km spatial resolution. Altitude (meters above sea level) was also obtained

from WorldClim at the same spatial resolution (~1 km).

Forecast predictions were provided by two general circulation models (GCMs) with good predictive ability for South America, the Hadley Global Environment Model 2 - Earth System (HadGEM2-ES) and the Model for Interdisciplinary Research on Climate version 5 (MIROC5) (Watanabe *et al.*, 2010).

We adopted two representative concentration pathways (RCPs) from IPCC Fifth Assessment Report. The RCP4.5 assumes growth in the greenhouse gas emissions trajectory is limited through several initiatives, while RCP8.5 corresponds to a "worst-case scenario", with the highest estimated magnitude of greenhouse gas emissions, and where no climate specific mitigation targets are set (Mcintyre *et al.*, 2017). These scenarios were also obtained from the WorldClim using the spatial resolution of 30 arc-seconds (~1 km).We configured our models to generate predictions for the years 2050 (average for 2041-2060) and 2070 (average for 2061-2080). For each future period, there are four different scenarios of climatic conditions (2 RCPs x 2 GCMs), but for ease of interpretability and in order to incorporate the uncertainties related to projections of climate change (Wang *et al.*, 2012; Zhang *et al.*, 2015), we averaged the results of these four projections outputs at each period.

Habitat suitability modeling and evaluating

We used the Random Forests (RF) package (Liaw & Wiener, 2002) in the R platform (R version x64 3.4.1; R Core Team, 2017) to model the relationship between abiotic variables and current known locations of *E. erythropappus* occurrence. We then used future climate data in the years 2050 and 2070 to generate predictions of the future *E. erythropappus* habitat niche.

Usually, when the database has unbalanced data (e.g. more data points for presence than absence), the results may be overestimated for the frequency class. To overcome this problem, we use the methodology of Wang *et al.* (2016), and

employed an ensemble of 100 RF models, each of which was built with randomly sampled data points for presence, while the data points for absence remained the same. The final prediction was the most voted class or mean of probabilities of all RF models.

After tests, we set 500 decision trees in each forest and used the default, square-root of the number of predictors at each split. We assess the predictive RF accuracy by the percentage of correctly classified data (overall accuracy), and the probability of presence and absence being correctly predicted (sensitivity and specificity) through out-of-bag data (OOB). We also used the importance values generated by RF (based on Gini Index) to identify the abiotic factors that were important for determining the ecological niche of *E. erythropappus (*Menze *et al.*, 2009). Subsequent to training, we input the final RF models with the environmental data of the study area to generate predictions for the current period and projections for future periods (2050 and 2070).

Our ENM predicted areas currently occupied by the species and also areas which are environmentally suitable for the species but with no known records of occupancy due to various factors including physical barriers, seed dispersal, migration, and/or human interference (Wang *et al.*, 2012). This should be interpreted as the quality of the site considering the preferences of the species, and we interpreted here as habitat suitability (HS). Because the classification of probability in binary data comparisons of range area changes among different scenarios require a threshold to classify probabilities in presence and absence, we set three thresholds for classifying the habitat suitability of *E. erythropappus*, considering the species as at a given grid if the habitat suitability (HS) is above or equal: 0.3 – broadly distributed; 0.5 – moderately distributed; and 0.7 – restricted distribution.

To track changes in latitudinal distributions, we compared geographic centroids of predictions and projections species range (Zhang *et al.*, 2015). The

geographic centroid weighted by probability was calculated using the mean center function in ArcGIS10.1 (ESRI Inc., http://www.esri.com/). This function calculates the average coordinates weighted by probability in the study area through the formulas (1) and (2), where x_i and y_i are the central coordinates for grid i, n is equal to the total number of grids, and w_i is the probability at the grid i.

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (1)$$

$$\overline{Y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad (2)$$

Finally, we used the predicted scenarios to map the areas of *E. erythropappus* with a habitat suitability >0.5 to account for areas with the presence of the species under full protection in conservation units within the Minas Gerais state. In total, there are 103 conservation units with full protection in the state, according to data updated in May 2017 by the State Forestry Institute (IEF) (http://www.ief.mg.gov.br/areas-protegidas/banco-de-dados-de-unidades-de-conservacao-estaduais).

Results

Assessment of ecological niche and important abiotic variables

Our environmental niche models for *Eremanthus erythropappus* presented reliable results, with an overall error rate average of 17.2%, and presence/absence error rate of 17.5% and 17.2%, respectively. As expected, the OOB error rate was similar for presence and absence, after the sampling rate was balanced for both classes through bootstrapping. Amongst the ten environmental input variables, precipitation of Wettest Quarter (BIO16), mean temperature of

coldest quarter (BIO11), annual mean temperature (BIO1), and altitude were the four most influential factors according to RF importance values (decrease impurity values of around 20%; Table 1). In contrast, the categorical variables of slope and texture were the least influential environmental variables (Table 1).

Table 1. Mean variables importance values for environmental variables generated from the Random Forests models based on Gini Index (or the Mean decrease impurity). The standard deviation of the Gini Index from 100 replicates are given in parentheses.

Variables	Mean decrease impurity
Precipitation of Wettest Quarter (BIO16)	26.32 (5.26)
Mean Temperature of Coldest Quarter (BIO11)	24.43 (3.42)
Annual Mean Temperature (BIO1)	23.76 (3.13)
Altitude	22.31 (4.11)
Mean Temperature of Warmest Quarter (BIO10)	18.38 (2.46)
Temperature Seasonality (BIO4)	14.36 (1.64)
Precipitation of Driest Quarter (BIO17)	12.61 (1.16)
Annual Precipitation (BIO12)	12.28 (1.11)
Slope	3.62 (0.85)
Soil texture	1.41 (0.39)

Current potential suitable habitat for *Eremanthus erythropappus*

E. erythropappus is currently distributed in the central-south portion of the state, with a particular concentration in mountainous environments (Figure 2). The habitat suitability thresholds revealed that the most suitable areas for the occurrence of *E. erythropappus* is the Espinhaço, Mantiqueira and Canastra mountain ranges, which we regard as core areas for the species . The area

covered by habitat suitability greater than 0.3, 0.5 and 0.7 corresponded respectively to 158,569 km², 61,467 km² and 24,188 km² (Table 2).

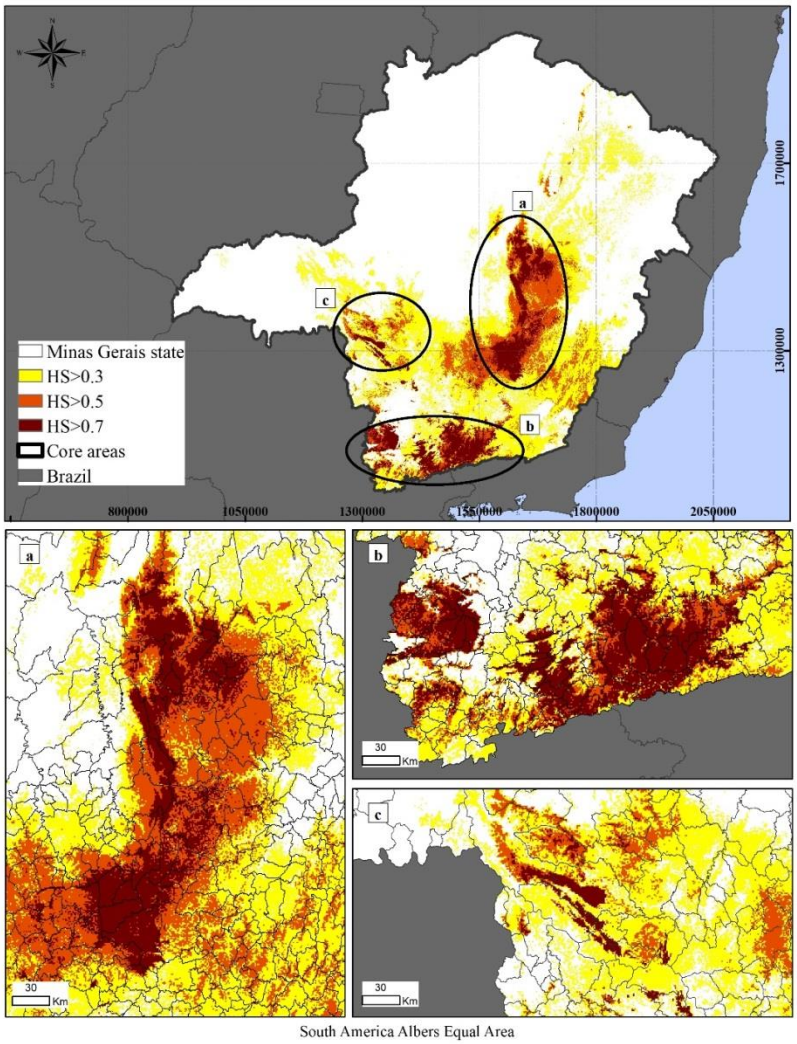


Figure 2. Predicted ecological niche of *Eremanthus erythropappus* trees in Minas Gerais, Brazil, at different habitat suitability (HS) thresholds in core areas of occurrence within the (a) Espinhaço, (b) Mantiqueira, and (c) Canastra mountain ranges.

Table 2. Area (km²) of current and future suitable habitat of *Eremanthus erythropappus* under three different habitat suitability (HS) thresholds (≥0.3 – broad distribution; ≥0.5 – moderate distribution; and ≥0.7 – restricted distribution). Where "Loss" represents the contraction area of the ecological niche of the species in the future scenario in relation to the current scenario; "Gain" represents the ecological niche expansion area in the future scenario in relation to the current scenario, and "Stable" represents the area covered by the ecological niche of the species both in the current scenario and in the future scenario. Total area (%) was calculated through the ratio of the area predicted in the future under the area predicted in the present multiplied by 100.

Period	HS	Current (1960-1990)	Loss	Stable	Gain	Total area	
		(km²)				(km²)	(%)
2050 (2041-2060)	0.3	158,569	- 64,082	94,486	+ 28,668	123,154	77.7
	0.5	61,466	- 50,595	10,871	+ 178	11,050	18.0
	0.7	24,187	- 22,781	1,406	+ 91	1,497	6.2
2070 (2061-2080)	0.3	158,569	- 47,030	111,538	+ 81,069	192,608	121.5
	0.5	61,466	- 55,693	5,773	+ 93	5,866	9.5
	0.7	24,187	- 24,175	12	0	12	0

Ecological niche of *Eremanthus erythropappus* under climate change scenarios

We found varying degrees of net losses and gains in ecological niche for *E. erythropappus* under different modelled habitat suitability (HS) thresholds. The loss in suitable habitat is particularly drastic in 2050 and 2070 under habitat suitability (HS) thresholds of 0.5 and 0.7 (Figure 3b, c, e, f; Table 2). Under HS ≥0.5, suitable habitat for *E. erythropappus* in 2070 is found only in core regions, such as Espinhaço and Mantiqueira (Figure 3b, e), and has vanished in Canastra mountain range. On the other hand, presence probabilities for *E. erythropappus*

under the ≥ 0.7 threshold shown serious decline in the state, covering only a small area in the Espinhaço and Mantiqueira mountain ranges in 2050 (Figure 3c) and practically extinguished in 2070 (Figure 3f).

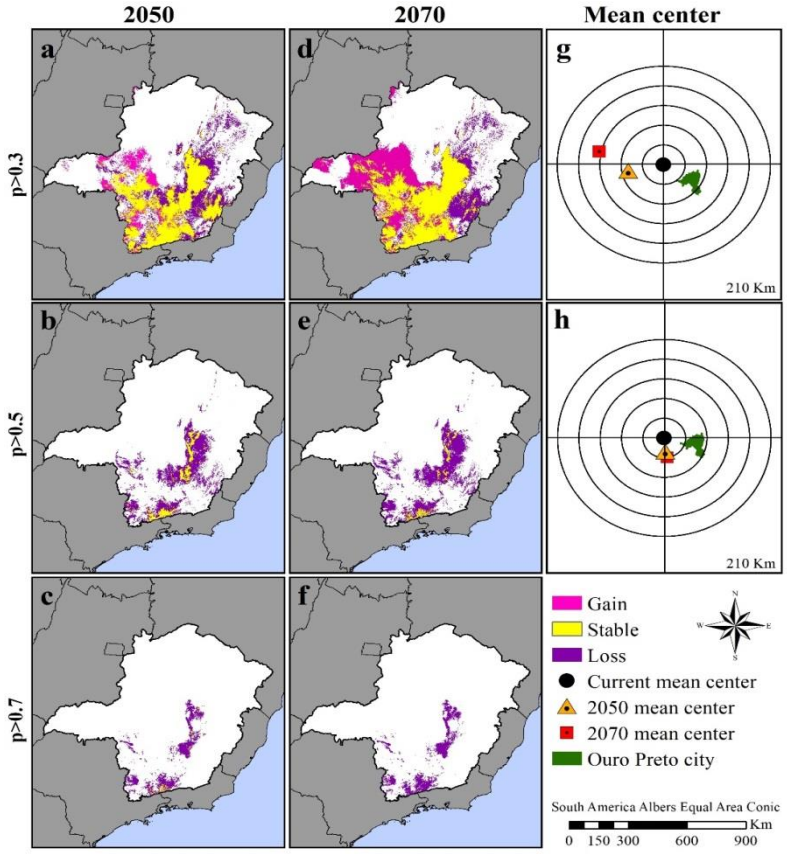


Figure 3. Predicted areas of occurrence of *Eremanthus erythropappus* in the years 2050 and 2070 and under three different habitat suitability thresholds (HS≥0.3, 0.5 and 0.7), and the gains and losses in areas in relation to predicted current distribution (a-f). Figures g and h illustrate the mean centers for current, 2050 and 2070 under the ≥0.3 and 0.5 thresholds respectively. The mean center for 2070 under HS≥ 0.7 is not shown due to the scarcity of remaining grids under that projection.

The substantial gain in areas of suitable habitat in 2050 and 2070 in relation to the current conditions appeared only when we adopted the HS ≥ 0.3 threshold. The average area for HS ≥ 0.3 was 158,569.00 km² (2018), decreasing to 123,155.00 (2050) and increasing to 192,608 km² in 2070 (Figure 3a, d). We also detected potential changes in distance and direction of *E. erythropappus* occurrence mean centers (Figure 3g and h) where all mean centers were located in the Espinhaço mountain range. Under the HS ≥ 0.3 threshold, there was a shift in suitable habitat towards to the west region, but under HS ≥ 0.5, the shift was towards the south region. We did not calculate the mean center for 2070 under HS≥ 0.7 due to the scarcity of remaining grids under this scenario.

Table 3. The predicted extent of suitable habitat area for *Eremanthus erythropappus* within conservation units (with total protection) in Minas Gerais, Brazil at a threshold HS ≥ 0.5. The classes of loss in area (%) were estimated based on the contraction area of the ecological niche of the species within each conservation unit in relation to the total area of the conservation unit, and represents de number of conservation units for each class of loss. The "Gain" corresponds the number of conservation units which presented expansion area of suitable habitat for the species.

Period	Total protected area (km²)	Number of conservation units					
		Current **(1960-1990)**	Classes of loss in area (%)				Gain
			0 – 25	25 - 50	50 - 75	75 - 100	
1960 - 1990	4,087.37	58	-	-	-	-	-
2041 - 2060	1,811.01	41	22	5	6	21	4
2061-2080	1,049.96	31	9	12	5	31	1

Minas Gerais state has 103 conservation units which encompass a wide range of vegetation structural types. However, only half of them (58 units) are suitable *E. erythropappus* habitat (HS ≥ 0.5 threshold). These conservation units encompass a land area of 4,087.37 km^{2}of protected suitable habitat for the species (Figure 4). We observed a 74.3% reduction in suitable habitat area for the species within these conservation units in the year 2070, with only 1,049.96 km² (Table 3) of suitable habitat remaining.

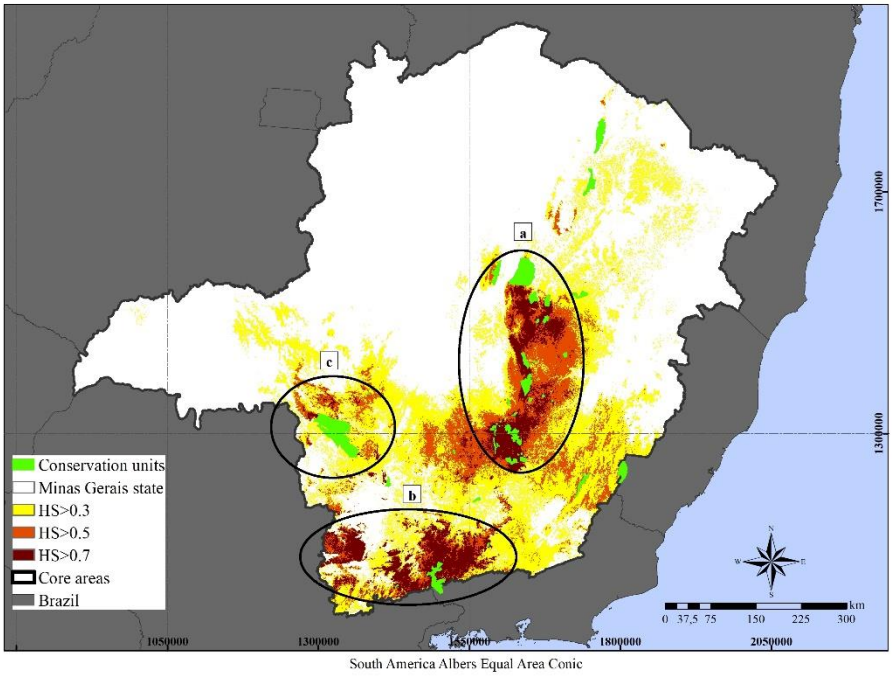


Figure 4. Predicted current suitable habitat of *Eremanthus erythropappus* trees in Minas Gerais, Brazil, under three different habitat suitability thresholds and within fully protected areas (conservation units) covered by the ecological niche of the species. The ellipses indicate the (a) Espinhaço, (b) Mantiqueira, and (c) Canastra mountain ranges.

Discussion

The bioclimatic variables BIO16 (precipitation of wettest quarter), BIO11 (the temperature of coldest quarter), BIO1 (Annual Mean Temperature) have the strongest bearing on our ecological niche models. According to Condit *et al.*, (2000) and Brenes-Arguedas *et al.*, (2011) this rainy period (BIO16) is crucial for plant photosynthesis, growth and resource allocation to reproduction, which affects the formation and quality of *E. erythropappus* seeds (Feitosa *et al.*, 2009). The mean temperatures, annual and the coldest quarter, indicates that *E. erythropappus* favors mild temperatures but does not tolerate ground frost in the winter (Magalhães *et al*., 2008, Siqueira & Pinto, 2009).

Altitude is the fourth most important variable and is of great importance for *E. erythropappus* trees in our model. Consistently, the upland areas have strong probabilities of *E. erythropappus* occurrence, in agreement with previous work (Pérez *et al.*, 2004; Silva *et al.*, 2008; Clark *et al.*, 2011; Scolforo *et al.*, 2012). It is worth noting that lowland or valley sites usually are not dominated by the tree species, where inter-species competition is likely the main reason for its exclusion. However, *E. erythropappus* occasionally may colonize these fields after deforestation (Scolforo *et al.*, 2012; Pádua *et al.*, 2016) and bushfires during the warmer months of the year (Oliveira Filho & Fluminhan Filho, 1999; Araújo *et al*., 2017).

The current and future suitable habitat of *E. erythropappus* is crucial for informing policies on the next steps in the conservation and sustainable management of the species. Eventually, these concerns will need to be addressed if climate change starts to affect the productivity of the species in certain areas, resulting in reduced population or lowered productivity. We therefore identified three core areas with high potential for the conservation of the species, namely areas within Mantiqueira, Espinhaço and Canastra mountain ranges. It is also noteworthy that these mountain complexes represent some of the most ancient

open vegetation areas in eastern South America, and are of high conservation priority (Silveira *et al.*, 2016). In fact, it is recognized that upland areas are less affected by climate changes and can still guarantee the species habitat. According to Gwitira *et al.* (2014), high-altitude areas will remain as biodiversity hotspots of savannah domain. Forest loss in warmer regions tends to accelerate the movement of species upslope, especially in the tropics where both climate and habitat loss stressors, will increase the risk of net lowland biotic attrition (Guo *et al.*, 2018).

Based on the most optimistic scenario of our models (HS≥ 0.3), these three core areas could maintain areas that will remain connected as a huge network. However, the pessimistic scenarios (HS≥ 0.5) for 2050 and 2070 show limited gains in suitable area, and the loss of continuous distribution of the species between the three cores mountain areas. Most worryingly, we detected a huge reduction in suitable habitat area the species in 2070 under the most pessimistic scenario (HS≥ 0.7), where there will practically be no suitable habitats for the species in Minas Gerais.

In contrast to losses in suitable area, we did detect some gains in habitat suitability in the west of the state, in the Canastra mountain. However, this happened only when we adopted a HS ≥ 0.3. These gains may be attributed to increasing precipitation of wettest quarter (BIO16) in this region, which is forecasted in climate scenarios.

Currently, conservation units within Minas Gerais state cover a considerable area of the ecological niche of *E. erythropappus*. Several of our future predictions have shown how deep the negative impacts on nature reserves are. The losses in suitable area and connections may lead to reduced genetic diversity in the species, thereby amplifying the risk of the species becoming extinct. Genetic studies indicate that even though *E. erythropappus* possess high genetic variability, trees typically have significant co-ancestry, where

individuals located in close proximity to each other are genetically similar (Pádua *et al.*, 2016). For this reason, conservation and restoration strategies efforts should be done at protection units around the three core distribution areas.

It is therefore timely to consider strategies for exploitation use of *E. erythropappus* that will be socially acceptable, economically viable, and ecologically sound. Since climate changes will impact negatively natural populations of the species, silviculture practices would likely serve as be an economically viable and ecological sustainable alternative to harvesting natural populations.

Several studies have already been published regarding the sustainability of plantations for oil extraction from *E. erythropappus* (Feitosa *et al.*, 2009; Silva *et al.*, 2014; Scolforo H *et al.*, 2016; Scolforo J *et al.*, 2016). Practically, plantations of *E. erythropappus* in the coming decades may be most feasible in the western parts of the Canastra mountain ranges, where we predicted some potential gains in suitable habitat for the species.

Additionally, *E. erythropappus* trees are easy to be propagated even in many soil types and conditions due to its ecological preferences (Amaral *et al.*, 2015; Meira Júnior *et al.*, 2015; Pereira *et al.*, 2015). We recommend therefore that management initiatives spearheaded by the Brazilian government should encourage the land owner to cultivate this species for commercial use or forest restoration areas. At present, Minas Gerais has 25,000 km² of rural areas and 4,000 km² of legal reserves under the Brazilian Forestry Service (SFB), which could benefit from reforestation programs incorporating this tree species (February 2018, http://www.florestal.gov.br/modulo-de-relatorios).

In conclusion, *Eremanthus erythropappus* is an ecologically and economically important tree species in Brazil, but our projections of how climate change could affect it is of concern for the sustainability of future natural

populations of the species in the state of Minas Gerais. We suggest that it may be prudent to engage in activities that will ensure sustainable harvesting of *E. erythropappus* in Minas Gerais state, and as an insurance against the potential negative impacts of climate change on the species.

References

Amaral C, Amaral W, Pereira I, Oliveira P, Machado V. Comparação florístico-estrutural dos estratos adultos e regenerantes em área minerada de campo rupestre, Diamantina, MG. Cerne 2015; 21(2): 183 – 190. http://dx.doi.org/10.1590/01047760201521021405.

Araújo F, Tng D, Apgaua D, Coelho P, Pereira D, Santos R. Post-fire plant regeneration across a closed forest-savanna vegetation transition. Forest Ecol Manag 2017; 400(15): 77 – 84. https://doi.org/10.1016/j.foreco.2017.05.058

Breiman L. Random Forests. Machine Learning 2001; 45(1): 5–32.

Brenes-Arguedas T, Roddy A, Coley P, Kursar T. Do differences in understory light contribute to species distributions along a tropical rainfall gradient? Oecologia 2011; 166(2): 443 – 456. http://dx.doi.org/10.1007/s00442-010-1832-9

Bucklin D, Basille M, Benscoter A, Brandt L, Mazzotti F, Romañach S, Speroterra C, Watling J. Comparing species distribution models constructed with different subsets of environmental predictors. Divers Distrib 2014; 21(1): 1 – 13. https://doi.org/10.1111/ddi.12247

Carnaval A C & Moritz C. Historical climate modeling predicts patterns of current biodiversity in the Brazilian Atlantic forest. J Biogeogr 2008; 35(7): 1187 – 1201. https://doi.org/10.1111/j.1365-2699.2007.01870.x

Carvalho M, Gomide L, Santos R, Scolforo JR, Carvalho L, Mello J. Modeling ecological niche of tree species in Brazilian tropical area. Cerne 2017; 23 (2): 229 – 240. http://dx.doi.org/10.1590/01047760201723022308

Clark A, Khweiss N, Salazar L, Verdadero L. Promoting Sustainability in the Value Chain of Natural Bisabolol, a Brazilian Rainforest Product. Columbia University, New York; 2011.

Cluter D, Edwards T, Beard K, Cluter A, Hess K, Gibson J. Random Forest for classification in ecology. Ecology 2007; 88(11): 2783 – 2792. https://doi.org/10.1890/07-0539.1

Coelho G, Tavares L, Gomide L. Modelagem preditiva de distribuição de espécies pioneiras no Estado de Minas Gerais. Pesqui Agropecu Bras 2016; 51(3): 207 – 214. http://dx.doi.org/10.1590/S0100-204X2016000300002.

Condit R, Ashton P, Baker P, Bunyavejchewin S, Gunatilleke S, Gunatilleke N, Hubbel S, Foster R, Itoh A, Lafrankie J *et al.* Spatial Patterns in the Distribution of Tropical Tree Species. Science 2000; 288 (5470): 1414 – 1418. http://dx.doi.org/10.1126/science.288.5470.1414

Curi N, Marques J, Marques A, Fernandes E. Solos, geologia, relevo e mineração. In: Zoneamento ecológico-econômico do estado de Minas Gerais: Componentes geofísico e biótico. Scolforo J, Carvalho L, Oliveira A (eds).pp: 73 – 88. Editora UFLA, Lavras; 2008.

Donadelli F. Motivações e resultados da certificação florestal: um estudo de caso cadeia de valor da Candeia. Ambiente & Sociedade 2012; 15 (3): 97 – 121. http://dx.doi.org/10.1590/S1414-753X2012000300007

Feitosa S, Davide A, Tonetti O, Fabricante J, Lui L. Estudos de viabilidade de sementes de Candeia *Eremanthus erythropappus* (DC.) MacLeish por meio de testes de germinação e raios X. Floresta 2009; 39 (2): 393 – 399. http://dx.doi.org/10.5380/rf.v39i2.14565

Garzón M, Blazek R, Neteler M, Dios R, Ollero H, Furlanello C. Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecol Model 2006; 197(3-4): 383 – 393. https://doi.org/10.1016/j.ecolmodel.2006.03.015

GFA Consulting Group. Análise de Mercado e da Cadeia Produtiva do Óleo de Candeia e do Alfabisabolol. Alemanha, Final Report, 72 pp; 2006.

Guo F, Lenoir J, Bonebrake T. Land-use change interacts with climate to determine elevational species redistribution. Nat Commun 2018; 9(1315). https://doi.org/10.1038/s41467-018-03786-9

Gwitira I, Murwira A, Shekede M, Masocha M, Chapano C. Precipitation of the warmest quarter and temperature of the warmest month are key to understanding the effect of climate change on plant species diversity in Southern African Savanna. Afr J Ecol 2014; 52 (2): 209 – 216. https://doi.org/10.1111/aje.12105

Hamann A, Wang T. Potential effects of climate change on ecosystem and tree species distribution in British Columbia. Ecology 2006; 87 (11): 2773–2786. https://doi.org/10.1890/0012-9658(2006)87[2773:PEOCCO]2.0.CO;2

Hijmans R, Cameron S, Parra J, Jones P, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 2005; 25 (15): 1965 – 1978. https://doi.org/10.1002/joc.1276

Kurz W, Dymond C, Stinson G, Rampley G, Neilson E, Carroll A, Ebata T, Safranyik L. Mountain pine beetle and forest carbon feedback to climate change. Nature 2008; 452 (24): 987– 990. https://doi.org/10.1038/nature06777

Liaw A, Wiener M. Classification and regression by Random Forest. R News 2012; 2 (3): 18 – 22.

Lorena A, Jacintho L, Siqueira M, De Giovanni R, Lohmann G, Carvalho A, Yamamoto M. Comparing machine learning classifiers in potential distribution modeling. Expert Syst Appl 2011; 38 (5): 5268 – 5275. https://doi.org/10.1016/j.eswa.2010.10.031

Magalhães C, Missagia R, Costa F, Costa M. Diversidade de fungos endofíticos em Candeia *Eremanthus erythropappus* (DC.) MacLeish. Cerne 2008; 14(3): 267 – 273.

Martins F, Gonzaga G, Santos D, Reboita M. Classificação Climática de Köppen

e de Thornthwaite para Minas Gerais: Cenário atual e projeções futuras. Revista Brasileira de Climatologia 2018; Edição Especial: Dossiê Climatologia de Minas Gerais – ano 14.

Mcintyre S, Rangel E, Ready P, Carvalho B. Species-specific Ecological niche modeling predicts different range contractions for *Lutzomyia intermedia* and a related vector of *Leishmania braziliensis* following climate change in South America. Parasit Vectors 2017; 10 (1): 10 – 157. https://doi.org/10.1186/s13071-017-2093-9

Meira Júnior M, Pereira I, Machado E, Mota S, Otoni T. Espécies potenciais para recuperação de áreas de floresta estacional semidecidual com exploração de minério de ferro na serra do Espinhaço. Biosci J 2015; 31(1): 283 – 295. http://dx.doi.org/10.14393/BJ-v31n1a2015-23414

Menze B, Kelm B, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht, 2009. A comparison of Random Forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMB Bioinformatics 10 (213). https://doi.org/10.1186/1471-2105-10-213

Mori C, Brito J, Scolforo J, Vidal E, Mendes L. Influence of altitude, age and diameter on yeld and Alpha-Bisabolol content of Candeia trees (*Eremanthus erythropappus*). Cerne 2009; 15 (3): 339 – 345.

Oliveira-Filho A, Fluminhan-Filho. Ecologia da vegetação do parque florestal quedas do Rio Bonito. Cerne 1999; 5(2): 51 – 64.

Pádua J, Brandão M, Carvalho D. Spatial genetic structure in natural populations of the overexploited tree *Eremanthus erythropappus* (DC.) Macleish (Asteraceae). Biochem Syst Ecol 2016; 66 (1): 307 – 311. https://doi.org/10.1016/j.bse.2016.04.015

Pereira I, Santos G, Carlos L, Silva N. Plantio de Candeia e uso de topsoil na recuperação de uma cascalheira no Parque Estadual do Biribiri em Diamantina,

MG. MG BIOTA 2015; 8(1): 22 – 53.

Pérez J, Scolforo J, Oliveira A, Mello J, Borges L, Camolesi JF. Sistema de manejo para a Candeia – *Eremanthus erythropappus* (dc.) Macleish – a opção do sistema de corte seletivo. Cerne 2004; 10 (2): 257 – 273.

Pouteau R, Meyer J, Taputuarai R, Stoll B. Support Vector machines to map rare and endangered native plants in Pacific islands forests. Ecol Inform 2012; 9 (1): 37 – 46. https://doi.org/10.1016/j.ecoinf.2012.03.003

Prasad A, Iverson L, Matthews S, Peters M. A multistage decision support framework to guide tree species management under climate change via habitat suitability and colonization models, and a knowledge-based scoring system. Landscape Ecol 2016; 31 (9): 2187 – 2204. https://doi.org/10.1007/s10980-016-0369-7

Ribeiro J, Fonseca C, Carvalho F. The woody vegetation of quartzite soils in a mountain landscape in the Atlantic forest domain (south-eastern Brazil): structure, diversity and implications for conservation. Edinburgh J Bot 2017; 74(1): 15–32. https://doi.org/10.1017/S096042861600024X

Rodrigues P, Eisenlohr P, Schaefer C. Climate change effects on the geographic distribution of specialist trees of the Brazilian tropical dry forests. Braz J Biol 2015; 75(3): 679 – 684. http://dx.doi.org/10.1590/1519-6984.20913.

Scolforo H, Scolforo J, Mello J, Ferraz Filho A, Rossoni D, Altoé T, Oliveira A, Lima R. Autoregressive spatial analysis and individual tree modeling as strategies for the management of *Eremanthus erythropappus*. J Forestry Res 2016; 27(3): 595 – 603. https://doi.org/10.1007/s11676-015-0185-y

Scolforo J, Altoé T, Scolforo H, Mello J, Silva C, Ferraz-Filho A. Management strategies of *Eremanthus erythropappus* (DC.) MacLeish under different initial spacing. Cienc Agrotec 2016; 40(3): 298 – 304. http://dx.doi.org/10.1590/1413-70542016403042715

Scolforo J, Oliveira A, Davide A. Manejo Sustentável da Candeia: o caminhar

de uma nova experiência florestal em Minas Gerais. Lavras: Editora UFLA, Lavras, BR, 329 pp; 2012.

Silva C, Oliveira A, Coelho Junior L, Scolforo J, Souza A. Viabilidade econômica e rotação florestal de plantios de Candeia (*Eremanthus erythropappus*) em condição de risco. Cerne 2014; 20(1): 113 – 122.

Silva M, Mello J, Scolforo J,Czanck L, Andrade I, Oliveira A. Análise da distribuição espacial da Candeia (*Eremanthus erythropappus* (DC.) MacLeish) sujeita ao sistema de manejo porta-sementes. Cerne 2008; 14(4): 311 – 316.

Silveira F, Negreiros D, Barbosa N, Lambers H. Ecology and evolution of plant diversity in the endangered campo rupestre: a neglected conservation priority. Plant Soil 2016; 403 (1-2): 129 – 152. http://dx.doi.org/10.1007/s11104-015-2637-8

Siqueira F, Pinto L. Semeadura direta de Candeia (*Eremanthus erythropappus*) sob diferentes adubações em Inconfidentes – MG. Agrogeoambiental 2009; 1(4): 64 – 69.

Sóberon J, Peterson, A T. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2005; 2: 1-10. https://doi.org/10.17161/bi.v2i0.4

Turchetto-Zolet A, Salgueiro F, Turchetto C, Cruz F, Veto N, Barros M, Segatto A, Freitas L, Margis R. Phylogeography and ecological niche modeling in *Eugenia uniflora* (Myrtaceae) suggest distinct vegetational responses to climate change between the southern and the northern Atlantic Forest. Bot J Linn Soc 2016; 182(3): 670 – 688. https://doi.org/10.1111/boj.12473

Vieira F, Fajardo C, Carvalho D. Floral biology of Candeia (*Eremanthus erythropappus*, Asteraceae). PFB 2012; 32(72): 477 – 481. https://doi.org/10.4336/2012.pfb.32.72.477

Wang T, Campbell E, O'Neill G, Aitken S. Projecting future distributions of ecosystem climate niches: uncertainties and management applications. Forest

Ecol Manag 2012; 279 (1): 128 – 140. https://doi.org/10.1016/j.foreco.2012.05.034

Wang T, Wang G, Innes J, Nitschke C, Kang H. Climatic niche models and their consensus projections for future climates for four major forest tree species in the Asia–Pacific region. Forest Ecol Manag 2016; 360 (15): 357 – 366. https://doi.org/10.1016/j.foreco.2015.08.004

Watanabe M, Suzuki T, O'ishi R, Komuro Y, Watanabe S, Emori S, Takemura T, Chikira M, Ogura T, Sekiguchi M *et al.*. Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. J Climate 2010; 23 (23): 6312 – 6335. https://doi.org/10.1175/2010JCLI3679.1

Zhang L, Shirong L, Sun P, Wang T, Wang G, Zhang X, Wang L. Consensus Forecasting of Species Distributions: The Effects of Niche Model Performance and Niche Properties. PLoS ONE 2015; 10(3). https://doi.org/10.1371/journal.pone.0120056

ARTIGO 3 - MODELAGEM DO ESTOQUE DE CARBONO CONTIDO NA VEGETAÇÃO: UMA ABORDAGEM ENVOLVENDO MINERAÇÃO DE DADOS

Modeling the spatial distribution of aboveground carbon stock using data mining.

Mônica Canaan Carvalho; Lucas Rezende Gomide; Eduarda Martiniano Silveira; Rafael Menali; Fausto Weimar Acerbi-Júnior; José Márcio de Mello; José Roberto Soares Scolforo.

Artigo redigido conforme a NBR 6022 (ABNT, 2003) e formatado de acordo com o Manual da UFLA de apresentação de teses e dissertações.

RESUMO

Estimativas confiáveis da taxa de desmatamento e do estoque de carbono da vegetação são essenciais para predizer a quantidade de carbono, emitida ou sequestrada, no tempo e no espaço, auxiliando no desenvolvimento de políticas de mitigação das mudanças climáticas. O objetivo deste artigo foi modelar o estoque de carbono da parte aérea da vegetação nativa presente na bacia do rio Grande - MG, indentificando as variáveis com potencial para tal objetivo. Com este intuito, utilizou-se técnicas de aprendizagem de máquina e grande conjunto inicial de dados, representando características climáticas, morfométricas, edáficas, espectrais e geográficas. Testou-se o algoritmo Random Forest (RF) com diferentes metodologias para seleção das variáveis principais: remoção recursiva e algoritmo genético, uni e multiobjetivo. A média do teor de carbono encontrado na bacia foi de 47,29 $Mg.ha^{-1}$, variando em uma faixa de 3,84 a 214,96 $Mg.ha^{-1}$, com desvio padrão de 24,65 $Mg.ha^{-1}$. Considerando o número de variáveis selecionadas e o menor erro, o melhor método para a seleção e modelagem do estoque de carbono foi a metodologia híbrida que combina algoritmo genético multiobjetivo e RF. Tal metodologia obteve raiz quadrada do erro médio percentual de 35,56% e selecionou apenas 4 variáveis dentro do conjunto inicial de 114. As variáveis selecionadas pela metodologia são do tipo espectral, dos satélites Landsat 8 oli e MODIS. São elas em ordem crescente de importância: latent heat flux, textura NBR2 correlação, índice de vegetação NDMI e Treecover. Diante dos resultados obtidos nesta pesquisa, indica-se o uso do algoritmo genético na seleção de variáveis para a modelagem do estoque de carbono acima do solo utilizando o algoritmo RF.

Palavras-chave: Algoritmo genético. Random Forest. Florestas nativas.

1 INTRODUÇÃO

O desmatamento e a degradação das florestas promovem a emissão de carbono para a atmosfera, o que reflete diretamente no aquecimento do planeta (FROLKING et al., 2009; HANSEN et al., 2013) e perdas de biodiversidade (BELLARD et al., 2012; PECL et al., 2017). Aproximadamente 25% do carbono liberado para a atmosfera derivam do desmatamento em todo o globo (PAIVA; FARIA, 2007). No Brasil, o desmatamento representa a principal fonte de emissões, o que coloca o país em 5° lugar no ranking dos países que mais liberam quantidades equivalentes de Dióxido de Carbono (CO_2) para a atmosfera (SILVEIRA et al., 2019a).

Estimativas confiáveis da taxa de desmatamento e do estoque de carbono da vegetação são essenciais para predizer a quantidade de carbono, emitida ou sequestrada, no tempo e no espaço. Ao precisar a taxa de desmatamento e a quantidade de carbono estocadas nas florestas, é possível gerar informações úteis no que tange às políticas de mitigação das mudanças climáticas (HIGUCHI et al., 2004; SCOLFORO et al., 2015). Diante dessa importância, os esforços científicos para a quantificação do estoque de carbono sequestrado pela vegetação são observados em Baccini et al. (2008), Lu et al. (2004), Pandit, Tsuyuki e Dube (2018), Scolforo et al. (2015), Seidel et al. (2011), Silveira et al. (2019a, 2019b) e Wang et al. (2011).

Entretanto os primeiros estudos com essa temática caracterizavam o estoque de carbono das florestas de forma pontual e com o emprego de técnicas destrutivas (HIGUCHI et al., 2004; MORAIS et al., 2013). Com o avanço da disponibilidade de dados de diversos satélites, essa barreira espacial pôde enfim ser ultrapassada (LU et al., 2004). As técnicas de sensoriamento remoto têm sido amplamente aplicadas para estimativas de biomassa e estoque de carbono em grandes escalas (PONZONI et al., 2015; WERE et al., 2015; WU et al., 2016),

devido a suas vantagens como grande quantidade de informações e disponibilidade espacial e temporal (CHEN, 2013; LU et al., 2012). O princípio da estimativa do teor de carbono através do sensoriamento remoto, baseia-se na reflectância do dossel da vegetação, que está diretamente relacionada com a biomassa aérea (*aboveground biomass* - AGB) e consequentemente, com o estoque de carbono acima do solo (LU et al., 2014).

Além dos valores de reflectância obtidos nas diferentes bandas das imagens de satélite e índices de vegetação, exite ainda a possibilidade de incorporar variáveis em escala regional/global, como clima, relevo, geografia, etc (BOLIVAR; GUTIERREZ-VELEZ; SIERRA, 2018; DUBE; MUTANGA, 2016; LU et al., 2017; REIS et al., 2018).

Trabalhos envolvendo macroescalas apresentam desafios compatíveis com seu tamanho. O primeiro deles é a confiabilidade dos dados de campo, o segundo é a identificação de padrões e correlações entre variáveis para, então, explicar a variável desejada em estudo. O uso de imagens de satélites e dados ambientais nesses trabalhos é uma prática recorrente, conforme observado em Gao et al. (2018), Reis et al. (2018) e Silveira et al. (2019b). Contudo, diante da grande disponibilidade de dados e de técnicas de modelagem, a pergunta que surge é qual o método e variáveis a serem utilizados para tal finalidade.

Uma técnica comumente utilizada nesses estudos é o algoritmo *Random Forest* (RF), com emprego de seleção das variáveis baseado no método de remoção recursiva em ordem crescente de importância no modelo (DIAZ-URIARTE, 2007; MILLARD; MURRAY, 2015). O algoritmo RF, desde sua concepção em 2001, tem sido empregado com sucesso em estudos de diversas áreas, principalmente na área de sensoriamento remoto (AHMED et al., 2015; LU et al., 2014; WU et al., 2016). Além do ótimo desempenho, o algoritmo apresenta validação interna e medida de contribuição de cada variável no modelo, facilitando seu entendimento.

Entretanto, outras abordagens para a seleção de variáveis têm sido pouco empregadas quando se trata da aplicação do RF. Algoritmos genéticos constituem uma dessas abordagens, e são comumente empregados para busca de soluções ótimas para problemas combinatórios, constituindo uma técnica consolidada para seleção de variáveis (KUMAR; SAHOO, 2017; LAFITI; NOTHDURF; KOCH, 2010). Ainda assim, são excassos os estudos que avaliam outras técnicas para seleção de variáveis, como os algoritmos genéticos, utilizando grande quantidade inicial de variáveis.

O trabalho buscou responder duas perguntas básicas, a primeira diz respeito à identificação das principais variáveis que explicam o comportamento do carbono na parte aérea da floresta.. A segunda está associada aos métodos de predição aplicados nas estimativas de carbono, considerando opções híbridas entre algoritmos e tipos de funções uni e multiobjetivas, com o propósito de aumento da acurácia. Assim, o objetivo geral foi modelar o estoque de carbono na parte aérea da floresta em macroescala, utilizando algoritmos de aprendizado de máquina.

2 MATERIAL E MÉTODOS

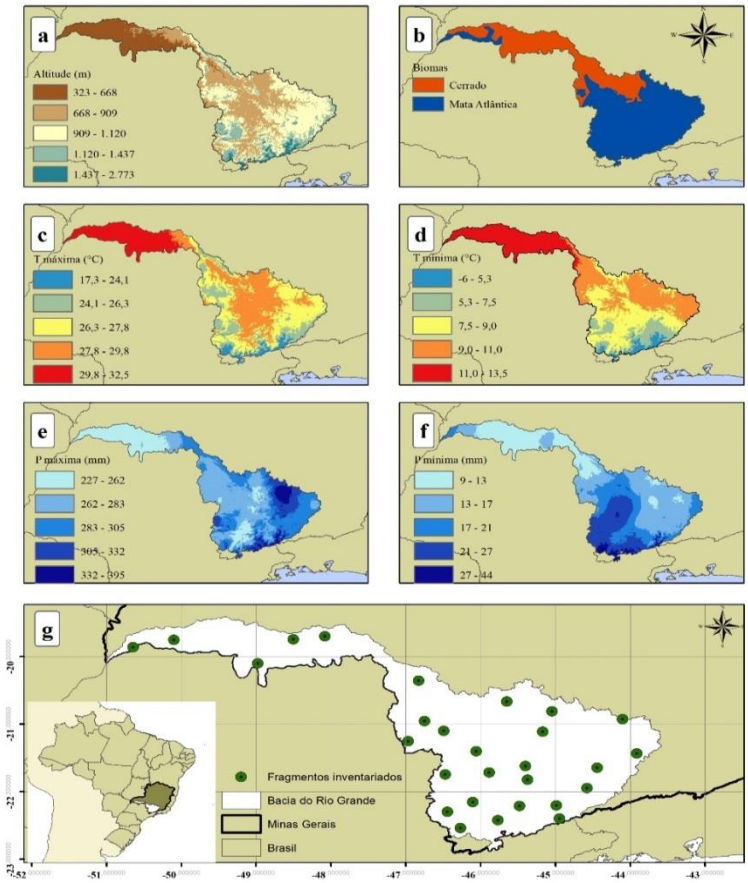
2.1 Descrição da área de estudo

A área de estudo compreende a bacia do Rio Grande situada no estado de Minas Gerais, com área de 86.110km². A bacia representa quase 15% do território mineiro, e está compreendida em parte do Centro-Sul de Minas e Triângulo Mineiro. A bacia situa-se em uma região de grande declividade, abrangendo as cadeias montanhosas da Mantiqueira e Canastra. As maiores altitudes, acima de 1.300 m, encontram-se na serra da Mantiqueira, ao extremo Sul de Minas Gerais. Em oposto situa-se a região do Triângulo Mineiro, com altitudes entre 300 e 700 m e relevo plano (FIGURA 1a). A abrangência da bacia hidrográfica do rio Grande garante a ela uma grande diversidade florística, compreendendo uma área de transição entre os biomas Cerrado e Mata Atlântica (FIGURA 1b).

De acordo com os dados mensais do WorldClim (HIJMANS et al., 2005) a temperatura média anual na bacia do Rio Grande fica em torno de 23 °C, é maior na região do baixo Rio Grande (Triângulo Mineiro) e menor no Sul de Minas. No extremo sul de Minas, na região da serra da Mantiqueira, a temperatura média do mês mais quente gira em torno de 17 a 22 °C e no mês mais frio pode chegar ao negativo enquanto que, no Triângulo Mineiro, a maior média mensal poder chegar até 10 °C a mais que na região serrana e a mínima não passa de 11 °C (FIGURAS 1c e 1d). A precipitação é outra característica bastante variável dentro da bacia. Considerando os extremos da serra da Mantiqueira e a parte oeste do Triângulo Mineiro, a diferença na precipitação média anual pode chegar a 1.200 mm. A região central da bacia apresenta uma média de 1.500 mm, a região mais seca uma média de 1.250 mm e a região mais úmida uma média de 2.200 mm. A média mensal do período chuvoso varia, de

220 mm a 400 mm enquanto que, nos meses do período de estiagem, a precipitação geralmente não ultrapassa 50 mm, e pode ser zero em alguns anos e regiões (FIGURAS 1e e 1f).

Figura 1 - Características topográficas, climáticas e vegetacionais da bacia do Rio Grande, sendo: a- Altitude (m); b- Área dos biomas Mata Atlântica e Cerrado; c- Temperatura máxima mensal; d- Temperatura mínima mensal; e- Precipitação média do mês mais úmido; f- Precipitação média do mês mais seco; g- Localização dos fragmentos inventariados.



Fonte: Da autora (2018).

2.2 Inventário florestal do Carbono da parte aérea

O inventário florestal foi realizado considerando a amostragem por conglomerado entre os anos de 2014 e 2015. O conglomerado foi formado por três subparcelas retangulares de área fixa de 250m² (10x25m), sendo cada unidade dispostass de forma sistematizada em transectos. De acordo com as classes climáticas e de altitude, ma amostra de 28 remanescentes de vegetação nativa foi selecionada (FIGURA 1g), onde foram lançadas 1007 subparcelas de conglomerados medindo todos os indivíduos com circunferência à altura do peito (CAP) superior a 15,7 cm (SCOLFORO et al., 2015).

A quantificação do estoque de carbono acima do solo foi realizada, a partir de uma cubagem rigorosa destrutiva envolvendo 232 árvores em 3 fitofisionomias, utilizando a metodologia proposta pela Food and Agricultural Organization of the United Nations - FAO (PICARD; SAINT-ANDRÉ; HENRY, 2012). Um modelo linear múltiplo para estimativa do teor de Carbono, utilizando as variáveis diâmetro a altura do peito (DAP) e altura (H), foi ajustado para cada fitofisionomia e aplicado a todas as árvores contidas nas respectivas parcelas, conforme sua fitofisionomia. Variáveis preditivas ambientais

A condição de macroescala estudada requer um trabalho investigativo de seleção das principais variáveis preditivas. No total, 114 variáveis foram consideradas, oriundas de diferentes fontes e representativas de características climáticas, morofométricas, geográficas, espectrais, texturais e edáficas (Tabela 1). A seleção inicial destas variáveis baseou-se na utilização em outros estudos para estimativa de variáveis dendrométricas em larga-escala (LU et al., 2004, 2014; BACCINI et al., 2008; SILVEIRA et al., 2019b). As variáveis climáticas foram extraídas do WorldClim, versão 1.4 (worldclim.org) (HIJMANS et al., 2005). No total, foram utilizadas as 19 variáveis que representam a média, mínimo, máximo e variação da temperatura e precipitação, com resolução

aproximada de 1km. A partir do modelo digital de elevação *Shuttle Radar Topography Mission* – SRTM, redimensionado para 100m de resolução espacial, foram calculadas 17 variáveis morfométricas, utilizando a ferramenta *Terrain Analysis* do software SAGA GIS (v. 6.3.0).

Os dados espectrais foram obtidos, 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. Do sensor OLI foram adquiridas 12 cenas para abranger toda a área de estudo, provenientes do *United States Geological Survey of Earth* (USGS/EROS), já apresentando as devidas correções geométricas e radiométricas. A partir desse mosaico foram calculados 7 índices de vegetação, utilizados como variáveis preditoras: 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). Sete texturas, utilizando a metodologia *Grey Level Co-occurrence Matrix* (GLCM) foram calculadas para cada índice de vegetação: variance (var), homogeneity (homog), contrast (contrast), dissimilarity (dissim), entropy (entrop), second moment (secmom), and correlation (correl). Para o cálculo destas, empregou-se uma janela de 61 linhas por 61 colunas (3721 pixels) (HAMUNYELA; VERBESSELT; HEROLD, 2016), processados no software ENVI Version 4.7 (*Exelis Visual Information Solutions, Boulder, Colorado*). As variáveis relacionadas à temperatura da superfície da Terra (emis32, lstd, lstn), atividade fotossintética (fpar, lai), evapotranspiração (et, le, pet, ple), produtividade

primária (gpp, psnnet) e porcentagem de cobertura vegetal (treecover) foram extraídas do sensor MODIS.

Características físico-químicas do solo no horizonte de 0-10cm também foram incorporadas ao conjunto de variáveis preditoras. A partir da amostragem do solo, nos fragmentos inventariados, obteve-se os teores de Dióxido de Silício (SiO_2) e Ferro Total (Fe), empregando o espectrômetro portátil de fluorescência de raios-X (pXRF) (SILVA et al., 2016). Outras variáveis como teor de matéria orgânica, pH, alumínio, argila e soma de bases foram retiradas diretamente da análise laboratorial dos solos. Esses dados pontuais foram interpolados utilizando valores de latitude e longitude, rasterizados com resolução espacial de 100m, por meio da ferramenta *Multilevel B-Spline* no software SAGA GIS (v. 6.3.0).

As coordenadas geográficas latitude (Y) e longitude (X) do centro das subparcelas foram adotadas visando a representar seus efeitos na modelagem matemática. A única variável categórica empregada, bioma, que apresenta duas classes - Cerrado e Mata Atlântica - foi obtida da base de dados do Zoneamento Ecológico Econômico de Minas Gerais (SCOLFORO; CARVALHO; OLIVEIRA, 2008).

Tabela 1 - Conjunto inicial de variáveis independentes utilizadas na modelagem do teor de Carbono acima do solo da vegetação nativa.

(continua)

TIPO	PREDITORES	SIGLA	RESOLUÇÃO (m)
Climática	Annual Mean Temperature	(BIO1)	1000
	Mean Diurnal Range	(BIO2)	
	Isothermality	BIO3	
	Temperature Seasonality	BIO4	
	Max Temperature of Warmest Month	BIO5	
	Min Temperature of Coldest Month	BIO6	
	Temperature Annual Range	BIO7	
	Mean Temperature of Wettest Quarter	BIO8	
	Mean Temperature of Driest Quarter	BIO9	
	Mean Temperature of Warmest Quarter	BIO10	
	Mean Temperature of Coldest Quarter	BIO11	
	Annual Precipitation	BIO12	
	Precipitation of Wettest Month	BIO13	
	Precipitation of Driest Month	BIO14	
	Precipitation Seasonality	BIO15	
	Precipitation of Wettest Quarter	BIO16	
	Precipitation of Driest Quarter	BIO17	
	Precipitation of Warmest Quarter	BIO18	
	Precipitation of Coldest Quarter	BIO19	

Tabela 1 - Conjunto inicial de variáveis independentes utilizadas na modelagem do teor de Carbono acima do solo da vegetação nativa.

(continuação)

TIPO	PREDITORES	SIGLA	RESOLUÇÃO (m)
Morfométrica	Altitude	altitude	100
	Analytical hillshading	hillshadin	
	Aspect	Aspect	
	Closed depressions	closed_dep	
	Channel network base level	cn_base_le	
	Convergence index	conv_index	
	Cross sectional curvature	c_sec_curv	
	Diffuse insolation	dif_insol	
	Direct insolation	direct_ins	
	Flow accumulation	flow_accum	
	Longitudinal curvature	long_curv	
	LS factor	ls_factor	
	Relative slope	relative_s	
	Valley depth	valley_dep	
	Vertical distance	vert_dist	
	Wetness index	wet_index	
	Slope (%)	slope_perc	
Espectral Landsat	EVI	evi	30
	Normalized Difference Vegetation Index	ndvi	
	Modified Soil-adjusted Vegetation Index	msavi	
	Normalized Burn Ratio	nbr	
	Normalized Burn Ratio 2	nbr2	
	Normalized difference moisture index	ndmi	
	Soil-adjusted Vegetation Index	savi	

Tabela 1 - Conjunto inicial de variáveis independentes utilizadas na modelagem do teor de Carbono acima do solo da vegetação nativa.

(conclusão)

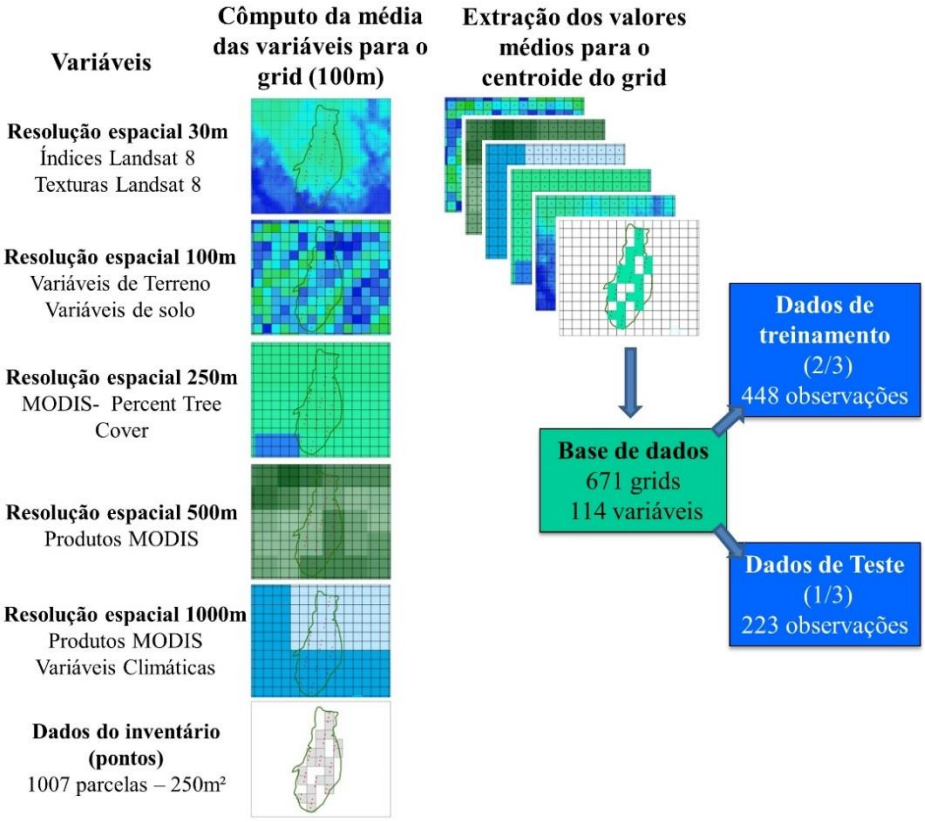
TIPO	PREDITORES	SIGLA	RESOLUÇÃO (m)
Espectral Texturas	Contrast	contra	30
	Correlation	correl	
	Dissimilarity	dissim	
	Entropy	entrop	
	Homogeneity	homog	
	Second Moment	sescom	
	Variance	var	
Espectral MODIS	Emissivity bands 32	emis32	1000
	Global evapotranspiration	et	500
	Fraction of photosynthetically active radiation	fpar	500
	Latent heat flux	le	500
	Land Surface Temperature day	lstd	1000
	Land Surface Temperature night	lstn	1000
	Potential global evapotranspiration	pet	500
	Potential latent heat flux	ple	500
	Percent Tree Cover	treecover	250
	Gross Primary Production	gpp	500
	Leaf Area Index	lai	500
	Net photosynthesis	psnnet	500
Edáfica	Clay	argila	100
	Aluminium	aluminio	
	Total iron	ferro_tota	
	Organic matter	m_o	
	Soil pH	ph	
	Sum of bases	sb	
	Silicon dioxide	sio2	
Geográfica	Latitude	X	-
	Longitude	Y	
	Biome	Biome	

Fonte: Da autora (2018).

2.3 Pré-processamento e padronização da base de dados

Todas as variáveis independentes bem como os dados coletados em campo foram primeiramente trabalhados na projeção *Albers Equal Area Conic* a fim de preservar as características de área e forma dos dados utilizados (DURO; FRANKLIN; DUBE, 2012; SILVEIRA et al., 2019b). Devido às diferentes fontes dos dados utilizados, bem como resoluções espaciais, optou-se por sobrepor os valores do teor de Carbono obtidos em campo ao conjunto de variáveis ambientais com o auxílio de um conjunto de grids, arquivo contendo uma malha quadrangular em formato shapefile, que abrange toda a área de estudo, com tamanho de 100 x 100m cada unidade do arquivo. Assim, para cada célula da malha (grid) em que uma ou mais parcelas estão inseridas, foram extraídos os valores médios das variáveis dependente e independentes para o centroide da célula. Para tal procedimento, foram utilizadas as ferramentas *Zonal Statistics* e *Extract Values* do software ArcGIS (v. 10.1). Do total de 1007 parcelas gerou 671 grids com valores médios do teor de Carbono ($Mg.ha^{-1}$) e das 114 variáveis ambientais, dados que serão utilizados no processamento dos algoritmos. Dessas 671 instâncias, 70% foram sorteadas para o desenvolvimento dos métodos (dados de treinamento) e o restante para sua avaliação (dados de teste) (FIGURA 2). O sorteio para a divisão da base de dados foi realizado de forma a manter a mesma distribuição dos dados originais (média, amplitude e percentis) em cada conjunto formado.

Figura 2 - Representação do pré-processamento da base de dados utilizada na modelagem do teor de carbono acima do solo da vegetação nativa.



Fonte: Da autora (2018).

2.4 Modelagem matemática do estoque de carbono

O algoritmo *Random Forest* (RF) (BREIMAN, 2001) foi selecionado para modelar o teor de Carbono neste estudo em razão de duas características principais: 1) Algoritmo com parametrização simples, o que facilita sua automatização no experimento com diferentes bases de dados; 2) Comprovada robustez e acurácia em diversas áreas, inclusive em estudos sobre a modelagem de dados de campo com a integração de dados espectrais em diversos tipos

florestais (AHMED et al., 2015; REIS et al., 2018; WU et al., 2016). O algoritmo é um método computacional, não paramétrico, utilizado em problemas de classificação e regressão, que lida com uma ampla gama de conjunto de variáveis (dados numéricos, categóricos, incompletos, multicolineares, com ruídos). Assim como uma série de algoritmos, este carece de uma parametrização inicial. A partir de testes iniciais, fixou-se o número de árvores (*ntrees*) em 1000 unidades, e número de atributos a serem sorteados (*mtry*) como a raiz quadrada do número total de atributos. Essa parametrização foi adotada em o experimento. Utilizou-se o pacote *randomForest* (LIAW; WIENER, 2002) para aplicar o algoritmo RF, bem como todos os códigos foram implementados no software R (R CORE TEAM, 2017).

Diante do subproblema metodológico, o uso do RF foi aplicado considerando 4 estratégias distintas. A primeira, considerou o uso puro do método, sem a remoção prévia de variáveis (*RFall*), ou seja, a forma clássica de aplicação. Na segunda estratégia, adotou-se uma remoção recursiva das variáveis (*RFrr*), em ordem crescente, de acordo com a importância da variável. As remoções são cumulativas, e a cada iteração é removida uma variável. Ao novo conjunto de variáveis formado, a cada iteração, o algoritmo RF é aplicado com 10 repetições e o erro médio quadrático dessas repetições (*out-of-bag* - OOB) é então memorizado. Após realizar a varredura de toda a base de dados, então se identifica o subconjunto de variáveis que propicia o menor erro quadrático médio do modelo, de acordo com o conjunto interno de validação do algoritmo (OOB).

Duas outras estratégias foram introduzidas nos testes, a partir da ideia de algoritmos híbridos, ou seja, a união de uma metaheurística e algoritmos de treinamento de máquina. Assim, optou-se por associar o algoritmo genético para a seleção de variáveis na aplicação do *Random Forest* (AGRF). O algoritmo genético (AG) foi parametrizado conforme Monti (2018), cujos operadores

genéticos foram adaptados para a forma binária. Adotou-se o operador de seleção torneio, com o operador de cruzamento executando trocas genéticas (bit-a-bit) entre pais selecionados, retornando novos indivíduos da população. As taxas de seleção e cruzamento dos operadores foi de 50%. O aumento da diversidade de bases da população foi gerado pelo operador de mutação, com taxa de 10% de indivíduos e 50% em troca aleatória dos genes. Os indivíduos foram dimensionados por um vetor fixo de 114 posições (*genes*), correspondendo às variáveis analisadas. A população inicial foi estabelecida em 100 indivíduos e 100 gerações como critério de parada.

Sob a ótica de funções uni e multiobjetivas, buscou-se aplicar sua forma pura de erro, considerando a condição uniobjetivo. Nesse caso, o *fitness* de cada indivíduo deriva do erro OOB gerado por cada RF processado (*AGRFuni*), de natureza uniobjetiva. A condição multiobjetiva (*AGRFmulti*), formada pela minimização do erro OOB e o número de variáveis, é descrita em equação (2). Para tal situação, o *fitness* consistiu na soma de dois termos, o primeiro, da razão entre o erro OOB obtido pelo RF e o seu máximo observado via testes, sendo aqui definido como uma constante. A segunda parte da expressão retrata a condição número de variáveis, e por isso foi acrescido a razão entre número de variáveis habilitadas (*n*) e o total de variáveis candidatas testadas.

$$fitness = \left(\frac{erro\ OOB}{1.060} + \frac{n}{114} \right) \tag{2}$$

2.5 Análise comparativa dos métodos

A natureza estocástica dos métodos testados requisitou o uso de repetições, para garantir uma comparação mais acertiva. Nesse caso, foram

realizadas 50 repetições do RF para cada metodologia empregada, foi obtido o erro OOB, derivado do conjunto interno de validação (treino) a cada repetição, bem como para o teste. As métricas adotadas para os dois conjuntos de dados foram: a) erro médio (ME – *mean error*) (3), b) raiz do erro quadrático (RMSE – *root mean squared error*) (4), c) erro médio absoluto percentual (MAPE – *mean absolute percentual error*) (5), e d) gráficos de resíduos. Sendo $i =$ número da instância, $n =$ número total de instâncias do conjunto de dados (treinamento 448 instâncias e teste 223), $y_i =$ valor real do teor de Carbono (Mg.ha^{-1}) da instância i; $\hat{y}_i =$ valor estimado médio do teor de Carbono (Mg.ha^{-1}) da instância i; $\overline{Y} =$ valor médio real do teor de Carbono (Mg.ha^{-1}).

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{3}$$

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - \hat{y}_i)^2} \tag{4}$$

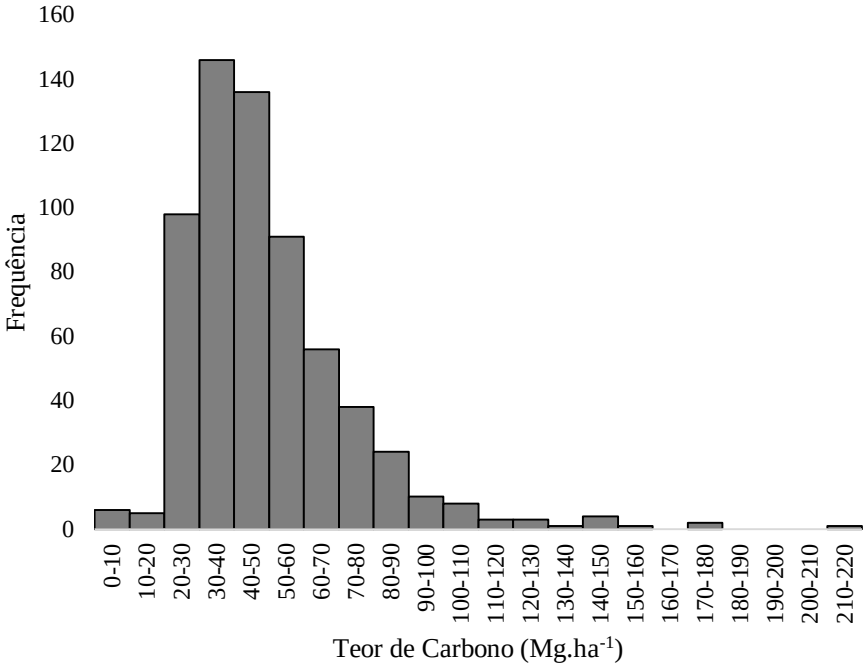
$$RMSE\% = \frac{RMSE}{\overline{Y}} \times 100 \tag{5}$$

O tempo médio de processamento destas metodologias também foi avaliado na análise comparativa. Todo o experimento foi processado em um computador com processador Intel® Core™ i3-2100 de 3.10 MHz e 8 Gb de memória RAM.

3 RESULTADOS

A média do teor de Carbono foi de 47,29 $Mg.ha^{-1}$, variando de 3,84 a 214,96 $Mg.ha^{-1}$, com desvio padrão de 24,65 $Mg.ha^{-1}$ e coeficiente de variação de 52,13%. De acordo com o histograma dos dados, valores abaixo de 20 $Mg.ha^{-1}$ e acima de 120 $Mg.ha^{-1}$ têm baixa ou até nenhuma representatividade no conjunto de dados (FIGURA 3).

Figura 3 - Histograma dos valores do teor de Carbono ($Mg.ha^{-1}$) acima do solo para a vegetação nativa da bacia do Rio Grande encontrados nas 671 grids de 1ha.

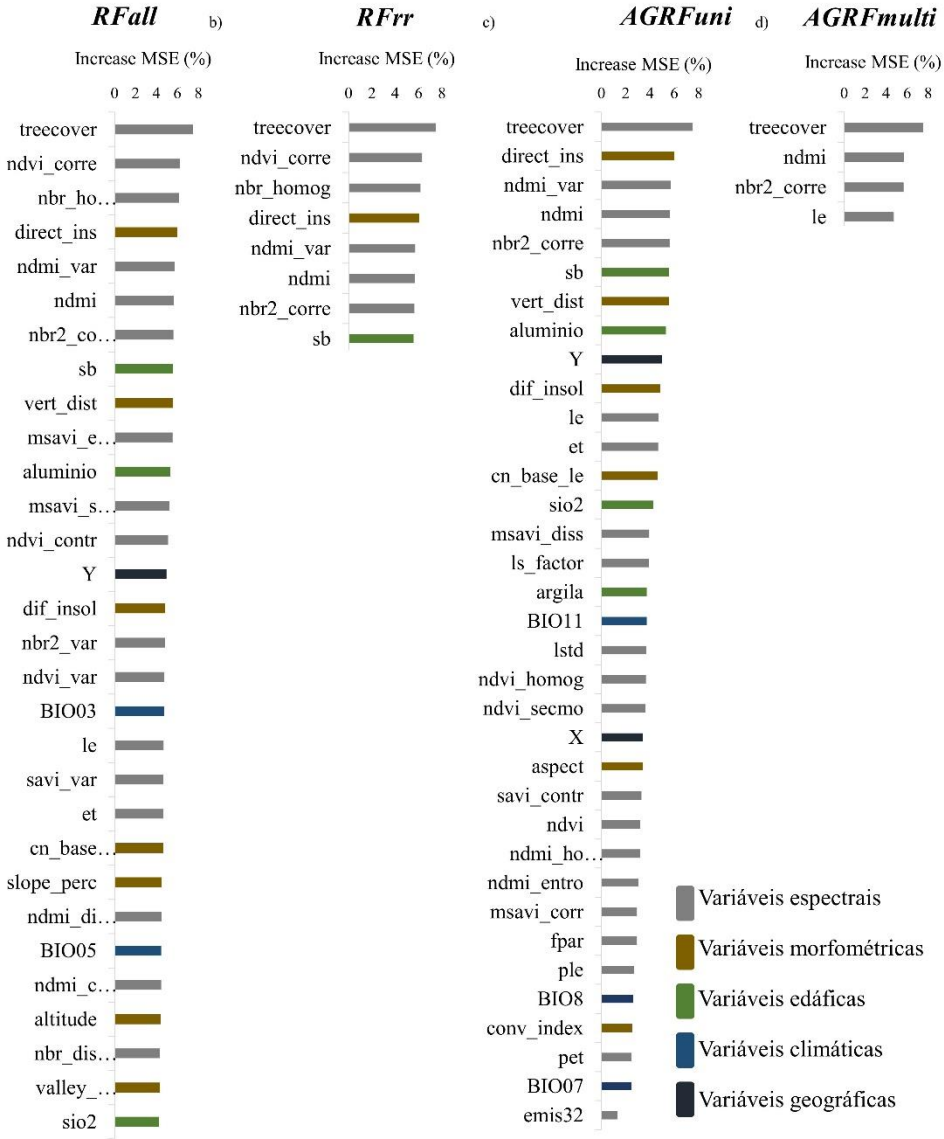


Fonte: Da autora (2018).

A primeira abordagem com o algoritmo *Random Forest* (*RFall*), além de possibilitar a estimativa da contribuição relativa de cada variável, serviu como

“testemunha” para comparações com as metodologias de seleção de variáveis empregadas nesta pesquisa. A média da importância de cada variável está representada na Figura 4a, na qual é possível observar as 30 variáveis com maior valor de importância. Destas 30, 17 são variáveis oriundas de imagens de satélites, 7 são variáveis do terreno, 3 são variáveis do solo, 2 são climáticas e 1 geográfica. As três variáveis que mais contribuem para o modelo são espectrais, dentre elas a varíavel TreeCover, do satélite MODIS, com maior valor de importância médio, de 7,51%, destacando-se das demais.

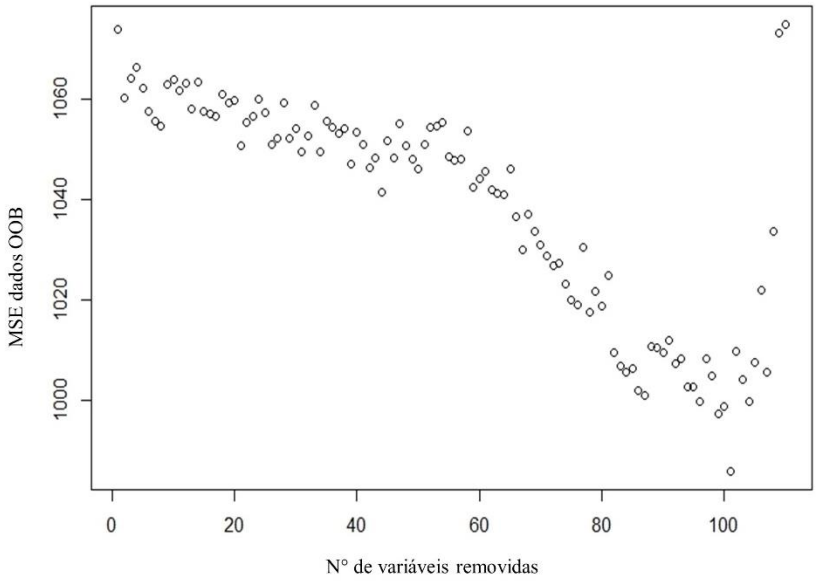
Figura 4 - Valores médios de importância das variáveis, sendo a) Importância média das 30 variáveis com maiores valores de importância, de acordo com a metodologia *RFall;* b) Variáveis selecionadas pela metodologia *RFrr;* c) Variáveis selecionadas pela metodologia *AGRFuni;* d) Variáveis selecionadas pela metodologia *AGRFmulti*.



Fonte: Da autora (2018).

De posse dos valores médios da importância de cada variável, executou-se a metodologia *RFrr*. O conjunto de variáveis que obteve menor erro médio foi formado pelas 8 variáveis com maior valor de importância (Figura 4b). Esse conjunto é constituído de 6 variáveis espectrais, 1 variável de terreno e 1 variável de solo. Todas as variáveis possuem valores de importância acima de 5,56%. É possível observar clara tendência de diminuição do erro OOB com a remoção de até cerca de 80 variáveis com menor valor de importância (Figura 5).

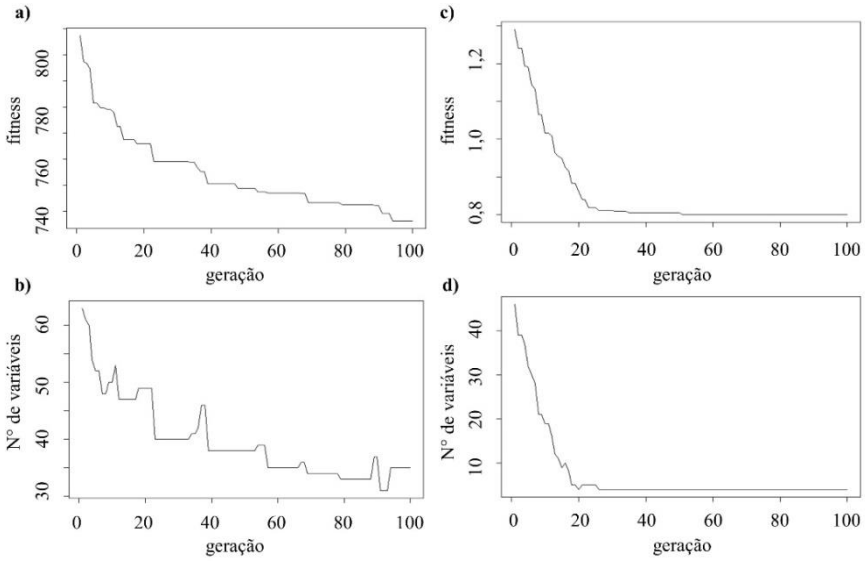
Figura 5 - Decaimento do erro médio quadrático interno do *RFrr* (MSE dados OOB) com a remoção recursiva das variáveis seguindo ordem crescente dos valores médios de importância provenientes da metodologia *RFall*.



Fonte: Da autora (2018).

A metodologia *AGRFuni*, empregando o algoritmo genético, encontrou um conjunto ótimo de 35 variáveis (Figura 4c). Cerca de 57% dessas variáveis (20 variáveis), são espectrais. A variável de maior valor de importância, TreeCover, foi selecionada pelo modelo, bem como outras variáveis de importância que também foram selecionadas pelo *RFrr*, como: Insolação direta (direct_insol), NDMI variância (ndmi_var), NDMI (ndmi), NBR2 correlação (nbr2_corre) e soma de bases (sb). No entanto, o algoritmo também selecionou grande quantidade de variáveis com valores de importância baixos, abaixo de 4% de aumento do erro (FIGURA 4a). O decaimento do erro (fitness) bem como a quantidade de variáveis testadas ao longo das gerações pelo *AGRFuni* pode ser observado na Figura 6.

Figura 6 - Comportamento do erro e do número de variáveis testadas ao longo das gerações para os algoritmos genéticos *AGRFuni* e *AGRFmulti*, onde a) e b) correspondem ao funcionamento do *AGRFuni* e c) e d) ao funcionamento do *AGRFmulti*.



Fonte: Da autora (2018).

O algoritmo genético multiobjetivo (*AGRFmulti*), que visa tanto a diminuição do erro quanto a diminuição do número de variáveis, encontrou um conjunto de menor erro com apenas 4 variáveis. São elas: TreeCover, NDMI, NBR2 correlação (nbr2_corre) e Latent heat flux (le), todas variáveis espectrais (FIGURA 4d). A variável de menor valor de importância selecionada é a *Latent heat flux*, do sensor MODIS, com 4,69% de aumento no erro. As outras três variáveis selecionadas pertencem ao grupo das oito variáveis com maior valor de importância. Em relação ao funcionamento do algoritmo genético, percebe-se que o valor da função fitness estabilizou perto da geração 50 (FIGURA 6c), e o número de variáveis por volta de 20 gerações antes (FIGURA 6d).

O algoritmo RF demonstrou desempenho consistente independente dos conjuntos de dados testados. As diferenças nos erros de predição são pequenas, quando comparadas as 4 metodologias, seguindo tendência similar no treinamento e teste (TABELA 4). Contudo, as três metodologias propostas garantiram uma redução no número de variáveis e uma redução absoluta dos erros. No treinamento, o *AGRFmulti* foi o que mais se destacou, com ME de -0,88 $Mg.ha^{-1}$ e RMSE de 16,92 $Mg.ha^{-1}$. Valores próximos foram observados pela metodologia *RFrr*. Por outro lado, a abordagem *AGRFuni* foi superior ao algoritmo *RFall*, nas métricas de avaliação e número de variáveis utilizadas.

Tabela 2 - Valores das métricas de avaliação das metodologias para os dados de treinamento e teste, onde ME – erro médio; RMSE – raiz do erro quadrático médio; RMSE – raiz do erro quadrático médio percentual; N – número de variáveis final.

Dados	Estratégia	ME	RMSE	RMSE (%)	Tempo* (s)	N
treino	*RFall*	-1,68	17,56	36,89	841,04	114
	RFrr	-1,26	16,89	35,49	10.220,66	8
	AGRFuni	-1,33	17,03	35,79	82.800,00	35
	AGRmulti	-0,88	16,92	35,56	16.560,00	4
teste	*Rfall*	-2,18	18,81	39,22	-	114
	RFrr	-1,61	18,11	37,74	-	8
	AGRFuni	-1,61	18,34	38,24	-	35
	AGRFmulti	-0,66	17,75	37,00	-	4

Fonte: Da autora (2018).
* tempo considerando 50 repetições do RF com a base selecionada.

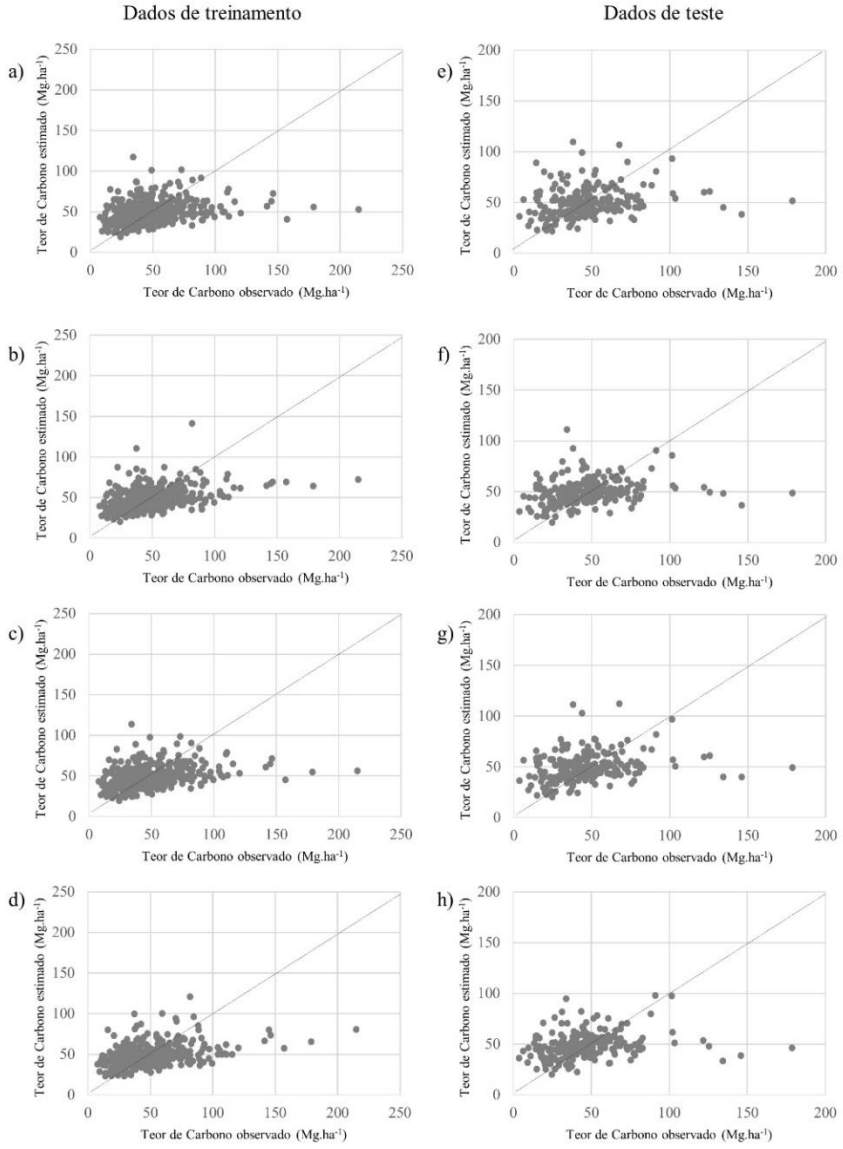
O desempenho dos algoritmos na validação (teste) foi inferior quando comparado à base de treinamento. A raiz do erro médio quadrático percentual dos algoritmos, em média de 35,9% na base treinamento, subiu para uma média de 38% na base de teste. No entanto, o ranqueamento dos algoritmos permanece o mesmo: *RFall* com pior desempenho; *RFrr* e *AGRFuni* com desempenho similares, com ligeira superioridade de *RFrr*; e *AGRFmulti* com melhor desempenho nas três métricas analisadas (TABELA 4).

Em relação ao tempo de processamento das metodologias, a lógica se inverte. O algoritmo com menor tempo de processamento médio foi *RFall*, com 841,04 segundos. Seguido pelo *RFrr*, com 10.220,66 segundos. As abordagens que utilizaram o algoritmo genético, superaram e muito o tempo médio das outras duas metodologias. O algoritmo genético uniobjetivo (*AGRFuni*) obteve maior tempo médio de processamento, 82.800 segundos. Já o algoritmo genético multiobjetivo diminuiu o tempo médio para 16.560 segundos (TABELA 4). Essa diferença do tempo de processamento está ligada ao número de variáveis

selecionadas para treinar o RF em cada geração, uma vez que, quanto maior o conjunto de dados em cada geração, maior o tempo de processamento.

Ao analisar os gráficos dos valores de teor de Carbono observados e estimados (FIGURA 7), percebe-se uma tendência para ambas as abordagens: superestimar valores abaixo da média dos teores de Carbono observados e subestimar valores acima, média esta em torno de 40-50 $Mg.ha^{-1}$. Os maiores erros situam-se nos maiores valores observados de teor de carbono, acima de 140 $Mg.ha^{-1}$. A perfomance das diferentes abordagens foram semelhantes ente si, com pequena diminuição da amplitude dos valores estimados para *RFrr* e *AGRFmulti*. As estimativas para os dados observados foram ligeiramente mais dispersas na base de teste em relação aos dados de treinamento .

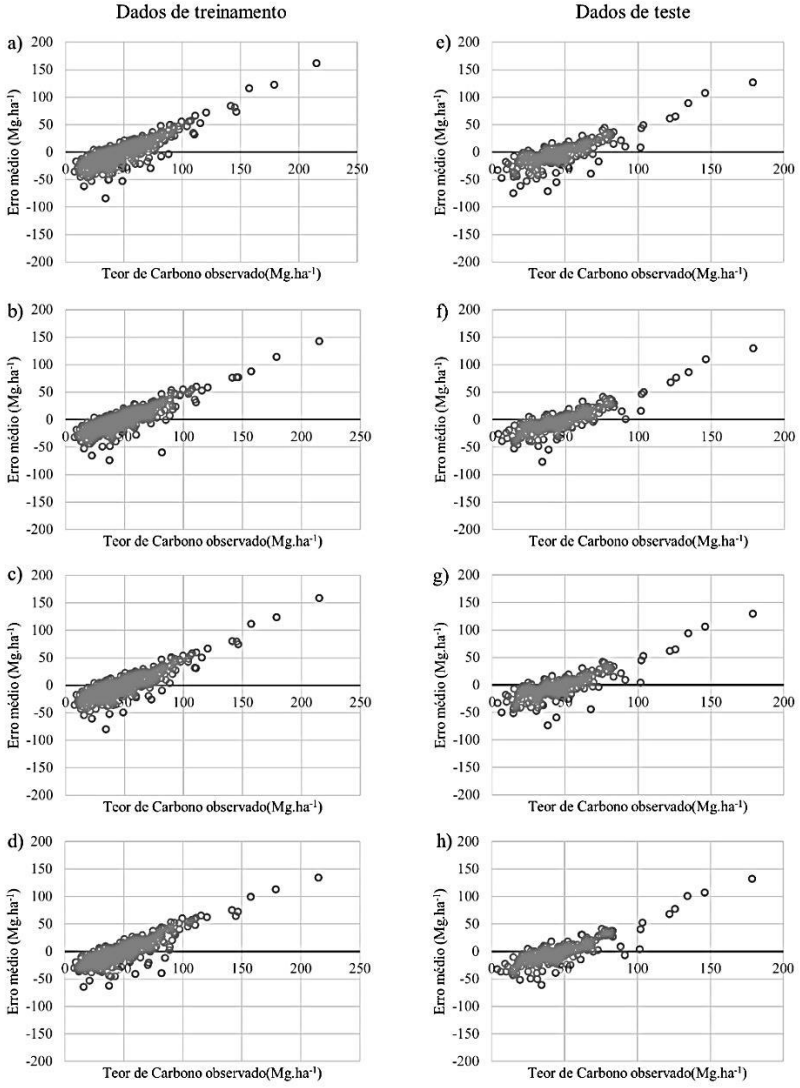
Figura 7 - Gráficos dos valores estimados do teor de Carbono em relação aos valores observados, sendo a, b, c, d valores estimados pelas metodologias RFall, RFrr, AGRFuni e AGRFmulti, respectivamente, para os dados de treinamento; e, f, g, h valores estimados pelas metodologias RFall, RFrr, AGRFuni e AGRFmulti, respectivamente, para os dados de teste.



Fonte: Da autora (2018).

A Figura 8 evidencia a propensão de maiores erros nos valores superiores de teor de Carbono, subestimando-os. No entanto, a grande quantidade de dados abaixo da média dos valores de teor de Carbono observados, resulta em uma superestimativa geral dos modelos (ME negativo). Mais uma vez *RFrr* e *AGRFuni* exibem modesta diminuição dos valores de EM, tanto para superestimativas quanto para subestimativas.

Figura 8 - Gráficos dos valores do erro médio em relação aos valores observados, sendo a, b, c, d valores estimados pelas metodologias *RFall*, *RFrr*, *AGRFuni* e *AGRFmulti*, respectivamente, para os dados de treinamento; e, f, g, h valores estimados pelas metodologias *RFall*, *RFrr*, *AGRFuni* e *AGRFmulti*, respectivamente, para os dados de teste.



Fonte: Da autora (2018).

4 DISCUSSÃO

O presente estudo se propôs a testar diferentes metodologias para seleção de variáveis na utilização do algoritmo RF para a modelagem do estoque de carbono acima do solo da vegetação nativa. Por meio da utilização da técnica mais utilizada na seleção de atributos para o RF, remoção recursiva das variáveis em ordem crescente de importância, e de algoritmos genéticos investigou-se o efeito destas seleções na performance do RF.

Os resultados obtidos indicam o algoritmo RF como um método robusto, pouco afetado pela inclusão de um grande número de variáveis relacionadas entre si. As diferenças entre as métricas de avaliação e gráficos de resíduos e estimativas são irrisórias. Mesmo com a pequena melhora nos erros do algoritmo, o uso das técnicas de seleção de variáveis se justifica quando diminui consideravelmente o número de variáveis a serem utilizadas. Essa redução facilita na aplicação e entendimento do algoritmo, bem como na geração posteriori dos mapas com as estimativas espacializadas.

Lafiti, Nothdurf e Koch (2010) aplicaram algoritmo genético para seleção de variáveis na predição de volume e biomassa de uma floresta temperada, comparando métodos não paramétricos, como medidas de distâncias Euclidiana, Mahalanobis e Vizinho mais próximo, além do algoritmo Random Forest. Os resultados obtidos apontaram uma diminuição do erro quando aplicado o algoritmo genético para os métodos de distância Euclidiana e de Mahalanobis, enquanto o Random Forest apresentou melhor desempenho na base de dados completa. Dados esses que respaldam a afirmação de que o RF consegue realizar boas predições com grande número de variáveis (BREIMAN, 2001; VAUHKONEN et al., 2010).

O método de remoção recursiva das variáveis em ordem crescente de importância (*RFRR*) mostrou-se de fácil utilização e processamento quando

comparado aos algoritmos genéticos, e desempenho similar em relação aos erros obtidos. No entanto, o método analisa subconjuntos de dados de forma unidirecional, seguindo a ordem crescente dos valores da importância das variáveis a cada processamento. Esse mecanismo limita o número de subconjuntos testados, restringindo as combinações entre as variáveis, e com isso tende a selecionar variáveis na mesma faixa de importância. Como efeito negativo, muitas vezes, o conjunto de variáveis selecionadas é altamente correlacionado, dividindo a base de dados de maneira similar.

No caso dos algoritmos genéticos, como esperado, expressaram maior dificuldade de processamento (maior tempo) e parametrização. O algoritmo genético com função única de diminuição do erro (*AGRFuni*) pode não ser uma boa escolha para ser aplicado na seleção de variáveis para o RF. Primeiro, o algoritmo RF mostra-se pouco sensível à diminuição do número de atributos. Segundo, o algoritmo genético, devido à pouca variação do erro, pode acabar selecionando grande subconjunto de dados, como aconteceu nesta pesquisa.

O desafio nesse estágio foi identificar qual estratégia seguir, no gerenciamento do algoritmo, com o propósito de redução do erro final, como ainda a seleção de variáveis explicativas do carbono na parte aérea. Nesse sentido, o uso de um número elevado de variáveis compromete a aplicação do método em macroescalas, pois exige-se maior tempo de processamento e memória para as predições. Logo, a busca pela redução das variáveis é um fator auxiliar na viabilidade do método. A abordagem *AGRFmulti* mostrou-se ser uma opção alternativa ao método comumente empregado (*RFRR*). Além de obter menores erros, a metodologia selecionou o menor conjunto de variáveis, com apenas a metade do número de variáveis selecionado por *RFrr*. Demonstrou também menor tempo de processamento quando comparado ao *AGRFuni*, mas ainda um pouco maior que *RFRR*. Essa diferença temporal em relação ao *AGRFuni* está relacionada ao tempo gasto para o treinamento do RF a cada

geração, é diretamente proporcional ao número de variáveis em cada subconjunto. As variáveis selecionadas por *AGRFMULTI* figuram entre as variáveis com maiores valores de importância, no entanto sem seguir uma ordem pré-estabelecida. Fato que demonstra a coerência do método, uma vez que os valores de importância não são apresentados ao algoritmo.

A aplicação de métodos híbridos utilizando o algoritmo *Random Forest* e algoritmos genéticos não é uma prática incomum no meio científico. Diversas abordagens foram propostas, seja para otimização dos parâmetros do RF utilizando algoritmo genético (MING et al., 2016), para seleção das melhores árvores dentro da floresta (BADER EL-DEN, 2014; BADER-EL-DEN; GABER, 2012) ou mesmo para seleção de atributos (JABBAR; BULUSU; PRITI, 2016; KUMAR; SAHOO, 2017; LAFITI; NOTHDURF; KOCH, 2010). Todas as abordagens, assim como nesta pesquisa, apontam para a otimização dos resultados, apesar da dificuldade de processamento dos métodos. Kumar e Sahoo (2017) testaram cinco métodos de seleção de atributos utilizando o RF, dentre eles algoritmo genético, para diagnóstico de doenças cardiovasculares. Dentre os métodos, os autores destacam a performance do algoritmo genético, que além de potencializar as estatísticas, diminuiu pela metade o número de variáveis selecionadas. No presente estudo, a redução foi de 96,5% considerando a estratégia multiobjetivo do algoritmo genético.

Relativamente às variáveis que mais contribuem para estimativas do teor de Carbono acima do solo, nota-se maior contribuição das espectrais, seguidas pelas variáveis morfométricas e do solo. As variáveis climáticas não exibiram relevância para os dados de Carbono na área estudada, assim como em Baccini et al. (2008). Mesmo com problemas de saturação da reflectância em vegetações com alto índice foliar e fechamento do dossel, como em florestas tropicais, relatados em diversos trabalhos para predição da biomassa e Carbono acima do solo (LU et al., 2004, 2014), os índices e produtos espectrais e suas texturas

apresentam dados mais informativos do que outras variáveis em contexto de Sistema de Informação Geográfica.

No que tange aos valores do teor de Carbono encontrados na bacia do Rio Grande e à sua modelagem, pode-se afirmar que a área de estudo apresenta grande heterogeneidade de valores e que estes não são homogeneamente representados no conjunto de dados, o que diminui a precisão da modelagem.

A área abrangida pela bacia hidrográfica do Rio Grande compreende a transição dos biomas Cerrado e Mata Atlântica, entretanto, exibe valores médios do estoque de Carbono mais próximos aos encontrados no bioma Mata Atlântica, nas regiões oeste e centro-sul de Minas Gerais. De acordo com Scolforo et al. (2015), os valores médios para o estoque de Carbono acima do solo para os biomas Cerrado e Mata Atlântica são de 21,5 e 55,0 $Mg.ha^{-1}$ para a região centro-sul, e 26,6 e 47,8 $Mg.ha^{-1}$para região oeste do estado, respectivamente. As fitofisionomias que compreendem essas regiões variam desde florestas com alto número de indivíduos e árvores grandes (floresta ombrófila, cerradão e até mesmo floresta estacional semi-decidual) a áreas com baixa densidade de indivíduos e pequeno porte das árvores (campo cerrado, cerrado sensu strictu). Isso garante aos dados de Carbono da região uma variação de mais de 50% entre os valores que compõe a base de dados para a modelagem.

Essa grande variação dos valores do teor de Carbono aparenta não ter sido acompanhada pela variação das variáveis independentes, ocasionando baixos valores de Importância da variável. O valor máximo de importância encontrado nesta pesquisa foi de 7,71% de aumento no erro do algoritmo. Pandit, Tsuyuki e Dube (2018), ao tentarem estimar os valores de biomassa acima do solo para vegetação nativa do Nepal utlizando bandas e índices provenientes do satélite Sentinel 2, obtiveram valores máximos de importância da variável, segundo a metodologia RF, em torno de 25%. Também modelando

o estoque de biomassa acima do solo, Martin Karlson et al. (2015), obtiveram valores de importância para o NDVI e EVI oriúndos de imagem Landsat 8, acima dos 30%. Tal constatação pode estar associada à grande variação dos teores de Carbono bem como à resolução das variáveis independentes em relação ao tamanho das parcelas em campo. Pandit, Tsuyuki e Dube (2018) utilizou 113 parcelas de 500m² cada, já Martin Karlson et al. (2015) utilizaram 75 parcelas de 2.500 m². Neste trabalho utilizamos parcelas de 250m² que foram extrapoladas em grids de 1ha (100x100m) por meio da média das parcelas dentro da grid.

O baixo poder de explicação das variáveis reflete também no desempenho do algoritmo que, mesmo em sua melhor abordagem (*AGRFmulti*), apresenta heterogeneidade dos resíduos, subestimando teores de Carbono acima da média e superestimando valores abaixo. Essa tendência, também encontrada por Baccini et al. (2008) e Gao et al. (2018), é intrínseca de modelos de regressão baseados em árvores, nos quais suas predições são a média dos valores nos nós terminais.

Mesmo com a robustez dos métodos e com a quantidade de variáveis testadas, o menor RMSE encontrado neste experimento foi de 16,92 $Mg.ha^{-1}$ (*AGRFmulti* treinamento). Valor este superior ao encontrado em Scolforo et al. (2015), de 16,21 $Mg.ha^{-1}$, ao modelar o teor de Carbono acima do solo da vegetação nativa para o estado de Minas Gerais utilizando regressão linear múltipla em função de longitude, latitude, altitude e bioma. Novamente, é válido salientar que em Scolforo et al. (2015) os teores de Carbono foram trabalhados em nível de fragmento, com média das parcelas no interior do fragmento extraídas para o seu centroide, totalizando 163 valores médios do teor de Carbono ($Mg.ha^{-1}$).

Diante disso surge uma hipótese para futuras pesquisas de se trabalhar com dados do inventário em escala maior, seja em nível de fragmento ou mesmo

com grids com resoluções mais grosseiras. Baccini et al. (2008) modelaram a biomassa acima do solo utilizando a integração de medições do inventário com grids de 1km² e bandas do MODIS, e obtiveram 82% da variação dos dados explicada pelo RF. Silveira et al. (2019b) conseguiram melhora na performance do RF para estimar a biomassa acima do solo em florestas tropicais utilizando a análise orientada a objeto em comparação com a metodologia pixel a pixel.

5 CONCLUSÕES

A média do teor de Carbono encontrada nos fragmentos florestais da bacia do Rio Grande foi de 47,29 $Mg.ha^{-1}$, variando em uma faixa de 3,84 a 214,96 $Mg.ha^{-1}$, e apresentando um coeficiente de variação dos dados de mais de 50%. Dentre as variáveis testadas para estimar o teor de Carbono acima do solo, nota-se maior contribuição das espectrais, seguidas pelas variáveis de terreno e do solo. A variável TreeCover (MODIS), além de demonstrar maior importância relativa de acordo com o Random Forest, foi selecionada pelas três metodologias de seleção de variáveis. O algoritmo Random Forest comprovou sua robustez, por ser pouco afetado pela inclusão de um grande número de variáveis relacionadas entre si. Os métodos de seleção de variáveis por remoção recursiva e algoritmo genético multiobjetivo, além de apresentarem ligeira diminuição do erro, diminuíram consideravelmente o número de variáveis, com vantagem para o último método. Diante dos resultados obtidos nesta pesquisa, indica-se o uso do algoritmo genético na seleção de variáveis para a modelagem do estoque de Carbono acima do solo utilizando o algoritmo Random Forest.

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.

BACCINI, A. et al. A first map of tropical Africa's above-ground biomass derived from satellite imagery. **Environmental Research Letters**, Bristol, v. 3, n. 4, 2008. Disponível em: <https://iopscience.iop.org/article/10.1088/1748-9326/3/4/045011/meta>. Acesso em: 10 dez. 2018.

BADER-EL-DEN, M. Self-adaptive heterogeneous Random Forest. In: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS, 11., 2014, New York. **Proceedings…** New York: IEEE/ACS, 2014. p. 640-646.

BADER-EL-DEN, M.; GABER, M. Garf: towards self-optimised Random Forests. In: _____. **Neural information processing**. New York: Springer, 2012. p. 506-515.

BELLARD, C. et al. Impacts of climate change on the future of biodiversity. **Ecology Letters**, Oxford, v. 15, n. 4, p. 365-377, 2012.

BOLIVAR, J.; GUTIERREZ-VELEZ, V.; SIERRA, C. Carbon stocks in aboveground biomass for Colombian mangroves with associated uncertainties. **Regional Studies in Marine Science**, New York, v. 18, p. 145-155, 2018.

BREIMAN, L. Random forest. **Machine Learning**, Boston, v. 45, p. 5-32, 2001.

CHEN, Q. Lidar remote sensing of vegetation biomass. In: WENG, Q.; WANG, G. (Ed.). **Remote sensing of natural resources**. Boca Raton: CRC Press; Taylor &Francis Group, 2013. p. 399-420.

DIAZ-URIARTE, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using Random Forest. **BMC Bioinformatics**, London, v. 8, p. 328, 2007.

DUBE, T.; MUTANGA, O. The impact of integrating WorldView-2 sensor and environmental variables in estimating plantation forest species aboveground biomass and carbon stocks in uMgeni Catchment, South Africa. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 119, p. 415-425, Sept. 2016.

DURO, D. C.; FRANKLIN, S. E.; DUBE, M. G. A comparison of pixel-based and objectbasedimage analysis with selected machine learning algorithms for the classificationof agricultural landscapes using SPOT-5 HRG imagery. **Remote Sensing Environment**, New York, v. 118, p. 259-272, 2012.

FROLKING, S. et al. Forest disturbance and recovery: a general review in the context of spaceborne remote sensing of impacts on aboveground biomass and canopy structure. **Journal of Geophysical Research:** Biogeosciences, v. 114, p. G00E02, 2009.

GAO, Y. et al. Comparative analysis of modeling algorithms for forest aboveground biomass estimation in a subtropical region. **Remote Sensing**, Basel, v. 10, p. 627, 2018.

HAMUNYELA, E.; VERBESSELT, J.; HEROLD, M. Using spatial context to improve early detection of deforestation from Landsat time series. **Remote Sensing of Environment**, New York, v. 172, p. 126-138, 2016.

HANSEN, M. C. et al. High-resolution Global Maps of 21st-century Forest Cover Change. **Science**, New York, v. 342, p. 850-853, 2013.

HIGUCHI, N. et al. Dinâmica e balanço do carbono da vegetação primária da amazônia central. **Floresta**, Curitiba, v. 34, n. 3, p. 295-304, 2004.

HIJMANS, R. et al. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, Chichester, v. 25, p. 1965-1978, 2005.

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

JABBAR, A.; BULUSU, D.; PRITI, C. Intelligent heart disease prediction system using Random Forest and evolutionary approach. **Journal of Network and Innovative Computing**, Auburn, v. 4, p. 175-184, 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.

KUMAR, S.; SAHOO, G. A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis. **International Journal of Engineering, Transactions B:** Applications, Karaj, v. 30, n. 11, p. 1723-1729, Nov. 2017.

LAFITI, H.; NOTHDURF, A.; KOCH, B. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. **Forestry:** An International Journal of Forest Research, London, v. 83, n. 4, p. 395-407, Oct. 2010.

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

LU, D. et al. Aboveground forest biomass estimation with landsat and LiDAR data and uncertainty analysis of the estimates. **International Journal of Forestry Research**, Cairo, v. 2012, p. 1-16, 2012.

LU, D. et al. Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin. **Forest Ecology and Management**, Amsterdam, v. 198, p. 149-167, 2004.

LU, D. et al. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. **International Journal of Digital Earth**, London, v. 9, p. 63-105, 2014.

LU, X. et al. Combining multi-source remotely sensed data and a process-based model for forest aboveground biomass updating. **Sensors**, New York, v. 17, p. 2062, 2017.

MARTIN KARLSON, M. et al. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian Woodlands Using Landsat 8 and Random Forest. **Remote Sensing**, Basel, v. 7, p. 10017-10041, 2015.

MILLARD, K.; MURRAY, R. On the importance of training data sample selection in random forest image classification: a case study in Peatland Ecosystem Mapping. **Remote Sensing**, Basel, v. 7, n. 7, p. 8489-8515, 2015.

MILLER, J. D.; THODE, A. E. Quantifying burn severity in a heterogeneous landscape 676 with a relative version of the delta Normalized Burn Ratio (dNBR). **Remote Sensing of Environment**, New York, v. 109, p. 66-80, 2007.

MING, D. et al. Land cover classification using Random Forestwith genetic algorithm-based parameter optimization. **Journal of Applied Remote Sensing**, Orlando, v. 10, n. 3, 2016. Disponível em: <https://www.spiedigitallibrary.org/journals/journal-of-applied-remote-sensing/volume-10/issue-03/035021/Land-cover-classification-using-random-forest-with-genetic-algorithm-based/10.1117/1.JRS.10.035021.full?SSO=1>. Acesso em: 10 dez. 2018.

MONTI, C. A. U. **Otimização aplicada à engenharia florestal**. 2018. 104 p. Dissertação (Mestrado em Engenharia Florestal)-Universidade Federal de Lavras, Lavras, 2018.

MORAIS, V. A. et al. Carbon and biomass stocks in a fragment of cerradão in Minas Gerais state, Brazil. **Cerne**, Lavras, v. 19, n. 2, p. 237-245, 2013.

PAIVA, A. O.; FARIA, G. E. Estoque de carbono do solo sob Cerrado Sensu Stricto no Distrito Federal, Brasil. **Revista Tropica:** Ciências Agrárias e Biológicas, Chapadinha, v. 1, p. 59-65, 2007.

PANDIT, S.; TSUYUKI, S.; DUBE, T. Estimating above-ground biomass in sub-tropical buffer zone community forests, Nepal, Using Sentinel 2 Data. **Remote Sensing**, Basel, v. 10, p. 601, 2018.

PECL, G. T. et al. Biodiversity redistribution under climate change: impacts on ecosystems and human well-being. **Science**, New York, v. 355, n. 6332, p. 9214, 2017.

PICARD, N.; SAINT-ANDRÉ, L.; HENRY, M. **Manual for building tree volume and biomass allometric equations:** from field measurement to prediction. Montpellier: FAO, 2012.

PONZONI, F. J. et al. Caracterização espectro-temporal de dosséis de Eucalyptus spp. mediante dados radiométricos TM/Landsat 5. **Cerne**, Lavras, v. 21, n. 2, p. 267-275, 2015.

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

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017.

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, 2018. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01431161.2018.1530808>. Acesso em: 10 dez. 2018.

ROUSE, J. et al. **Monitoring the vernal advancements and retrogradation (greenwave effect) of nature vegetation**. Greenbelt: NASA/GSFC, 1973.

SCOLFORO, H. F. et al. Spatial distribution of aboveground carbon stock of the arboreal vegetation in Brazilian biomes of Savanna, Atlantic Forest and Semi-Arid Woodland. **Plos One**, San Francisco, v. 10, n. 6, p. 1-20, 2015.

SCOLFORO, J. R. S.; CARVALHO, L. M. T. de; OLIVEIRA, A. D. de. **Zoneamento ecológico-econômico do Estado de Minas Gerais:** componentes geofísico e biótico. Lavras: Ed. UFLA, 2008. 161 p.

SEIDEL, D. et al. Review of ground-based methods to measure the distribution of biomass in forest canopies. **Annals of Forest Science**, Les Ulis, v. 68, n. 2, p. 225- 244, 2011.

SILVA, S. H. G. et al. Proximal sensing and digital terrain models applied to digital soil mapping and modeling of Brazilian Latosols (Oxisols). **Remote Sensing**, Basel, v. 8, n. 614, p. 1-22, 2016.

SILVEIRA, E. M. O. et al. Estimating aboveground biomass loss from deforestation in the savanna and semi-arid biomes of Brazil between 2007 and 2017. In: _____. **Tropical forests in transition:** the role of deforestation and impacts from community composition to regional climate change. London: IntechOpen, 2019a. Disponível em: <https://www.intechopen.com/online-first/estimating-aboveground-biomass-loss-from-deforestation-in-the-savanna-and-semi-arid-biomes-of-brazil>. Acesso em: 20 abr. 2019.

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, June 2019b.

VAUHKONEN, J. et al. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. **Remote Sensing Environment**, New York, v. 114, p. 1263-1276, 2010.

WANG, G. et al. Uncertainties of mapping aboveground forest carbon due to plot locations using national forest inventory plot and remotely sensed data. **Scandinavian Journal of Forest Research**, Stockholm, v. 26, n. 4, p. 360-373, 2011.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, London, v. 52, p. 394-403, 2015.

WILSON, E. H.; SADER, S. A. Detection of forest harvest type using multiple dates of 759 Landsat TM imagery. **Remote Sensing of Environment**, New York, v. 80, p. 385-396, 2002.

WU, C. et al. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. **Journal of Applied Remote Sensing**, Orlando, v. 10, n. 3, p. 1-18, 2016.