SEVERINO JOSÉ MACOO

ALGORITMOS EVOLUCIONÁRIOS NA PREDIÇÃO DE ESTOQUE DE CARBONO ACIMA DO SOLO EM FLORESTAS DE MOPANE - MOÇAMBIQUE

LAVRAS – MG
2023

SEVERINO JOSÉ MACOO

ALGORITMOS EVOLUCIONÁRIOS NA PREDIÇÃO DE ESTOQUE DE CARBONO ACIMA DO SOLO EM FLORESTAS DE MOPANE - MOÇAMBIQUE

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

Prof. Dr. Lucas Rezende Gomide

Orientador

LAVRAS – MG
2023

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

Macoo, Severino José.
Algoritmos evolucionários na predição de estoque de carbono acima do solo em florestas de mopane- Moçambique. / Severino José Macoo. - 2023.
80 p. : il.

Orientador(a): Lucas Rezende Gomide.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. *Genetic Algorithm and Random Forest* (GARF). 2. Programação Genética (PG). 3. Carbono Acima do Solo (AGC). I. Gomide, Lucas Rezende. II. Título.

SEVERINO JOSÉ MACOO

ALGORITMOS EVOLUCIONÁRIOS NA PREDIÇÃO DE ESTOQUE DE CARBONO ACIMA DO SOLO EM FLORESTAS DE MOPANE – MOÇAMBIQUE

EVOLUTIONARY ALGORITHMS FOR PREDICTING ABOVEGROUND CARBON STOCK IN MOPANE WOODLANDS - MOZAMBIQUE

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

APROVADA em 07 de julho de 2023.

Prof. Dr. Lucas Rezende Gomide	UFLA
Prof. Dr. Bruno Henrique Groenner Barbosa	UFLA
Prof. Dr. Luciano Cavalcante De Jesus França	UFU

Prof. Dr. Lucas Rezende Gomide

Orientador

LAVRAS – MG
2023

À minha esposa Vânia Theylla que me acompanhou em toda caminha

Ao meu filho Allan Nasson Severino Macoo

À minha Deusa (mãe), Celeste Afonso Maunge

Aos meus irmão e a toda família Macoo

DEDICO.

AGRADECIMENTOS

Em primeiro gostaria de prestar o meu agradecimento ao Instituto Superior Politécnico de Gaza (ISPG) pela concessão da Bolsa de estudos para o mestrado. À Organização das Nações Unidas para Agricultura e Alimentação (FAO), através do projeto *"Food Security and Nutrition Program (FSNP) In Gaza Province-GCP/MOZ/116/BEL*" que financiou a coleta de dados, e especialmente ao Eng. Campos Ferro. Aos professores do Programa de Pós-graduação em Engenharia Florestal da Universidade Federal de Lavras (UFLA) pelos ensinamentos, e especialmente ao LEMAF pela disponibilização de recursos para a minha formação.

Especial agradecimento vai para o meu Orientador, Prof. Lucas Rezende Gomide, pelos ensinamentos, acompanhamento minucioso e apoio incondicional na minha formação e na elaboração do presente trabalho. Ao Evandro Miranda pelo apoio no processamento de dados. Ao meu colega de classe Fabiano Rodrigues Pereira pela companhia e apoio durante as matérias. Ao meu amigo e companheiro Sérgio Alfredo Bila, pelo companheirismo, amizade e apoio durante a formação. Aos colegas do laboratório, Delano (pela revisão do trabalho), André, Pedro, Kléber e Jessy pela companhia. Aos amigos do LEMAF, Rebecca, Maria Sueliane e Bruno pelo apoio. O agradecimento estende-se também para todos os colegas de classe e amigos que direta ou indiretamente contribuíram para o alcance dos meus objetivos.

À família Macoo pelo apoio moral. Especial agradecimento vai para minha família (Vânia Theylla e Allan Nasson) pelo amor, apoio moral e paciência durante o tempo em que estive fora de casa.

Aos colegas e estudantes do ISPG que participara na coleta de dados. Aos membros das comunidades de Nwamandzele e Chihondzoene que apoiaram nas atividades de campo.

MUITO OBRIGADO!

RESUMO GERAL

As florestas tropicais desempenham papel importante na regulação do clima global e do ciclo de carbono. Mopane é uma floresta tropical seca, que ocorre na África Austral, ocupando cerca de 555000 km^2. A exploração do Mopane para carvão vegetal, em Moçambique, causa degradação e redução de estoques de carbono. Estudos de carbono neste tipo florestal podem auxiliar no monitoramento das emissões de CO_2, no âmbito do combate às mudanças climáticas, incluindo a Redução de Emissões por Desmatamento e Degradação Florestal (REDD+). No presente estudo foram testados métodos de *Machine Learning*, aplicando algoritmos evolucionários e dados de sensoriamento remoto, cobertura florestal, biofísicas e bioclimáticas para predizer estoques de Carbono Acima do Solo (AGC) na floresta de Mopane, nos distritos de Mabalane e Chicualacuala, província de Gaza, Moçambique. A amostra de campo foi composta por 114 *clusters* e foram usadas imagens de Sentinel-2, Sentinel-1, MODIS e de World.Clim para extração das variáveis. Foram testadas 139 variáveis de diferente natureza para predizer o AGC, usando (*i*) método hibrido entre Algoritmo Genético-AG para seleção de variáveis e *Random Forest* - RF para predição (GARF) e (*ii*) Programação Genética (PG) via regressão simbólica. Ambos métodos reduziram o tamanho da base de dados em 95.6%. O GARF aderiu-se mais a variáveis bioclimáticas e de sensores ópticos, enquanto a PG combinou variáveis independentemente de sua natureza e pode gerar modelos mistos e segmentados. Os valores de AGC (em $MgC.ha^{-1}$) medidos no campo variaram de 1.313 a 28.476, média = 10.988. O AGC estimado por GARF variou de 2.910 a 19.459, média = 10.235, raiz do erro quadrado médio normalizado – nRMSE = 0.427 e erro médio de viés - BEM = 0.08. Para PG variou de 1.721 a 23.503, nRMSE = 0.428 e BEM = 2.731×10^{-17}. Ambos métodos mostraram eficiência na seleção de variáveis e potencial para predição de AGC em florestas tropicais secas. A PG é mais prática em relação ao GARF, por fornecer um modelo com estrutura visível e facilmente replicável.

Palavras-chaves: Mopane. Carbono Acima do Solo. Algoritmos Evolucionários. Algoritmo Genético e *Random Forest* (GARF). Programação Genética.

GENERAL ABSTRACT

Tropical forests play an important role in regulating the global climate and the carbon cycle. Mopane is a tropical dry forest, occurring in southern Africa, occupying about 555000 km^2. Mopane harvesting for charcoal in Mozambique is a principal driver for forest degradation and carbon stocks reduction. The estimation of Carbon stocks in this type of forest can help to assess and monitoring CO_2 emissions, in the context of climate change, including Reduction of Emissions from Deforestation and Forest Degradation (REDD+). In this study, we tested Machine Learning methods by applying evolutionary algorithms and remote sensing, forest cover data, biophysical and bioclimatic data to predtic Aboveground Carbon (AGC) in the Mopane forest, in the districts of Mabalane and Chicualacuala, Gaza province, Mozambique. The sample was composed of 114 clusters and we used satellites images from Sentinel-2, Sentinel-1, MODIS and World.Clim dataset to extract the predictor variables. A set of 139 variables of different nature has been tested to predict the AGC, using (i) the hybrid method between Genetic Algorithm-AG for variable selection and Random Forest - RF for prediction (GARF) and (ii) Genetic Programming (PG) via symbolic regression. Both methods were able to reduce the database size by 95.6%. The GARF adhered more to bioclimatic variables and optical sensors, while the PG combined variables regardless of their nature and can generate mixed and segmented models. The AGC values (in MgC.ha-1) from field survey ranged from 1.313 to 28.476, mean = 10.988. The AGC estimated by GARF ranged from 2.910 to 19.459, mean = 10.235, normalized root mean square error – nRMSE = 0.427 and mean bias error - BEM = 0.08. For PG it ranged from 1.721 to 23.503, nRMSE = 0.428 and BEM = 2.731×10-17. Both methods showed efficiency for variables selection and potential for predicting AGC in tropical dry forests. The PG algorithm is more practical than GARF, as it provides a model with a visible and easily replicable structure.

Keywords: Mopane. Aboveground Carbon. Evolutionary Algorithms. Genetic Algorithm and Random Forest (GARF). Genetic programming.

LISTA DE FIGURA

CAPÍTULO 1

CAPÍTULO 2

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

SUMÁRIO

CAPÍTULO 1 (INTRODUÇÃO GERAL E REFERENCIAL TEÓRICO)

1. INTRODUÇÃO

Os efeitos das mudanças climáticas como consequência do aumento global da temperatura, devido a ação antrópica, já se fazem sentir em todo planeta e se denotam pelo aumento da frequência e intensidade dos eventos climáticos extremos nas últimas décadas, tais como, elevadas ondas de calor, precipitações elevadas, cheias e inundações, secas prolongadas, ciclones tropicais, entre outros (AERENSON, 2018; IPCC, 2021; MENG; JIA, 2018; YAMAGUCHI, et al., 2020). A redução das emissões de gases de efeito estufa, principalmente do dióxido de carbono (CO_2), constitui o principal desafio a nível mundial para reduzir os riscos e impactos das mudanças climáticas. Dentre as diversas iniciativas globais, tem-se as metas estabelecidas no Acordo de Paris, que incluem limitar o aumento da temperatura média global a 1,5°C em relação aos níveis pré-industriais (UNITED NATIONS – UN, 2015), e mais recentemente no Pacto Climático de Glasgow (2021). Este último, as metas visam a redução das emissões globais de CO_2 e de outros gases de estufa em cerca de 45% (em relação ao nível de 2010) até 2030, sendo zeradas as emissões até 2050.

Jackson et al. (2018) relatam que as emissões de CO_2 aumentaram na ordem de 1.6% a 2.7%, o equivalente a 36.2 a 37.2 Gt, entre os anos de 2017 a 2018. Este panorama apresenta tendências de aumento para níveis cada vez mais alarmantes, nos anos subsequentes, se confirmado este cenário, e com isso o limite estabelecido no Acordo de Paris poderá ser excedido até 2040 (INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE – IPCC, 2021; UNITED NATIONS ENVIRONMENT PROGRAMME (UNEP) AND INTERNATIONAL UNION FOR CONSERVATION OF NATURE (IUCN) – 2021). Como forma de contribuir para a redução da concentração de CO_2 na atmosfera, a UNEP e IUCN (2021) apontam para a potencialidade das soluções baseadas na natureza, que englobam diferentes ações como a proteção dos ecossistemas, o manejo sustentável e a restauração dos ecossistemas (naturais ou modificados) de forma absorver e armazenar o CO_2 atmosférico, proporcionando simultaneamente benefícios socioeconômicos e ambientais.

As florestas tropicais desempenham um papel fundamental na regulação do clima global e do ciclo de carbono, pois tem maior contribuição no sequestro do carbono terrestre, armazenando cerca de 45% do carbono disponível (BONAN, 2008; EPPLE et al., 2016; FISCHER, 2021; TEXEIRA. et al., 2016). Contudo, o desflorestamento, degradação e fragmentação florestal contribuem negativamente na conservação das florestas tropicais, o que resulta na redução da sua capacidade de sequestro de carbono (SONEJI; SUDARSHANA, 2012; HOLM; KUEPPERS; CHAMBERS, 2017; SONEJI; SUDARSHANA, 2012; THOMAS;

BALTZER, 2002; YESUF; BROWN; WALFOR, 2019). Assim sendo, a quantificação e o monitoramento de estoques de carbono nos remanescentes florestais é uma tarefa imprescindível, para o monitoramento das emissões de carbono, pode auxiliar na tomada de decisão sobre ações de mitigação e adaptação às mudanças climáticas (a nível local, regional e global). Além disso, serve como de guia para implementação de programas ambientais como a Redução de Emissões por Desmatamento e Degradação Florestal, mais Manejo de Florestas, Conservação e Aumento de Estoques de Carbono (REDD+), acesso ao mercado de carbono (ROZENDAAL et al, 2022), assim como ferramenta de verificação da contribuição de cada país nos esforços de mitigação e reversão das mudanças climática.

Do ponto de vista metodológico, o protocolo desenvolvido pelo IPCC é o mais empregado para a quantificação/estimativa de estoque de carbono. Segundo Chave et al. (2014), a proposta foi desenvolvida considerando uma menor rede amostral ou sub-amostragem, o que levanta incertezas na precisão das estimativas. Portanto, um dos maiores desafios consiste em encontrar uma metodologia que gera estimativas mais precisas do estoque de biomassa/carbono. Até então, o uso de modelos alométricos (envolvendo diâmetro, altura, densidade de madeira, etc.) ou uso do sensoriamento remoto (sensores ópticos e de Radar) e inventário florestal são os métodos mais usados para estimar biomassa e carbono ao nível da árvore e do povoamento. Por outro lado, a introdução de metodologias que envolvem a inteligência computacional, tais como Redes Neurais Artificiais (*Artificial neural Network* – ANN), *Random Forest* (RF), Algoritmo Genético (AG) e Programação Genética (PG) tem ganhado cada vez mais espaço (ARJASAKUSUMA; KUSUMA; PHINN; 2020; GUERRA-HERNÁNDEZ et al., 2016; MIRANDA, et al., 2022; PHAM et al, 2020; WIEGAND; PELL; COMAS, 2009).

Diante do exposto, a presente pesquisa visa implementar métodos de inteligência computacional (algoritmos evolutivos) na predição de estoques de carbono arbóreo acima do solo, na floresta de Mopane, nos distritos de Chicualacuala e Mabalane, província de Gaza, Moçambique. O estudo abrange uma área total de 62462 ha, sendo 37983 ha em Chicualacuala e 30001 ha em Mabalane. Os objetivos específicos incluem: (*i*) Aplicar métodos de aprendizagem de máquinas (Algoritmo Genético - AG, *Random Forest* - RF e Programação Genética - PG) para predição de carbono acima de solo (AGC) com base nos dados do inventário florestal, variáveis bioclimáticas e dados de sensoriamento remoto; (*ii*) Avaliar e comparar o desempenho dos métodos testados e; (*iii*) Analisar a influência das variáveis selecionadas pelos modelos na variação do AGC, relacionando-as com as condições da vegetação de cada local de estudo, incluído influências antrópicas.

2. REVISÃO DE LITERATURA

2.1. Florestas tropicais e sequestro de carbono

A fitofisionomia das florestas tropicais depende da distribuição da precipitação ao logo do ano, da altitude e do tipo de solos (THOMAS; BALTZER, 2002). Portanto destacam-se dois grandes tipos de biomas tropicais: (*i*) as florestas tropicais úmidas (florestas tropicais sempre verdes) que ocorrem ao longo do equador, entre 10º norte e sul do equador, onde não se verifica muita variação em termos de foto período, temperatura e precipitação ao longo do ano, e (*ii*) as florestas tropicais sazonais (semi-deciduais e deciduais) que ocorrem entre as latitudes 10º e 23º norte e sul do equador, em regiões onde a estação chuvosa e seca são bem distintas. As principais formações florestais na zona tropical incluem: florestas tropicais úmidas ou florestas ombrófilas, florestas semi-deciduais, florestas decíduas, savanas tropicais e mangue (HOLZMAN, 2008).

As florestas tropicais ocorrem na região tropical, entre as latitudes 23º norte e sul, e em algumas exceções que apresentam influências oceânicas ou climáticas favoráveis. Podem ser encontradas no centro e sul da América, na África ocidental, central e austral incluindo Madagáscar, Ásia-Pacifico, sudeste da Ásia, Nova Guiné, Índia, e nordeste de Austrália (HOLZMAN, 2008). Ocupam uma extensão total estimada em 2.1 milhões de hectares, o correspondente a cerca de 10% da superfície terrestre (ALLABY, 2006; HOLZMAN, 2008; NAGESWARA-RAO; SONEJI; SUDARSHANA, 2012; THOMAS; BALTZER, 2002). Apesar de ocuparem menor área em relação aos outros biomas terrestres, elas apresentam a mais alta biodiversidade (HOLZMAN, 2008) e proporcionam múltiplos bens e serviços que incluem a regulação dos ciclos biológico, serviços ambientais, provisão de produtos diversos (material de construção, medicamentos, energia doméstica) assim como dos alimentos para o Homem (THOMAS; BALTZER, 2002).

Estima-se que cerca de um terço do total da atividade metabólica da superfície terrestre concentra-se nas florestas tropicais e, portanto, elas possuem um papel fundamental na funcionalidade dos sistemas terrestres, principalmente na regulação do ciclo global do carbono (CANO *et al*. 2021; MALHI, 2012; TEXEIRA et al., 2016, THOMAS; BALTZER, 2002). Cerca de metade do carbono existente no globo terrestre (aproximadamente 45%) encontra-se armazenado em florestas tropicais (FISCHER, 2021; TEXEIRA et al., 2016), portanto são as que mais contribuem para o sequestro de carbono na superfície terrestre (BONAN, 2008; TEXEIRA et al., 2016). A complexidade das florestas tropicais resulta numa ampla variação em termos do estoque e densidade de carbono florestal (MULLER-LANDAU et al., 2020), devido a vários fatores (solos, precipitação, temperatura, altitude, estrutura florestal,

perturbação antrópica, etc.) que influenciam na dinâmica espacial e temporal de biomassa/carbono (BEHERAA et al., 2016; CASTANHO, et al.; 2020; LINDSELL e KLOP, 2013; NAVARRETE-SEGUEDA et al., 2018; RODIG et al., 2017). Estes *drivers* ambientais aumentam a variação na capacidade de sequestrar carbono. Conceitualmente, a produtividade de biomassa/carbono é maior nas florestas úmidas que nas florestas secas e em baixas altitudes (embora limitado por elevadas temperaturas) em relação a altas (MULLER-LANDAU et al., 2020), muito provavelmente devido à influência de árvores de grande copa (MEYER et al., (2018). Em geral, nas florestas húmidas de baixa altitude, a produtividade primária bruta de carbono varia entre 30 e 40 $MgC.ha^{-1}$ por ano, sendo que os solos mais férteis tendem a mostrar maior produtividade, sendo o efeito contrário em zonas mais secas ou com maior sazonalidade da precipitação (MALHI, 2012).

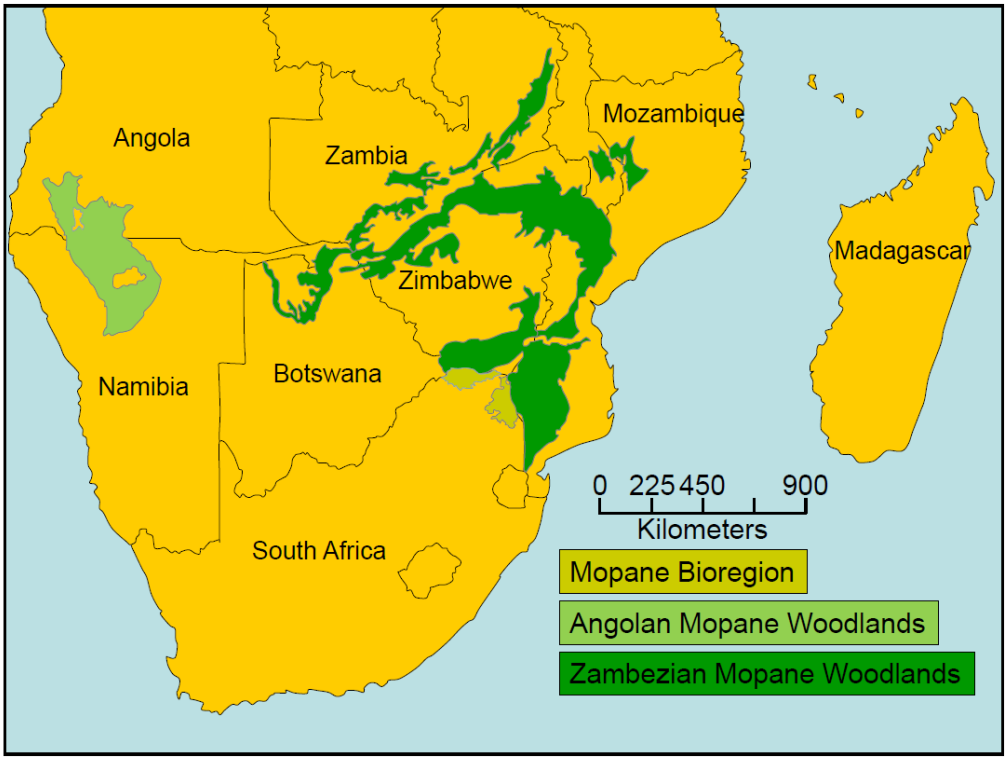
Por outro lado, a perturbação antrópica também constitui um fator preponderante no fluxo de biomassa e carbono. Hu; et al., (2016), estudando a influência da rede de estradas na biomassa florestal na China, verificaram que a concentração de biomassa e carbono reduz em função da distância, embora Lindsell e Klop, (2013) tenham notado o contrário na África ocidental, provavelmente devido à menor densidade de estradas. Não obstante, estes últimos autores observaram influência significativa da exploração florestal e da distância até aos assentamentos ao redor da floresta, provavelmente porque, no local, a população depende muito dos recursos florestais para satisfação das suas necessidades básicas (alimentos, energia, construção).

A ocorrência de incêndios florestais também leva a mudanças estruturais, a alteração da composição de espécies, a sucessão florestal e ao balanço de carbono nos ambientes. Fischer (2021) verificou uma redução da biomassa florestal pelos incêndios na ordem dos 46% a 80%, ao aplicar uma modelagem a longo prazo no monte Kilimanjaro na Tanzânia. O autor ainda relata em seu estudo que a reincidência dos incêndios exige um tempo de 150 anos para recuperar as emissões totais de carbono causadas pelo incêndio. Logo após os distúrbios, os estados sucessionais desempenham um papel importante para as relações entre a produtividade e a biomassa (RODIG et al., 2017). Ao avaliar a dinâmica de carbono florestal, em diferentes estágios sucessionais após a perturbação do desmatamento, Texeira et al. (2016) notaram que a biomassa acima do solo aumentou rapidamente durante os primeiros 20 anos, sendo mais lento após este período. O comportamento similar foi observado no estudo desenvolvido por Twery e Weiskittel (2013).

2.2. Floresta de Mopane

A floresta de Mopane é um tipo de floresta tropical seca, caracterizado pela dominância de *Colophospermum mopane* (J. Kirk ex Benth.) J. Léonard (Fabaceae), que ocorre na região da África Austral, ocupando numa extensão de cerca de 555000 km^2 e faz parte região do Centro de Endemismo do Zambeze. Encontra-se distribuída entre sul de Angola e norte de Namíbia, norte de Botswana e Zimbabwe, estendendo-se até a região central e sul de Moçambique e norte da África do Sul, sul da Zâmbia e centro de Malawi (FIGURA 1.1). Esta vegetação ocorre geralmente em regiões semiáridas, com temperaturas quentes (22 a 28 °C), baixa precipitação (200 a 800 mm/ano) e baixas altitudes (200 a 600 m) (BILA; MABJAIA, 2012; DE SOUSA et al., 2021; MAKHADO et al., 2014, 2012, NGAREGA; MASOCHA; SCHNEIDER, 2021; STEVENS, 2021). Em Moçambique a vegetação de Mopane forma uma das regiões fitogeográficas do país e cobre parte considerável das províncias de centro Tete e Manica, no centro do país, e Norte de Gaza no Sul (DE SOUSA et al., 2021, MAQUIA, et al., 2019).

Figura 1.1 – Distribuição da floresta de Mopane na África Austral e em Moçambique.



Fonte: Makhado et al. (2012).

A vegetação de Mopane tem uma importância ecológica e socioeconômica, fornecendo bens e serviços que incluem a regulação do ciclo de carbono, provisão de alimentos, material de construção, pasto, combustível lenho e medicamentos, além servir como habitat para a vida selvagem (DE SOUSA et al., 2021; MAKHADO et al., 2014; STEVENS, 2021; WOOLLEN et al., 2016). Devido ao alto poder caloríficos das espécies que compõem o Mopane, em

Moçambique tem sido amplamente explorada para a produção de carvão vegetal para alimentar as grandes cidades. Esta atividade constitui uma ameaça para conservação do ecossistema de Mopane, conduzindo ao desflorestamento e degradação florestal, consequentemente a redução dos estoques de carbono. Segundo Sedano et al. (2016), a produção de carvão em Tete, foi responsável pela liberação de 37,545 ±4,826 MgC. de carbono para atmosfera, no período de 2011 e 2014.

2.3.Estimativa de carbono

O carbono florestal encontra-se armazenado na forma de compostos orgânicos constituintes da biomassa vegetal e matéria orgânica morta (incluindo liteira) e no solo (CIAIS, et al., 2013; MAGNUSSEN; REED, 2004; MALHI, 2012). A quantidade de carbono do material vegetal está diretamente relacionada à biomassa, que é dada pela massa (Kg ou toneladas) do material vegetal seco na estufa (MAGNUSSEN; REED, 2004). A conversão de biomassa para carbono consiste na multiplicação do teor de carbono pelo valor da biomassa. Estima-se que 45 a 50% da biomassa do material vegetal corresponde ao carbono (ASSEFFA et al., 2013; SUBEDI, et al., 2010). Contudo, para o material orgânico morto ou necromassa, o teor de carbono depende do estágio de decomposição, sendo que em alguns casos há necessidade de estimar com base em outros métodos laboratoriais (MAGNUSSEN; REED, 2004). As metodologias para a estimativa de carbono vão desde os métodos que envolvem a modelagem matemática clássica, através de equações alométricos, à modelagem usando dados de sensoriamento remoto por meio de índices espectrais, imagens de hiperespectrais assim como dados LiDAR e Radar, até aos métodos contemporâneos que envolvem o uso da inteligência computacional ou combinação destes. Ambos requisitando medições no campo para o estabelecimento dessas relações funcionais.

2.4. Modelagem do carbono arbóreo acima do solo

2.4.1. Modelos alométricos nível árvore

Estimar carbono de uma floresta implica na quantificação da biomassa da componente arbórea, arbustiva, herbácea, matéria orgânica morta e do solo (incluindo raízes). A literatura geralmente subdivide em biomassa acima do solo e abaixo do solo (biomassa das raízes) (MAGALHÃES; SEIFERT, 2015; MATE; JOHANSSON; SITOE, 2014; RYAN; WILLIAMS; GRACE, 2011). A biomassa de uma árvore individual pode ser estimada pelo método direto ou pelo método indireto (CHAVE et al., 2005). O método direto ou destrutivo requere o abate, secionamento e pesagem de todas as componentes da árvore para obtenção do peso úmido e extração de amostras para secagem na estufa, para posteriormente calcular a biomassa.

Contudo, apesar de oferecer resultados mais precisos de biomassa, o método é muito oneroso e requere muito tempo. O método indireto ou não destrutivo baseia-se na relação entre a biomassa e outras variáveis dendrométricas de fácil mensuração, como o diâmetro, altura, diâmetro da copa, densidade, etc., e estima-se a biomassa através de modelos matemáticos. Este método é relativamente menos oneroso e economiza tempo. Contudo, o ajuste dos modelos requere dados provenientes do método destrutivo (CHAVE at al., 2005; MAGNUSSEN; REED, 2004). Na região tropical, as equações de Brown, Gillespie e Lugo (1989) foram amplamente usadas para estimativa de biomassa em inventários florestais. Posteriormente surgiu o trabalho de Chave at al. (2005), que foi posteriormente atualizado em Chave at al. (2014), sendo, atualmente, a equação mais usada e aceite pela comunidade científica para estimativa de biomassa e carbono no nível árvore na região pan-tropical. A mesma contou com um conjunto de dados mundiais em uma ampla rede amostral composta por 58 países, incluindo Moçambique. Em Moçambique apesar de existirem alguns modelos alométricos para estimativa de biomassa em árvores individuais, parte desses foram desenvolvidos para determinadas espécies (MATE; JOHANSSON; SITOE, 2014; MAGALHÃES; SEIFERT, 2015) e outros são direcionados para algumas fisionomias (GUEDES; SITOE; OLSSON, 2018; LISBOA et al., 2018; RYAN; WILLIAMS; GRACE, 2011; SITOE; MANDLATE; GUEDES, 2014). Por outro lado, maior parte dos estudos foi realizado na zona centro do país (TABELA 1.1).

Tabela 1.1 – Equações de biomassa em Moçambique.

(Continua)

#	Fisionomia / espécie	Equação de biomassa	R^2	Autores	Local
1	Miombo (mistura de espécies)	$log(Y_{tronco}) = -3{,}629 + 2{,}601 * log(D)$ $log(Y_{raizes}) = -3{,}370 + 2{,}262 * log(D)$ $log(Y_{total}) = -3{,}018 + 2{,}545 * log(D)$	93,0 94,0 98,0	Ryan; Williams; Grace, 2010	Gorongosa (Sofala, centro de Moçambique)
2	Chanfuta (*Afzelia quanzensis* Welw.),	$Y_{tronco} = 0{,}4369 * D^{2{,}0033}$ $Y_{ramos} = 22{,}7577 * D^{0{,}7335}$ $Y_{folhas} = 19{,}9625 * D^{-0{,}0836}$ $Y_{total} = 3{,}1256 * D^{1{,}5833}$	91,0 79,0 40,0 97,0	(Mate; Johansson; Sitoe, 2014)	Centro de Moçambique
3	Jambire (*Millettia stuhlmannii* Taub. and D.C.)	$Y_{total} = 5{,}7332 * D^{1{,}4567}$ $Y_{tronco} = 4{,}8782 * D^{1{,}4266}$ $Y_{ramos} = 0{,}3587 * D^{1{,}8091}$ $Y_{folhas} = 77{,}0114 * D^{-0{,}5511}$	95,0 94,0 78,0 72,0	(Mate; Johansson; Sitoe, 2014)	Centro de Moçambique
4	Umbila (*Pterocarpus angolensis*	$Y_{total} = 0{,}2201 * D^{2{,}1574}$ $Y_{tronco} = 0{,}0083 * D^{2{,}8923}$ $Y_{ramos} = 2{,}3596 * D^{1{,}2690}$ $Y_{folhas} = 4{,}0400 * D^{0{,}1680}$	89,0 95,0 70,0 71,0	(Mate; Johansson; Sitoe, 2014)	Centro de Moçambique

Legenda: Y - biomassa ou massa seca (Kg), D - diâmetro a altura do peito (cm), H – altura total (m), log – logaritmo na base 10, N/A – não disponível.

Fonte: Do Autor (2023).

Tabela 1.2 – Equações de biomassa em Moçambique.

(Conclusão)

#	Fisionomia / espécie	Equação de biomassa	R^2	Autores	Local
	Mangal (mistura de espécies)	$Y = 3{,}254 * \exp(0{,}065 * D)$	89,0	(Sitoe; Mandlate; Guedes, 2014)	Sofala (Centro de Moçambique)
5	*Androstachys johnsonnii* (Mecrusse)	$Y_{raizes} = 0{,}2522 + 0{,}0097 * D^2 * H$ $Y_{tronco} = 0{,}6616 + 0{,}0251 * D^2 * H$ $Y_{casca} = 0{,}1895 + 0{,}0028 * D^2 * H$ $Y_{copa} = 0{,}3033 + 0{,}0118 * D^2 * H$ $Y_{total} = 1{,}4066 + 0{,}0494 * D^2 * H$	95,0 97,5 84,4 82,3 97,6	(Magalhães; Seifert, 2015)	Gaza e Inhambane (Sul de Moçambique)
6	Floresta Sempre-verde de Montanha (mistura de espécies)	$Y = 0{,}0613 * D^{2{,}7133}$	N/A	(Lisboa et al., 2018)	Moribane (Manica)
7	Miombo de baixas altitudes (mistura de espécies)	$Y = 0.1754 * D^{2.3238}$	98,5	(Guedes; Sitoe; Olsson, 2018)	Manica e Sofala

Legenda: Y - biomassa ou massa seca (Kg), D - diâmetro a altura do peito (cm), H – altura total (m), log – logaritmo na base 10, N/A – não disponível.

Fonte: Do Autor (2023).

2.4.2. Modelos de biomassa/carbono nível povoamento

A modelagem de biomassa e carbono nível povoamento envolve o uso de variáveis de povoamento ou características inerentes a este (cobertura de copa, rugosidade, refletância espectral, etc) para estabelecer relações funcionais entre as variáveis preditoras e a variável de interesse. Geralmente emprega-se informações de biomassa/carbono ao nível de parcela ou fragmento florestal como variável dependente. As variáveis independentes podem ser de natureza variada, desde informações de solos, clima, variáveis de sensores, até informações derivadas da influência antrópica, como queimadas, regimes de manejo, entre outras.

Neste caso, o sensoriamento remoto (SR) é uma técnica amplamente usada em diversos estudos, como também para a modelagem de biomassa/carbono a nível de povoamentos florestais. É baseado na combinação entre a resposta espectral da vegetação medida pelos sensores acoplados aos satélites, aeronaves ou em veículos aéreos não tripulados (VANT´s) com dados medidos no campo em inventários florestais (CHEN et al, 2018; DANG et al., 2019; HARMSE; GERBER; VAN NIEKERK, 2022; JIANG et al., 2022). Os valores espectrais podem ser usados diretamente na modelagem ou convertido em índices espectrais (DANG et al., 2019). Para mais detalhes sobre os índices, ver Xue e Su (2017). A biomassa ou carbono medido no campo para modelagem via SR pode ser obtida pelo método direto ou através de

equações alométricas. O SR tem vantagem de permitir monitorar a variação espacial e temporal dos elementos biofísicos de uma floresta, possibilitando deste modo o monitoramento dos estoques de biomassa/carbono acima do solo no espaço e no tempo.

A fonte de dados de SR para estimativa de biomassa e carbono varia desde dados de sensores ópticos, de baixa resolução espacial (ex. Landsat, ASTER e MODIS), média resolução (ex. série SENTINEL) e imagens de alta resolução (ex. Quickbird, IKONOS, WorldView e GeoEye) até aos dados de Radar (*Radio Detection and Rangin*) e LiDAR (*Light Detection and Ranging*). Os dados de Radar e LiDAR têm sido largamente usados na modelagem de biomassa/carbono acima de solo em florestas, estes sensores ultrapassam a limitação dos sensores ópticos em fornecer uma referência vertical do dossel das árvores (CHEN et al, 2018; FATOYIMBO et al., 2018; JIANG et al., 2022; TARAVAT; WAGNER; OPPELT, 2019; ZAKI; LATIF, 2017).

A série de satélites Sentinel, pertencente a Agência Espacial Europeia (ESA)/Comissão Europeia no âmbito do Programa Copernicus, foi lançada em 2014 para monitoramento dos recursos naturais terrestres, uso e cobertura de terra, ambientes marinhos, clima e desastres naturais (DANG et al., 2019; TORRES et al., 2012). A série 1 foi destinada ao monitoramento terrestre e oceânico, série 2 a vegetação, solos e áreas costeiras, série 3 ao mar (com sensores Radar específicos) e as séries 4 e 5 a qualidade do ar. O Sentinel-1 é equipado com sensores de Radar (*Sinthetic Aperture Range* - SAR), que opera na banda C (entre 3,8 – 7,5 cm) com polarização VV, VH, HH e HV (ØSTERGAARD et al., 2011; TORRES et al., 2012). O Sentinel-2 é equipado pelo sensor multiespectral MSI, com 13 bandas espectrais, variando de 443 a 2190 nm, com resolução espacial de 10m para as bandas do visível, 20m para o infravermelho e 60m para as bandas de correção atmosférica (CHEN et al., 2018; HARMSE; GERBER; VAN NIEKERK, 2022). Mais detalhes sobre estes produtos encontram-se no *site* oficial da ESA. Atualmente, estão disponíveis gratuitamente coleções de imagens Sentinel referentes a diferentes períodos, possibilitando o seu uso para análise de séries temporais, incluindo na plataforma *Google Earth Engine* (GEE).

As imagens do Sentinel 1 e 2 já foram usadas para várias aplicações, tais como monitoramento de culturas agrícolas e pastagens (TARAVAT; WAGNER; OPPELT, 2019, HARMSE; GERBER; VAN NIEKERK, 2022), uso e cobertura de terra e modelagem biofísica em florestas incluindo biomassa/carbono (CHEN et al., 2018, DANG et al., 2019; QIAN, et al., 2021; MACAVE et al., 2022; SINGH et al., 2022), monitoramento desastres naturais (TARPANELLI; MONDINI; CAMICI, 2022), entre outras aplicações.

Em Moçambique há um número reduzido de estudos de biomassa e carbono florestal, envolvendo o uso de SR. Macave et al. (2022) combinando dados de sensores ópticos (Landsat 8/OLI e Sentinel 2A/MSI) e Radar (Sentinel -1B e ALOS/PALSAR-2) modelou a biomassa florestal acima do solo na vegetação de Miombo na Reserva de Niassa, onde a biomassa média estimada foi de 56 $Mg.ha^{-1}$ (R^2 = 87.5% e RMSE de 11.56 $Mg.ha^{-1}$). Gou, Ryan e Reiche (2022) melhoraram a predição de biomassa acima de solo na floresta de Miombo e Mopane no centro de Moçambique, com a inclusão do modelo semi-empírico WCM (*water cloud model*) para contabilizar a retro difusão da umidade do solo, usando dados de Radar (ALOS PALSAR-1). Os autores verificaram que as estimativas da média de biomassa reduziram em 18.6%. Outros estudos de biomassa e carbono envolvendo o uso de SR foram realizados em diferentes tipos de vegetação, como o exemplo de Carreiras, Melo e Vasconcelos (2013) e Ribeiro et al. (2008), na vegetação de Miombo, e Fatoyimbo et al. (2008, 2018) e Mitchard et al. (2009) no manguezal.

2.5. Avanços na modelagem de atributos florestais

A inteligência artificial ou computacional está cada vez mais ganhando espaço na área da modelagem, devido a sua capacidade de lidar com problemas complexos. Logo, detecta-se na literatura uma expansão de trabalhos desta natureza, baseados na inteligência computacional. Os algoritmos de aprendizagem de máquinas (*Machine Learnning - ML*) mais usados incluem: Redes Neurais Artificiais (*Artificial Neural Networks – ANN*), Máquinas de Suporte de Vetores (*Support Vector Machines – SVM*), Mapa Auto-Organizado (*Self-Organizing Map – SOM*), Árvores de Decisão (*Decision Trees* – DT), Florestas Aleatórias (*Random Forests – RF*), Algoritmo Genético (AG), Programação Genética (PG), *Splines* de Regressão Adaptativa Multivariada (*Multivariate Adaptive Regression Splines* - MARS), entre outros algoritmos (LARY et al., 2016).

2.5.1. *Random Forest*

Radom Forest (RF) ou Floresta Aleatória é um algoritmo de aprendizagem de máquinas baseado em um conjunto de Árvores de Decisão, desenvolvido por Breiman, (2001), podendo ser usado para classificação e para regressão. No algoritmo RF, as Árvores de Decisão são construída a partir de amostras *bootstrap*, usando o algoritmo CART (*Classification and Regression Trees*) e a predição final é obtida agregando todo o conjunto, através do voto maioritário, para classificação, onde a classe mais votada pelas árvores é selecionada, ou pela média das saídas de todas as árvores, para regressão (BIAU, 2012; SVETNIK et al., 2003). Este processo é designado por *bootstrap aggregating* ou simplesmente Bagging. *Bootstrap* é um

método de reamostragem que usa amostragem aleatória com reposição, e é usado para quantificar as incertezas associadas a um determinado estimador. O Bagging resulta numa floresta não correlacionada de árvores, cuja predição feita pelo conjunto é mais precisa do que a de qualquer árvore individual (BIAU, 2012; BREIMAN, 2001; LIAW; WIENER, 2002; MARTINS SILVA et al., 2019).

Ao contrário do que acontece na maioria dos algoritmos, onde cada nó é dividido usando a melhor divisão entre todas as variáveis preditoras, no algoritmo RF, Breiman (2001) introduziu uma pequena modificação, que consiste em dividir cada nó usando o melhor subconjunto de variáveis preditoras, entre várias selecionadas aleatoriamente nesse nó. Esse processo resulta na melhoria do desempenho em relação aos outros classificadores mais populares, incluindo SVM e ANN, sendo menos susceptível a *overfitting* (LIAW; WIENER, 2002).

O RF é um algoritmo muito simples de implementar, embora a sua formulação matemática continua um mistério (BIAU, 2012), pois apenas três parâmetros é que necessitam de ser ajustados (*tuning*), para fornecer maior precisão: (*i*) o número de variáveis no subconjunto aleatório em cada nó (m_{try}); (*ii*) o número de árvores na floresta (n_{tree}) e (*iii*) tamanho da árvore, medido pelo menor tamanho de nó para divisão ou o número máximo de nós terminais (CUTLER; CUTLER; STEVENS, 2012). Porém, segundo Liaw e Wiener (2002) o algoritmo RF geralmente não é muito sensível aos valores desses parâmetros, exceto m_{try} ao qual é ligeiramente sensível, sedo que na classificação, o padrão é $m_{try} = \sqrt{M}$ e na regressão $m_{try} = N/3$, onde M é o número total de preditores e N é o tamanho da amostra.

A implementação do algoritmo RF segue os seguintes passos: (*i*) gerar amostras *bootstrap* com tamanho n_{tree} a partir dos dados de treinamento; (*ii*) para cada uma das amostras *bootstrap*, construir uma árvore de decisão não podada, onde em cada nó, gera-se uma amostra aleatória de m_{try} variáveis preditoras e selecionar a melhor divisão entre elas e (*iii*) estimar novos dados agregando as saídas de todas as árvores geradas, através da classe mais votada (classificação) ou pela média (regressão). De madeira geral, o desempenho de um algoritmo de predição é avaliado usando um conjunto de dados de teste independente, que não foi usado no treinamento, ou através de uma validação cruzada, quando os dados são limitados. Portanto, segundo Svetnik et al. (2003) o RF efetua uma validação cruzada em paralelo com a etapa de treinamento, com base nas chamadas amostras fora da mochila (*Out-Of-Bag*, ou simplesmente OOB). Considerando que no processo de treinamento, cada árvore é construída usando uma amostra *bootstrap* específica, onde uma parte dos dados de treinamento não é incluída na amostra, enquanto outra parte é repetida, os dados deixados de fora constituem a amostra OOB

(geralmente cerca de um terço das vezes uma amostra está fora da mochila). Assim sendo, como as amostras OOB não foram usadas na construção da árvore, podem ser usadas para avaliar o desempenho do algoritmo, seguindo os seguintes passos: (*i*) a cada iteração *bootstrap*, estimar os dados que não estão na amostra de *bootstrap* (amostra OOB) usando a árvore construída com a amostra de *bootstrap*; (*ii*) calcular uma estimativa da taxa de erro (ER_{OOB}) para classificação ou o erro quadrático médio (MSE_{OOB}) para regressão, agregando as estimativas OOB feitas no passo anterior (CUTLER; CUTLER; STEVENS, 2012; LIAW; WIENER, 2002, SVETNIK et al., 2003) através das fórmulas 1 e 2.

$$ER_{OOB} = N^{-1}\sum_{i=1}^{N} I\left(y_i \neq \hat{f}_{oob}(x_i) \right) \quad (1)$$

$$MSE_{OOB} = N^{-1}\sum_{i=1}^{N}\left(y_i - \hat{f}_{oob}(x_i) \right)^2 \quad (2)$$

Em que: $f_{oob}(x_i)$ é a predição OOB para a observação i; y_i é o valor observado da variável de interesse e; N – tamanho da amostra OOB e I é a função indicadora.

Quanto ao desempenho computacional, o RF tem se destacado em relação a outros algoritmos de aprendizagem de máquinas pela capacidade resolver problemas de diferentes naturezas e de alta dimensão, velocidade computacional, necessidade de poucos parâmetros de ajuste, possibilidade de generalização do erro e facilidade de implementação em paralelo. Do ponto de vista estatístico, o RF permite medir a importância de variáveis (através da permutação de variáveis), ponderação das classes quando forem dados desbalanceados, visualização, detecção de *outliers* e aceita uma aprendizagem não supervisionada (CUTLER; CUTLER; STEVENS, 2012; PHAN; KUCH; LEHNERT, 2020).

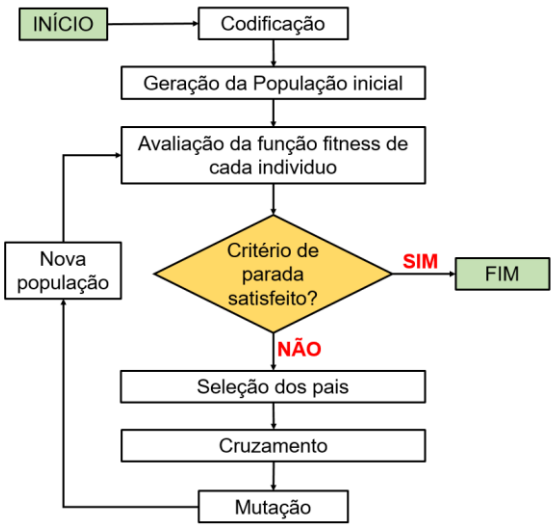
2.5.2. Algoritmo genético

O Algoritmo Genético (AG) é uma técnica de computação evolucionária que busca soluções ótimas para problemas complexos, imitando a teoria de evolução de Darwin (DEB, 2001; HAUPT; HAUPT, 2004; PHAM et al., 2020; WIEGAND; PELL; COMAS, 2009). O método foi desenvolvido por John Holland (1975) e a sua aplicabilidade na optimização foi demostrada por De Jong (1975), um dos estudantes de Holland. Em 1985, David Goldberg (outro estudante de Holland), destacou-se quando conseguiu aplicar o AG para resolver um problema complicado envolvendo controle de *pipeline* de gás e, mais tarde (em 1989), o método popularizou com a publicação do seu livro intitulado: *Genetic Algorithms in Search, Optimization, and Machine Learning* (DEB, 2001; GOLDBERG, 1989; HAUPT; HAUPT,

2004; MITCHELL, 1996; REEVES, 2003; ROTHLAUF, 2006). O AG é bastante aplicado para resolução de problemas de optimização e em outras aplicações (REEVES, 2003).

O AG baseia-se em princípios genéticos e de seleção natural na busca de uma solução ótima, partindo de uma população aleatória (também designada de solução candidata), aplicando operadores genéticos de seleção, cruzamento (*crossover*) e mutação, iterativamente substituindo uma população por outra em cada iteração, até alcançar a convergência (ARJASAKUSUMA; KUSUMA; PHINN, 2020; COLEY, 1999; GOLDBERG, 1989; MITCHELL, 1996; REEVES, 2003; ROTHLAUF, 2006). A composição de um AG apresenta os seguintes elementos: (*i*) codificação, (*ii*) geração da população inicial, (*iii*) avaliação da função fitness, (*iv*) aplicação dos operadores de seleção, crossover e mutação para geração de descendentes (nova população) e (*v*) verificação do critério de parada (FIGURA 1.2).

Figura 1.2 – Esquema básico de funcionamento de um algoritmo genético simples.



Fonte: Adaptado de Gomide (2009) e Miranda (2020).

Para cada situação, existem várias decisões a serem tomadas relativas aos aspectos a considerar ao aplicar um AG, dentre eles, Coley (1999) destaca os seguintes: (*i*) método de codificação dos parâmetros, (*ii*) tipo de crossover, (*iii*) o tamanho da população, (*iv*) a aplicação do conceito de mutação e representação e (*v*) a definição do critério parada. Dependendo da variável em questão, a codificação dos parâmetros pode ser feita usando variáveis binárias, contínuas, inteiras ou caracteres (GOMIDE, 2009; MITCHELL, 1996; HAUPT; HAUPT, 2004). Deb (2001) descreve com detalhe a função de mapeamento das variáveis inteiras ou contínuas em variáveis binárias. Em todos os casos, para codificação correta do problema é necessário entender alguns conceitos genéticos básicos no contexto do AG (genótipo, fenótipo, gene, cromossoma, gene e alelo). O genótipo é a representação codificada das variáveis, isto

é, representa toda informação armazenada nos cromossomas. O fenótipo é a aparência externa de um determinado indivíduo e corresponde ao conjunto das variáveis que compõem a solução. O fenótipo determina o sucesso de um indivíduo, portanto, a comparação entre os indivíduos é feita ao nível do fenótipo (avaliação do *fitness*), embora os descendentes não herdam as propriedades fenotípicas, mas sim as propriedades genotípicas, razão pela qual, os operadores genéticos atuam ao nível do genótipo. O cromossoma armazena toda informação do indivíduo, sendo composta por uma sequência de genes (FIGURA 1.3). A maioria dos AG´s usa apenas um cromossoma para formar um indivíduo, sendo que cada indivíduo representa uma solução candidata. O gene é a região do cromossoma que contém informação responsável por uma determinada característica fenotípica, ou seja, codifica um elemento particular da solução, e pode ser composta por um ou mais alelos. Os alelos são unidades mínimas de informação dentro de um cromossoma e são representados por um símbolo (números, letras, etc.), na codificação binária corresponde ao valor numérico de cada bit podendo ser 0 ou 1. A posição de cada alelo no cromossoma chama-se *locus* (GOMIDE, 2009; LI et al., 2021; MITCHELL, 1996, ROTHLAUF, 2006).

Figura 1.3– Exemplo básico de codificação de um cromossoma com 3 genes, cada composto por 4 alelos, na codificação binária.

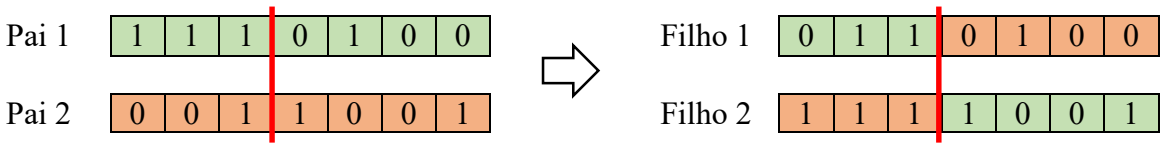
Fonte: Adaptado de Gomide (2009).

O conjunto de cromossomas é denominado de população, a recombinação dos cromossomas é feita pelos operadores genéticos de crossover e mutação. A população inicial é selecionada aleatoriamente. A busca de solução ótima é guiada pela função *fitness (f)* em cada cromossoma da população. O *fitness* de cada cromossoma depende da sua capacidade de resolver o problema em questão. Com base no valor do *fitness*, indivíduos com melhor solução são identificados e dados oportunidade para reproduzir (MITCHELL, 1996, REEVES, 2003). Não existe uma regra clara para definição do tamanho da população, contudo Reeves (2003) afirma que um tamanho de população menor restringe o espaço de busca comprometendo a eficiência na busca de solução ótima, por outro lado um tamanho de população muito grande estende o tempo de busca da solução resultando num tempo de processamento demasiadamente longo.

O operador de seleção exerce pressão sobre a população, imitando a seleção natural (COLEY, 1999) e seleciona com mais frequência indivíduos mais aptos para reprodução, favorecendo-os em relação aos menos aptos, desta forma, indivíduos altamente aptos tem maior probabilidade de criar descendentes do que indivíduos inferiores (GOLDBERG, 1989), desta feita, após algumas gerações os indivíduos inferiores são eliminados da população e como consequência a aptidão média de uma população aumenta ao longo das gerações (ROTHLAUF, 2006). Existem diferentes formas de implementação da seleção descritas na literatura e segundo Gomide (2009) os métodos mais conhecidos são: elitista, proporção, roleta (*roulette wheel*), torneio (*tourament*), duplo torneio (*double tounament*), *stochastic universal sampling*, truncamento, boltzman, *sigma scaling, sigma sharing, sigma scaling* truncada, normalizada, *ranking*, relacionado a diversidade, bi-classista, aleatória salvacionista e aleatório não salvacionista.

O *crossover* permite a troca de informação genética entre os indivíduos (cromossomas) em analogia a reprodução sexual, para gerar descendentes com características presumidamente superiores aos seus progenitores, que garantem a sua sobrevivência (COLEY, 1999, LI et al., 2021, MITCHELL, 1996, REEVES, 2003, ROTHLAUF, 2006). No entanto, segundo Miranda (2020) este operador é considerado o mecanismo de evolução mais importante do algoritmo, pois ao produzir novos cromossomas possibilita-o a escapar de ótimos locais. Diferentes formas de implementação do operador de *crossover* podem ser encontradas na literatura, a destacar os seguintes: uniforme, segmentado, único ponto e pontos múltiplos, sendo os dois últimos os mais usados (ROTHLAUF, 2006). A Figura 1.4 ilustra um esquema de funcionamento de crossover de um ponto.

Figura 1.4 – Exemplo funcionamento de crossover de um ponto (A linha vermelha indica o ponto de corte).

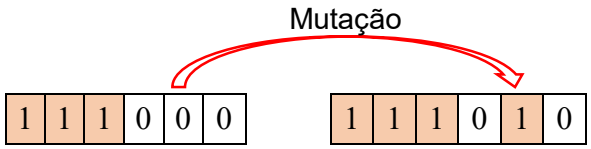


Fonte: Adaptado de Gomide (2009).

A mutação consiste na troca aleatória da informação genética contida em determinado lócus, alterando ligeiramente o genótipo e produzindo novos indivíduos (COLEY, 1999, MITCHELL, 1996), desta forma, o operador de mutação contribui para manutenção da diversidade genética e amplia o espaço de busca, ajudando o algoritmo a escapar de ótimos locais (DEB, 2001, LI et al., 2021) e permite a recuperação do material genético eventualmente perdido durante a seleção e cruzamento (GOLDBERG, 1989). A probabilidade de ocorrência

da mutação é baixa (menor que 1%) por isso é considerado um evento secundário no AG (GOLDBERG, 1989). Por outro lado, elevadas taxas de mutação no AG podem alterar tanto o genótipo dos indivíduos, que resulta numa geração muito diferente dos seus progenitores (ROTHLAUF, 2006). A Figura 1.5 ilustra o processo de mutação num cromossoma codificado de forma binária, alterando o valor do bit.

Figura 1.5– Exemplo de mutação de mutação através da mudança do valor do bit de 0 para 1.



Fonte: Adaptado de Gomide (2009).

Após a aplicação dos operadores genéticos a nova população passa pela avaliação da função *fitness*, que consiste na transformação dos bits de um determinado cromossoma (x) em um valor numérico real (y), que corresponde ao valor função nesse determinado ponto (MITCHELL, 1996). A metodologia detalhada sobre a operação de transformação foi descrita por Haupt e Haupt, 2004. Os processos de seleção, *crossover* e mutação são repetidos até satisfazer o critério de parada. A escolha do critério de parada, embora seja complicada, depende da natureza do problema e segundo Gomide (2009) deve ser feita de tal forma que maximize o algoritmo. Os critérios de parada mais usuais são: (*i*) número máximo de iterações, (*ii*) número máximo de iterações sem melhorias na função fitness ou (*iii*) tempo de processamento (ARJASAKUSUMA; KUSUMA; PHINN, 2020). O AG, em relação aos outros algoritmos, tem como vantagens: (*i*) capacidade de lidar com um elevado número de combinações variáveis; (*ii*) permite escapar de soluções locais; (*iii*) não necessita de um ponto de partida, dado que o processo é iterativo; (*iv*) lida com diferentes tipos de problemas envolvendo variáveis tanto binárias assim como contínuas. A maior desvantagem é que não garante a solução ótima (LI et al.; 2021; MCROBERTS et al., 2016).

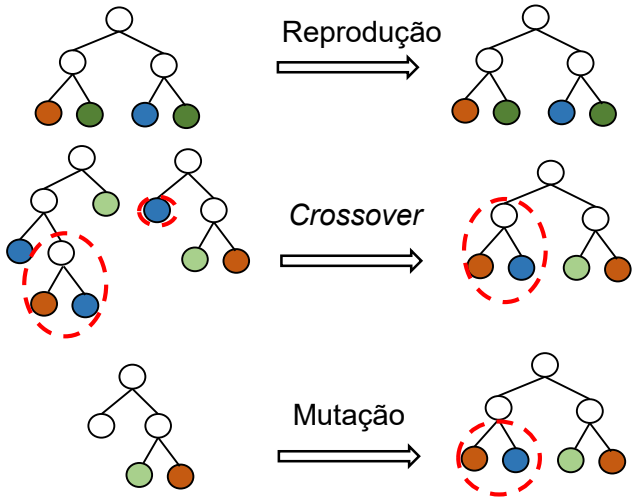
2.5.3. Programação genética

A Programação Genética (PG) é um algoritmo evolucionário e probabilístico de aprendizagem de máquinas, baseado na teoria darwiniana de seleção natural, inventado por Koza (1992), para resolução de problemas complexos. A PG é praticamente uma evolução do Algoritmo Genético na qual a população consiste em programas de computador, compostos por um conjunto de funções primitivas e terminais previamente fornecidos em função do problema (KOZA, 2003, POLI; LANGDON; MCPHEE, 2008; TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016). As funções podem ser formadas por operações aritméticas

padrão, operações de programação padrão, funções matemáticas padrão, funções lógicas ou funções específicas de domínio. Os programas de computador podem assumir valores booleanos, inteiros, reais, complexos, vetoriais, simbólicos ou múltiplos, e a sua arquitetura varia em tamanho e formato (KOZA, 2003, POLI; LANGDON; MCPHEE, 2008).

O algoritmo PG inicia com uma população inicial de programas de computador gerados através de uma busca aleatória e cega, usando um conjunto de funções primitivas e terminais previamente fornecidos em função do problema. Cada indivíduo (programa de computador) na população constitui uma solução candidata ao problema. A população evolui automaticamente ao longo de gerações, mediante o princípio de sobrevivência do mais apto e aplicação de operadores de reprodução, *crossover* e mutação (FIGURA 1.6). Para uma descrição detalhada sobre o funcionamento dos operadores refere-se a Poli, Langdon e Mcphee (2008) e Koza (1992). Assim como no AG, em cada geração, a população é submetida a uma avaliação do *fitness*, os indivíduos mais aptos são selecionados para formar a nova geração (seleção dos indivíduos baseada no *fitness*), uma fração destes indivíduos (cerca de 10%) é reproduzida (cópia dos pais), maior parte (cerca de 90%) é selecionada para o cruzamento (*crossover*) para formar novos indivíduos diferentes dos pais e presumivelmente superiores, e uma ínfima parte (cerca de 1%) sofre mutação gerando também indivíduos diferentes. As operações de avaliação, seleção, reprodução, *crossover* e mutação são repetidas iterativamente de geração em geração até alcançar o critério de parada. O resultado final é definido a partir do indivíduo ou programa mais adaptado, contudo, este pode ser uma aproximação ou uma solução ótima para o problema (CABRAL et al., 2018; KOZA, 2003, 1992; LONDHE, et al., 2022; MEHR e KAHYA, 2017; MOGHADDAM, et al., 2021; POLI; LANGDON; MCPHEE, 2008; TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016).

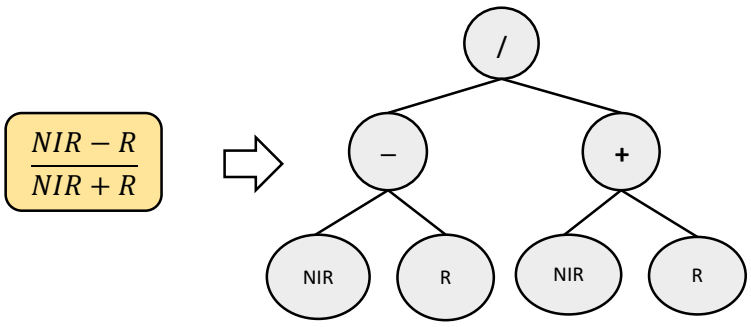
Figura 1.6– Exemplo de funcionamento dos operadores genéticos na PG.



Fonte: Adaptado de Wen et al. (2022).

Existem inúmeras aplicações da PG e a terminologia usada em função da natureza do problema e do produto gerado (KOZA, 1992). Portanto, quando o objetivo é desenvolver um modelo empírico, que forneça um bom ajuste, a uma determinada amostra de dados adquiridos de um processo ou sistema, a PG é designada como Regressão Simbólica (KOZA, 1992; MEHR e KAHYA, 2017). Diferentemente da regressão linear e não linear, onde o algoritmo busca os coeficientes numéricos para um modelo especifico e já conhecido, na PG via regressão simbólica procura-se automaticamente construir o modelo e buscar os coeficientes numéricos que se ajustam ao modelo construído, isto é, não é necessário que a estrutura da solução seja conhecida. Neste caso, a regressão simbólica procura selecionar variáveis, operadores matemáticos e booleanos, funções matemáticas assim como estimativa dos parâmetros para o modelo construído (CABRAL et al., 2018; KOZA, 1992; TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016). Na regressão simbólica os indivíduos são representados em forma de árvores de análise, criadas a partir da combinação de funções e terminais escolhidos em função da natureza e complexidade do problema (FIGURA 1.7).

Figura 1.7 – Esquema de uma árvore de análise representando a expressão de NDVI na PG.



Fonte: Adaptado de Batista et al. (2021).

A estrutura e arquitetura das árvores dependem do conjunto dos elementos iniciais (funções e terminais) usados para sua construção (CABRAL et al., 2018) e consiste em um nó em cada raiz, ramos, que são derivados de cada função e terminam em um terminal ou folha (BAYAT et al, 2019; MEHR e KAHYA, 2017). Especial atenção deve ser prestada na criação dos indivíduos tanto da população inicial, como da descendência pelos operadores genéticos (*crossover* e mutação), de modo a garantir que os indivíduos criados (ex. programas, expressões) sejam executáveis e sintaticamente válidos. Detalhes sobre os procedimentos específicos incluindo as restrições na estrutura e arquitetura dos programas encontram-se descritos na obra de Poli, Langdon e Mcphee (2008).

Resumidamente, uma modelagem GP padrão necessita de 4 principais entradas: (*i*) dados de treino e validação; (*ii*) uma função *fitness* (ex. erro quadrático médio, coeficiente de determinação) para avaliação e seleção dos indivíduos, geralmente por torneio; (*iii*) conjuntos funcionais (nós internos) e terminais (folhas) para identificação estrutural; e (*iv*) parâmetros GP para formação de uma árvore sintática (um programa/solução potencial), que podem ser operadores aritméticos básicos (adição, subtração, multiplicação, divisão), operadores boleanos ou funções matemáticas mais complexas (trigonométricas, logarítmicas, exponencial), em alguns casos inclui funções automaticamente definidas (ADF), entre outros (MEHR e KAHYA, 2017; POLI; LANGDON; MCPHEE, 2008).

Um aspecto importante a não ignorar está relacionado com a dimensão e complexidade das soluções criadas (tamanho e profundidade da árvore) pela PG, assim como à precisão dos resultados. Segundo Koza (20003), na PG é difícil especificar ou restringir o tamanho e a forma da solução final antecipadamente, sob risco de limitar o algoritmo a convergir em boas soluções. Não obstante, Poli, Langdon e Mcphee (2008) apresentam técnicas e procedimentos que podem ser usados para controlar o crescimento das árvores ao mesmo tempo que se minimiza o viés (*bias*), através de restrições de tamanho, profundidade e nas operações genéticas (*crossover* e mutação). Teoricamente, a PG pode resolver qualquer problema cujas soluções candidatas possam ser avaliadas e comparadas (CABRAL et al., 2018). A melhor solução do algoritmo pode ser escolhida de diferentes formas, dependendo da experiência do modelador. Por exemplo, para PG multiobjectivo, pode-se escolher a melhor solução usando a fronteira de Pareto (KOZA, 2003).

Devido à sua capacidade de resolver problemas de diferentes naturezas e complexos, a PG tem sido amplamente usada em diferentes áreas de conhecimento científico, como por exemplo na hidrologia (MEHR e KAHYA, 2017), engenharia civil (LONDHE et al., 2022), mineração de dados (LENSEN; XUE; ZHANG, 2021; ZHOU et al., 2023), manejo florestal (BAYAT et al., 2019; GHOSH; BEHERA; PARAMANIK, 2020) e sensoriamento remoto (BATISTA et al., 2021). Cabral et al. (2018) usaram a PG para estimar áreas afetadas por incêndios, usando sensoriamento remoto em florestas naturais no Brasil, Guiné Bissau e República Democrática do Congo, onde a metodologia provou ser prática comparando com as metodologias clássicas. BAYAT et al. (2019) recorreram a PG para estudar os fatores biofísicos que controlam o crescimento em diâmetro de *Fagus orientalis*, no Norte de Irã. A PG via regressão simbólica mostrou melhor desempenho em relação ao algoritmo RF, na estimativa da altura da copa do manguezal na Índia, usando imagens satélites SENTINEL (GHOSH; BEHERA; PARAMANIK, 2020). Em problemas de classificação, a PG mostrou melhor

desempenho em relação a outros algoritmos mais conceituados como *K-Nearest Neighbor* (KNN), *Decision Tree* (DT), Naïve Bayes (NB), *Support Vector Machine* (SVM) e *Random Forest* (RF) no mapeamento de áreas agrícolas, usando imagens de alta resolução (WEN, C. et al., 2011). Batista et al. (2021) provaram a aplicabilidade de PG usando o algoritmo M3GP na construção de variáveis para classificação de uso e cobertura de terra em Angola, Brasil, República Democrática de Congo, Guiné Bissau e Moçambique, onde verificaram que as variáveis multiespectrais construídas pela PG foram superiores aos índices tradicionalmente conhecidos como NDVI, NDWI e NBR na explicação dos padrões de uso de terra.

3. CONSIDERAÇÕES FINAIS

A questão das mudanças climáticas é preocupante a nível mundial, no entanto, os esforços estão sendo envidados com vista a mitigação e reversão do atual cenário. As convenções mundiais sobre o clima como o Protocolo de Kyoto, Acordo de Paris e o Pacto Climático de Glasgow mostraram avanços no sentido de luta contra as mudanças climáticas, com maior enfoque para a redução das emissões de gases de efeito estufa, principalmente o CO_2, através de iniciativas como a REDD+. Neste contexto, os países são chamados a mostrar suas contribuições, quantificando os estoques de carbono nos remanescentes florestais. Por outro lado, as florestas tropicais desempenham um papel importante na regulação do clima global e do ciclo de carbono, pois são os maiores reservatórios do carbono terrestre. As metodologias usadas para estimativa dos estoques de carbono incluem o uso de equações alométricas, sensoriamento remoto combinado com inventário florestal através da modelagem clássica. A inclusão da variável ambiental que considera o *stress* na modelagem clássica foi um avanço na redução das incertezas associadas a estimativas de biomassa e carbono (CHAVE, et al., 2014). A modelagem usando sensoriamento remoto é a forma mais prática e barata de monitorar os estoques de carbono, espacialmente e temporalmente. A evolução computacional traz consigo uma tendência de migração na abordagem da modelagem biofísica, com o uso da inteligência computacional na mineração de dados e modelagem, ultrapassando as limitações da modelagem clássica, relacionada com a escolha de variáveis, distribuição estatísticas dos dados, multicolinearidade, complexidade dos modelos, robustez e precisão das estimativas. Assim sendo, a inteligência computacional constitui uma potencialidade para a modelagem de carbono a nível local, regional e global. A combinação entre o sensoriamento remoto e algoritmos de inteligência computacional é potencial para o futuro da modelagem biofísica, principalmente em florestas topicais que são caracterizadas por uma alta complexidade.

REFERÊNCIAS

AERENSON, T. et al. Changes in a suite of indicators of extreme temperature and precipitation under 1.5 and 2 degrees warming. **Environmental Research Letter.** v. 13, n. 3, p. 1-20, 2018.

ALLABY, M. **Tropical Forests.** In: Biome of the Earth. Chelsea: Chelsea House. 272p, 2006.

ARJASAKUSUMA, S.; KUSUMA, S. S.; PHINN, S. Evaluating Variable Selection and Machine Learning Algorithms for Estimating Forest Heights by Combining Lidar and Hyperspectral Data. **International Journal o Gero-Information**. v. 9, n. 9, 2020.

BATISTA, J.E.; et al. Improving Land Cover Classification Using Genetic Programming for Feature Construction. **Remote Sensing**. v. 13, 2021.

BAYAT, M. et al. "A Semi-empirical Approach Based on Genetic Programming for the Study of Biophysical Controls on Diameter-Growth of *Fagus orientalis* in Northern Iran. **Remote Sensing**. v. 11, n. 14, 2019.

BEHERAA S. K., et al. Aboveground biomass and carbon stock assessment in Indian tropical deciduous forest and relationship with stand structural attributes. **Ecological Engineering**. v. 99, p. 513–524, 2016.

BIAU, G. Analysis of a Random Forests Model. **Journal of Machine Learning Research**. v. 13, p. 1063 – 1095, 2012.

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

BREIMAN, L. Random Forests. **Machine Learning**. v. 45, p. 5 – 31, 2001.

BRONISZ, K.; MEHTÄTALO, L. Seemingly Unrelated Mixed Efects Biomass Models for Young Silver Birch Stands on Post-Agricultural Lands. **Forests**. v. 11, n. 4, p. 381, 2020.

BROWN, S.; GILLESPIE, A. J. R.; LUGO, A. E. Biomasss estimation methods for tropical forests with applications to forest inventory data. **Forest Ecology**, v.35, n. 4, p.881-902, 1989.

CABRAL, A. I. R. et al. Burned area estimations derived from Landsat ETM+ and OLI data: Comparing Genetic Programming with Maximum Likelihood and Classification and Regression Trees. **ISPRS Journal of Photogrammetry and Remote Sensing.** V. 142, p. 94-105, 2018.

CANO, I. M. at al. Allometric constraints and competition enable the simulation of size structure and carbon fluxes in a dynamic vegetation model of tropical forests (LM3PPA-TV). **Global change biology**. v. 26, n.8, p. 4478-4494, 2020.

CARREIRAS, J.M.B.; MELO, J.B.; VASCONCELOS, M.J. Estimating the Above-Ground Biomass in Miombo Savanna Woodlands (Mozambique, East Africa) Using L-Band Synthetic Aperture Radar Data. **Remote Sensing**. v. 5, p. 1524-1548, 2013.

CASTANHO, A. D. A. et al. Potential shifts in the aboveground biomass and physiognomy of a seasonally dry tropical forest in a changing climate. **Environmental Research Letters**. v. 15, n. 03, 2020.

CHAVE, J. et al. Improved allometric models to estimate the aboveground biomass of tropical trees. **Global Change Biology. v.** 20 (10), p. 3177-3190, 2014.

CHAVE, J. et al. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. **Oecologia**. v. 145, p. 87–99, 2005.

CHEN, L.; et al. Estimation of Forest Above-Ground Biomass by Geographically Weighted Regression and Machine Learning with Sentinel Imagery. **Forests**. v. 9, p. 582, 2018.

CHOI, W.; et al. Feature Optimization for Gait Phase Estimation with a Genetic Algorithm and Bayesian Optimization**. Applied Sciences**. v. 11, p. 8940, 2021.

CIAIS, P., C. et al. **Carbon and Other Biogeochemical Cycles**. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University, New York, 2013.

COLEY, D. **An Introduction to Genetic Algorithms for Scientists and Engineers**. Singapore: World Scientific. 223p, 1999.

CUTLER, A., CUTLER, D.R., STEVENS, J.R. **Random Forests**. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, Boston, MA. pp. 157-175, 2012.

DANG, S. N. et al. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. **Ecological Informatics**. v. 50, p. 24-32, 2019.

DEB, K. **Multi-Objective Optimization using Evolutionary Algorithms**. Chichester: John Wiley. 497p, 2001.

EPPLE, C. et al. **Managing ecosystems in the context of climate change mitigation: A review of current knowledge and recommendations to support ecosystem-based mitigation actions that look beyond terrestrial forests.** Technical Series No.86. Secretariat of the Convention on Biological Diversity, Montreal. 55p, 2016.

FATOYINBO, T. E. et al. Estimating mangrove aboveground biomass from airborne LiDAR data: a case study from the Zambezi River delta. **Environmental Research Letters**. v. 13, n. 2, 2018.

FATOYINBO, T. E. et al. Landscape-scale extent, height, biomass, and carbon estimation of Mozambique's mangrove forests with Landsat ETM+ and Shuttle Radar Topography Mission elevation data. **Journal of Geophysical Research**. v. 113, 2008.

FISCHER, R. The Long-Term Consequences of Forest Fires on the Carbon Fluxes of a Tropical Forest in Africa. **Applied Sciences**. Basel: MPDI. v. 11, n. 4696, 2021.

GHOSH, S.M.; BEHERA, M.D.; PARAMANIK, S. Canopy Height Estimation Using Sentinel Series Images through Machine Learning Models in a Mangrove Forest. **Remote Sensing**. v. 12, 2020.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. New York: Addison-Wesley. 412p, 1989.

GOMIDE, L. R. **Planejamento Florestal Espacial**. Universidade Federal do Paraná. Tese de doutorado. Curitiba, 2009.

GOU, Y.; RYAN, C.M.; REICHE, J. LARGE. Area Aboveground Biomass and Carbon Stock Mapping in Woodlands in Mozambique with L-Band Radar: Improving Accuracy by Accounting for Soil Moisture Effects Using theWater Cloud Model. **Remote Sensing**. v. 14, p. 404, 2022.

GUEDES, B. S.; SITOE A. A.; OLSSON, B. A. Allometric models for managing lowland miombo woodlands of the Beira corridor in Mozambique. **Global Ecology and Conservation**. v. 13, p. e00374, 2018.

GUERRA-HERNÁNDEZ, J. et al. **Comparison of ALS based models for estimating aboveground biomass in three types of Mediterranean forest. European Journal of Remote Sensing. v. 4**9, p. 185-204, 2016.

HARMSE, C.J.; GERBER, H.; VAN NIEKERK, A. Evaluating Several Vegetation Indices Derived from Sentinel-2 Imagery for Quantifying Localized Overgrazing in a Semi-Arid Region of South Africa. **Remote Sensing**. v. 14, p. 1720, 2022.

HAUPT, R. L.; HAUPT, S. H. **Practical Genetic Algorithms**. 2nd ed. New Jersey: John Wiley. 253p, 2004.

HOLM, J. A.; KUEPPERS, L. M.; CHAMBERS, J. Q. Novel tropical forests: response to global change. **New Phytologist**. v. 213, n. 3, p. 988–992, 2017.

HOLZMAN, B. A. **Tropical forest biomes.** In, WOODWARD, S. L. (Ed.). Greenwood guides to biomes of the world. London: Greenwood. 242p, 2008.

HU, X. et al. Locating spatial variation in the association between road networkand forest biomass carbon accumulation. **Ecological Indicator**. n. 73, p. 214–223, 2017.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE – IPCC. **Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis.** Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. MASSON-DELMOTTE, et al (Eds.). Cambridge, 2021.

JACKSON, R. B. et al. Global energy growth is outpacing decarbonisation. **Environmental Research Letter**. v. 13, n.12, 2018.

JIANG, F. et al. Improving aboveground biomass estimation of natural forests on the Tibetan Plateau using spaceborne LiDAR and machine learning algorithms. **Ecological Indicators**. v. 143, 2022.

KOZA, J. R. **Genetic Programming: Automatic Synthesis of Topologies and Numerical Parameters**. In: GLOVER, J.; KOCHENBERGER, G; A (Ed). Handbook of Metaheuristics. New York: Kluwer Academic. 556p, 2003.

KOZA, J. R. **Genetic programming: on the programming of computers by means of natural selection**. Namco, 1992.

LARY, D. J. et al. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**. v. 7, n. 1, p. 3 – 10, 2016.

LENSEN, A.; XUE, B.; ZHANG, M. Genetic Programming for Evolving a Front of Interpretable Models for Data Visualization. **IEEE Transactions on Cybernetics**. v. 51, n. 11, p. 5468 – 5482, 2021.

LI, Z. et al. Improved Estimation of Bio-Oil Yield Based on Pyrolysis Conditions and Biomass Compositions Using GA- and PSO-ANFIS Models. **BioMed Research International**. p. 1-9, 2021.

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

LINDSELL J. A; KLOP E. Spatial and temporal variation of carbon stocks in a lowland tropical forest in West Africa. **Forest Ecology and Management**. v. 289, p. 10–17, 2013.

LISBOA, S.N. et al. Biomass allometric equation and expansion factor for a mountain moist evergreen forest in Mozambique. **Carbon Balance Manage**. v. 13, 2018.

LONDHE, S. et al. Tree Based Approaches for Predicting Concrete Carbonation Coefficient. ***Applied Sciences***. v. 12, n. 8, p. 3874, 2022

MACAVE, O.A. et al. Modelling Aboveground Biomass of Miombo Woodlands in Niassa Special Reserve, Northern Mozambique. **Forests**. V. 13, p. 311, 2022.

MAGALHÃES, T. M.; SEIFERT, T. Estimation of Tree Biomass, Carbon Stocks, and Error Propagation in Mecrusse Woodlands. **Open Journal of Forestry**. v. 5, p. 471-488, 2015.

MAGNUSSEN, S., REED, D. **Modelling for Estimation and Monitoring**. Chapter in *National Forest Assessment Knowledge Reference*. FAO, Rome, Italy, 2004.

MALHI, Y. The productivity, metabolism and carbon cycle of tropical forest vegetation. **Journal of Ecology**. v.100, p.65–75, 2012

MARTIN, M. A. et. al. Ten new insights in climate science 2021: a horizon scan. **Global Sustainability**. Cambridge: Cambridge University. v. 4, p. 1 – 20, 2021.

MARTINS SILVA, J. P. et al. Computational techniques applied to volume and biomass estimation of trees in Brazilian savanna. **Journal of Environmental Management**. v. 249, p. 1 – 12, 2019.

MATE, R.; JOHANSSON, T.; SITOE, A. Biomass Equations for Tropical Forest Tree Species in Mozambique. **Forests**. v. 5, n. 3, p. 535-556, 2014.

MEHR, A. D.; KAHYA, E. A Pareto-optimal moving average multigene genetic programming model for daily streamflow prediction. **Journal of Hydrology**. v. 549, p. 603-615, 2017.

MENG, Y.; JIA, L. 2018. Global warming causes sinkhole collapse – Case study in Florida, USA. **Natural Hazards and Earth System Science**. p. 1 – 8, 2018.

MEYER, V., et al. Canopy area of large trees explains aboveground biomass variations across neotropical forest landscapes. **Biogeosciences**. v.15, n. 15, p. 3377–3390, 2018.

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

MIRANDA, E. N. **Hipsometria: Seleção de Variáveis e Mineração de Dados por Métodos de Inteligência Computacional**. Dissertação de mestrado. UFLA. 87p, 2020.

MITCHARD, E. T. A. et al. Using satellite radar backscatter to predict above-ground woody biomass: A consistent relationship across four different African landscapes. **Geophysical Research Letters**. v. 36, 2009.

MITCHELL, M. **An Introduction to Genetic Algorithms**. Cambridge: Massachusetts Institute of Technology. 158p, 1996.

MOGHADDAM, S. A. et al. An automatic feature construction method for salient object detection: A genetic programming approach. **Expert Systems with Applications**. v. 186, p. 603-615, 2021.

MULLER-LANDAU, et al. Patterns and mechanisms of spatial variation in tropical forest productivity, woody residence time, and biomass. **New Phytologist**. v. 229, n. 6, p. 3065–3087, 2021.

NAGESWARA-RAO, M.; SONEJI, J. R.; SUDARSHANA, P. **Structure, Diversity, Threats and Conservation of Tropical Forests.** In: Sudarshana, P. Nageswara-Rao, M.; Soneji, J. R. Tropical Forests. Croatia: InTech. 388p, 2012.

NAVARRETE-SEGUEDA et al. Variation of main terrestrial carbon stocks at the landscape-scale are shaped by soil in a tropical rainforest. **Geoderma**. v. 13, p. 57–68, 2018.

ØSTERGAARD, P. et al. C-band SAR for the GMES Sentinel-1 mission. **2011 8th European Radar Conference**. Manchester, UK, pp. 234-240, 2011.

PHAM, T. D. et al. Comparison of Machine Learning Methods for Estimating Mangrove Above-Ground Biomass Using Multiple Source Remote Sensing Data in the Red River Delta Biosphere Reserve, Vietnam. **Remote Sensing**. v. 12, n. 1334, p. 1- 24, 2020.

POLI, R.; LANGDON, W. B.; MCPHEE, N. F. **A field guide to genetic programming**. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, (With contributions by J. R. Koza). GPBiB, 2008.

QIAN, C. et al. Estimation of Forest Aboveground Biomass in Karst Areas Using Multi-Source Remote Sensing Data and the K-DBN Algorithm. **Remote Sensing**. v. 13, p. 5030, 2021.

REEVES, C. **Genetic Algorithms**. In: GLOVER, J.; KOCHENBERGER, G; A (Ed). Handbook of Metaheuristics. New York: Kluwer Academic. 556p, 2003.

RIBEIRO, N. S. et al. Aboveground biomass and leaf area index (LAI) mapping for Niassa Reserve, northern Mozambique. **Journal of Geophysical Research**. v. 113, 2008.

RODIG. E. et al. The importance of forest structure for carbon fluxes of the Amazon rainforest. **Environmental Research Latter.** v.13, n. 5, 2017.

ROTHLAUF, F. **Representations for Genetic and Evolutionary Algorithms**. 2nd ed. New York: Springer. 325p, 2006.

ROZENDAAL, et. al. Aboveground forest biomass varies across continents, ecological zones and successional stages: refined IPCC default values for tropical and subtropical forests. **Environmental Research Letter.** v. 17, n. 1, 2022.

RYAN, C. M.; WILLIAMS, M.; GRACE, J. Above- and Belowground Carbon Stocks in a Miombo Woodland Landscape of Mozambique. **Biotropica**. v. 43, n. 4, p. 423–432, 2011.

SINGH, C. et al. Remote sensing-based biomass estimation of dry deciduous tropical forest using machine learning and ensemble analysis. **Journal of Environmental Management**. v. 308, 2022.

SITOE, A. A.; MANDLATE, L. J. C.; GUEDES, B. S. Biomass and Carbon Stocks of Sofala Bay Mangrove Forests. **Forests**. v. 5, p. 1967-1981, 2014.

SUBEDI, B. P. et al. **Forest Carbon stock measurement: Guidelines for measuring carbon stocks in community-managed forests**. Asia Network for Sustainable Agriculture and Bioresources (ANSAB). Federation of Community Forest Users, Nepal (FECOFUN). International Centre for Integrated Mountain Development (ICIMOD). Norwegian Agency for Development Cooperation (NORAD). 69p, 2010.

SVETNIK, V. et al. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. **J. Chem. Inf. Comput. Sci.** v. 43, p. 1947-1958, 2003.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; KERRY, R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. **Geoderma**. v. 266, p. 98-110, 2016.

TARAVAT, A.; WAGNER, M.P.; OPPELT, N. Automatic Grassland Cutting Status Detection in the Context of Spatiotemporal Sentinel-1 Imagery Analysis and Artificial Neural Networks. **Remote Sensing.** v. 11, n. 6, p. 711, 2019.

TARPANELLI, A.; MONDINI, A. C.; CAMICI, S. Effectiveness of Sentinel-1 and Sentinel-2 for flood detection assessment in Europe, **Nat. Hazards Earth Syst**. **Sci.,** v. 22, n.8, p. 2473–2489, 2022.

TEXEIRA, K. A. et al. Carbon dynamics of mature and regrowth tropical forests derived from a pantropical database (TropForC-db). **Global Change Biology**. v. 22, p. 1690–1709, 2016.

THOMAS, S. C.; BALTZER, J. L. Tropical Forests. **Encyclopaedia of Life Sciences**. John Wiley. p. 1-8, 2002.

TORRES, R. et al. GMES Sentinel-1 Mission. **Remote sensing of environment**. v.120, p. 9-24, 2012.

TWERY, M. J.; WEISKITTEL, A. R. **Forest-Management Modelling. In, Environmental Modelling: Finding Simplicity in Complexity.** 2nd ed. Wainwright, J. e Mulligan, M. (Ed). John Wiley. p.279 – 398, 2013.

UNITED NATIONS ENVIRONMENT PROGRAMME AND INTERNATIONAL UNION FOR CONSERVATION OF NATURE (2021). **Nature-based solutions for climate change mitigation**. Nairobi and Gland. 34p, 2021.

UNITED NATIONS. **Paris Agreement**. 25 p, 2015.

WEN, C. et al. An Object-Based Genetic Programming Approach for Cropland Field Extraction. **Remote Sensing**. v. 14, 1275, 2022.

WIEGAND, P.; PELL, R. COMAS, C. Simultaneous variable selection and outlier detection using a robust genetic algorithm. **Chemometrics and Intelligent Laboratory Systems.** v. 98, p. 108 – 114, 2009.

XUE, J.; SU, B. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. **Journal of Sensors**. v. 2017, 2017.

YAMAGUCHI, M. et al. Global warming changes tropical cyclone translation speed. **Nature Communications**. v. 11, n. 47, p. 1- 20, 2020.

YESUF, G.; Brown, K. A.; Walfor, N. Assessing regional-scale variability in deforestation and forest degradation rates in a tropical biodiversity hotspot. **Remote Sensing in Ecology and Conservation**. London: John Wiley. v. 5, n. 4, p. 346-359, 2019.

ZAKI, N. A. M; LATIF, Z. A. Carbon sinks and tropical forest biomass estimation: a review on role of remote sensing in aboveground-biomass modelling. **Geocarto International**. v. 32, n. 7, p. 701-716, 2017.

ZHOU, Z.; YANG, Y.; ZHANG, G.; XU, L.; WANG, M. EBM3GP: A novel evolutionary bi-objective genetic programming for dimensionality reduction in classification of hyperspectral data. **Infrared Physics & Technology**. v. 129, 2023.

CAPÍTULO 2 (ARTIGO)

TÍTULO

Algoritmos Evolucionários na Predição de Estoque de Carbono Acima do Solo em Florestas de Mopane - Moçambique

TITLE

Evolutionary Algorithms for Predicting Aboveground Carbon Stock in Mopane Woodlands - Mozambique

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

RESUMO

As florestas tropicais desempenham papel importante na regulação do clima global e do ciclo de carbono. Mopane é um tipo de floresta tropical seca, que ocorre na África Austral, com importância socioeconômica a nível local. A exploração do Mopane para carvão vegetal, em Moçambique, causa degradação e redução de estoques de carbono. Estudos de carbono neste tipo florestal podem auxiliar no monitoramento dos estoques de carbono, no âmbito do combate às mudanças climáticas, incluindo Redução de Emissões por Desmatamento e Degradação Florestal, mais Manejo de Florestas, Conservação e Aumento de Estoques de Carbono (REDD+). No presente estudo foram testados métodos de *Machine Learning*, aplicando algoritmos evolucionários e dados de sensoriamento remoto, cobertura florestal, biofísicas e bioclimáticas para predizer estoques de Carbono Acima do Solo (AGC) na floresta de Mopane, nos distritos de Mabalane e Chicualacuala, província de Gaza, Moçambique. A amostra de campo foi composta por 114 *clusters* e foram usadas imagens de Sentinel-2, Sentinel-1, MODIS e de World.Clim para extração das variáveis. Foram testadas 139 variáveis de diferente natureza para predizer o AGC, usando (*i*) método hibrido entre Algoritmo Genético-AG para seleção de variáveis e *Random Forest* - RF para predição (GARF) e (*ii*) Programação Genética (PG) via regressão simbólica. Ambos métodos reduziram o tamanho da base de dados em 95.6%. O GARF aderiu-se mais a variáveis bioclimáticas e de sensores ópticos, enquanto a PG combinou variáveis independentemente de sua natureza e pode gerar modelos mistos e segmentados. Os valores de AGC (em $MgC.ha^{-1}$) medidos no campo variaram de 1.313 a 28.476, média = 10.988. O AGC estimado por GARF variou de 2.910 a 19.459, média = 10.235, raiz do erro quadrado médio normalizado – nRMSE = 0.427 e erro médio de viés - BEM = 0.08. Para PG variou de 1.721 a 23.503, nRMSE = 0.428 e BEM = 2.731×10^{-17}. Ambos métodos mostraram eficiência na seleção de variáveis e potencial para predição de AGC em florestas tropicais secas. A PG é mais prática em relação ao GARF, por fornecer um modelo com estrutura visível e facilmente replicável.

Palavras-chaves: Mopane. Carbono Acima do Solo. Algoritmos Evolucionários. Algoritmo Genético e *Random Forest* (GARF). Programação Genética.

ABSTRACT

Tropical forests play an important role in the global climate regulation and the carbon cycle. Mopane is a tropical dry forest, occurring in southern Africa, with socioeconomic importance at the local level. Mopane harvesting for charcoal production in Mozambique is a main driver for forest degradation and carbon stocks reduction. Estimating carbon stocks in this forests can help monitoring carbon emissions, in this type of forest can help to assess and monitoring CO_2 emissions, in the context of climate change, including Reduction of Emissions from Deforestation and Forest Degradation (REDD+). In this study, we tested Machine Learning methods by applying evolutionary algorithms and remote sensing, forest cover data, biophysical and bioclimatic data to predtic Aboveground Carbon (AGC) in the Mopane forest, in the districts of Mabalane and Chicualacuala, Gaza province, Mozambique. The sample was composed of 114 clusters and we used satellites images from Sentinel-2, Sentinel-1, MODIS and World.Clim dataset to extract the predictor variables. A set of 139 variables of different nature has been tested to predict the AGC, using (i) the hybrid method between Genetic Algorithm-AG for variable selection and Random Forest - RF for prediction (GARF) and (ii) Genetic Programming (PG) via symbolic regression. Both methods were able to reduce the database size by 95.6%. The GARF adhered more to bioclimatic variables and optical sensors, while the PG combined variables regardless of their nature and can generate mixed and segmented models. The AGC values (in MgC.ha-1) from field survey ranged from 1.313 to 28.476, mean = 10.988. The AGC estimated by GARF ranged from 2.910 to 19.459, mean = 10.235, normalized root mean square error – nRMSE = 0.427 and mean bias error - BEM = 0.08. For PG it ranged from 1.721 to 23.503, nRMSE = 0.428 and BEM = 2.731×10-17. Both methods showed efficiency for variables selection and potential for predicting AGC in tropical dry forests. The PG algorithm is more practical than GARF, as it provides a model with a visible and easily replicable structure.

Keywords: Mopane. Aboveground Carbon. Evolutionary Algorithms. Genetic Algorithm and Random Forest (GARF). Genetic programming.

1. INTRODUÇÃO

As florestas tropicais desempenham papel importante na regulação do clima global e do ciclo de carbono, sendo o maior reservatório de carbono terrestre (BONAN, 2008; FISCHER, 2021; HOLZMAN, 2008; TEXEIRA. et al., 2016). O carbono encontra-se armazenado na forma de compostos orgânicos, constituintes da biomassa e matéria orgânica morta e no solo (CIAIS, et al., 2013; MAGNUSSEN; REED, 2004; MALHI, 2012). Porém, o desmatamento e a degradação florestal resultam na redução da capacidade de sequestro e liberação de carbono para atmosfera, contribuindo para o acúmulo de gases de efeito estufa, o principal fator das mudanças climáticas (YESUF; BROWN; WALFOR, 2019).

A floresta de Mopane é um tipo de floresta tropical seca, caracterizado pela dominância de *Colophospermum mopane* (J. Kirk ex Benth.) J. Léonard (Fabaceae), que ocorre na região da África Austral (Angola, Namíbia, Botswana, Zimbabwe, Moçambique, África do Sul, Zâmbia e Malawi), ocupando cerca de 555 mil km^2 e faz parte do Centro de Endemismo do Zambeze (DE SOUSA et al., 2021; MAKHADO et al., 2014; NGAREGA; MASOCHA; SCHNEIDER, 2021; STEVENS, 2021). Em Moçambique, o Mopane ocorre nas províncias de Tete, Manica e Gaza, e forma uma das regiões fitogeográficas do país (DE SOUSA et al., 2021; MAQUIA, et al., 2019). Além dos serviços ecossistêmicos, o Mopane é importante para a economia local, servindo como fonte de alimentos, medicamentos, pasto, material de construção e combustível lenhoso (DE SOUSA et al., 2021; GARA et al., 2023; STEVENS, 2021; WOOLLEN et al., 2016). No entanto, a exploração seletiva para produção de carvão vegetal constitui o principal *driver* de degradação e redução dos estoques de carbono no Mopane (SEDANO et al., 2020, WOOLLEN et al., 2016). Segundo Sedano et al. (2016), a produção de carvão em Tete, foi responsável pela liberação de 37,545 ±4,826 MgC, entre 2011 e 2014. A quantificação dos estoques de carbono no Mopane e outras florestas tropicais é importante para o monitoramento das emissões de CO_2, no âmbito dos esforços de mitigação das mudanças climáticas e iniciativas globais como a Redução de Emissões por Desmatamento e Degradação Florestal, mais Manejo de Florestas, Conservação e Aumento de Estoques de Carbono (REDD+), do qual Moçambique é signatário. Moçambique possui estudos dessa natureza e já foram desenvolvidos modelos de biomassa/carbono nível árvore (por espécies e fisionomia) e nível povoamento (CARREIRAS, MELO; VASCONCELOS, 2013; GOU, RYAN; REICHE, 2022; GUEDES; SITOE; OLSSON, 2018; LISBOA et al., 2018; MACAVE ET AL., 2022; MAGALHÃES; RYAN; WILLIAMS; GRACE, 2011; SEIFERT, 2015; MATE; JOHANSSON; SITOE, 2014; SITOE; MANDLATE; GUEDES, 2014). Porém, desses estudos, uma limitada fração foi direcionada particularmente para floresta de Mopane.

A abordagem clássica de modelagem de biomassa/carbono envolve dados coletados em campo, dados de sensores remotos e ajuste de equações. No entanto, com a evolução computacional, a combinação entre Sensoriamento Remoto e algoritmos de *Machine Learning* (ML) apresenta potencial para predição de biomassa/carbono. Os algoritmos de ML têm capacidade de identificar e modelar as relações complexas (incluindo não lineares) entre variáveis, com menor esforço computacional (CHEN et al., 2018; FERNÁNDEZ-CARRILLO et al., 2022). Nos últimos anos, o uso de ML na seleção de variáveis e modelagem de biomassa/carbono tem proporcionado bons resultados (CHEN et al., 2018; GARA et al., 2023; MIRANDA, et al., 2022; QIAN et al., 2021). Porém, o desafio consiste na seleção de variáveis adequadas, principalmente quando o volume de dados é grande, devido à natureza combinatória das possibilidades (MIRANDA, et al., 2022) e à complexidade de fatores que afetam a biomassa/carbono em florestas tropicais (BEHERAA et al., 2016).

Random Forest (RF) é um algoritmo de ML, baseado em árvores de decisão para problemas de classificação e regressão (BREIMAN, 2001), já conceituado para modelagem de biomassa/carbono. A maioria dos trabalhos de modelagem via RF, com redução de variáveis, efetuam a seleção usando o método de *stepwise* por adição/remoção recursiva, baseada na importância relativa da variável, dada pelo RF (DANG et al., 2019; JIANG et al., 2022; LI et al., 2019; SILVEIRA et al., 2019). Porém, esse processo é oneroso e seu resultado é influenciado pela sequência de testagem. O uso do Algoritmo Genético (AG) para seleção de variáveis e RF para o ajuste, pode melhorar o desenvolvimento da modelagem (CARVALHO et al., 2022; MIRANDA, et al., 2022; TAVASOLI; AREFI, 2021). O AG é um algoritmo evolucionário que usa princípios genéticos para resolução de problemas (ARJASAKUSUMA; KUSUMA; PHINN; 2020; GOLDBERG, 1989; PHAM et al., 2020). Este método tem a vantagem de automatização do processo e eficiência na seleção de variáveis (MIRANDA, et al., 2022). Carvalho et al. (2022) combinaram AG e RF (GARF) na modelagem de carbono acima de solo (*Above Ground Carbon* – AGC), em floresta natural na Bacia do Rio Grande (Brasil), resultando na redução de 95% da base de dados sem comprometer a precisão. Tavasoli e Arefi (2021) também reportaram ganhos na predição de AGC usando o método de GARF em termos de número de variáveis, precisão e performance. Outro algoritmo evolucionário com potencial para modelagem de biomassa/carbono é a Programação Genética (PG) via regressão simbólica. A PG usa mesmos princípios que o AG, sendo que a diferença entre os métodos está na necessidade de um modelo pré-definido no AG, enquanto a PG (regressão simbólica) gera a estrutura do modelo, seleciona variáveis e coeficientes adequados em função do problema proposto (KOZA, 2003, 1992; POLI; LANGDON; MCPHEE, 2008; TAGHIZADEH-

MEHRJARDI; NABIOLLAHI; KERRY, 2016). Apesar da PG ter sido aplicada em diferentes áreas de conhecimento (LONDHE et al., 2022; MEHR e KAHYA, 2017; ZHOU et al., 2023; WEN et al., 2022), sua aplicabilidade foi pouco explorada na modelagem de biomassa/carbono na área florestal. Um dos exemplos da sua aplicação na área florestal consiste em um estudo conduzido por Ghosh; Behera e Paramanik (2020), no qual os autores reportaram melhor desempenho da PG (regressão simbólica) na predição da altura da copa de manguezais em relação a RF (PG: RMSE=1.48 m e R^2=0.62; RF: RMSE=1.57 m; R^2=0.6). Outro exemplo foi do estudo de Cabral et al., (2018) sobre classificação de áreas queimadas usando imagens satélites, em savanas tropicais (Brasil, Guiné-Bissau e Congo), onde a PG proporcionou melhor acurácia em relação as metodologias clássicas (MaxVer e CART).

No que concerne a fonte de dados para predição, estudos comprovaram que a combinação entre dados de fontes diferentes gera melhores estimativas em relação a uma única fonte (CHEN et al., 2018; MACAVE et al., 2022; GARA et al., 2023). A sinergia entre informações de naturezas distintas, ajuda a ultrapassar as limitações associadas a cada uma delas, melhorando o desempenho da predição (CHEN et al., 2018). Essa condição promove uma complementação de visões sobre as variáveis preditoras. Assim, no presente estudo buscou-se aplicar algoritmos evolucionários, bem como o uso de informações originárias do inventário florestal, sensoriamento remoto, cobertura florestal e variáveis ambientais (biofísicas e bioclimáticas) na modelagem do carbono arbóreo acima do solo (AGC) na floresta de Mopane, nos distritos de Mabalane e Chicualacuala, província de Gaza, Moçambique. Foram testadas duas estratégias metodológicas: (*i*) método híbrido que combina o AG com RF (GARF) e (*ii*) PG via regressão simbólica. Além disso, avaliar e comparar o desempenho dos métodos dos métodos testados, verificando a influência das variáveis selecionadas pelos modelos na variação de AGC.

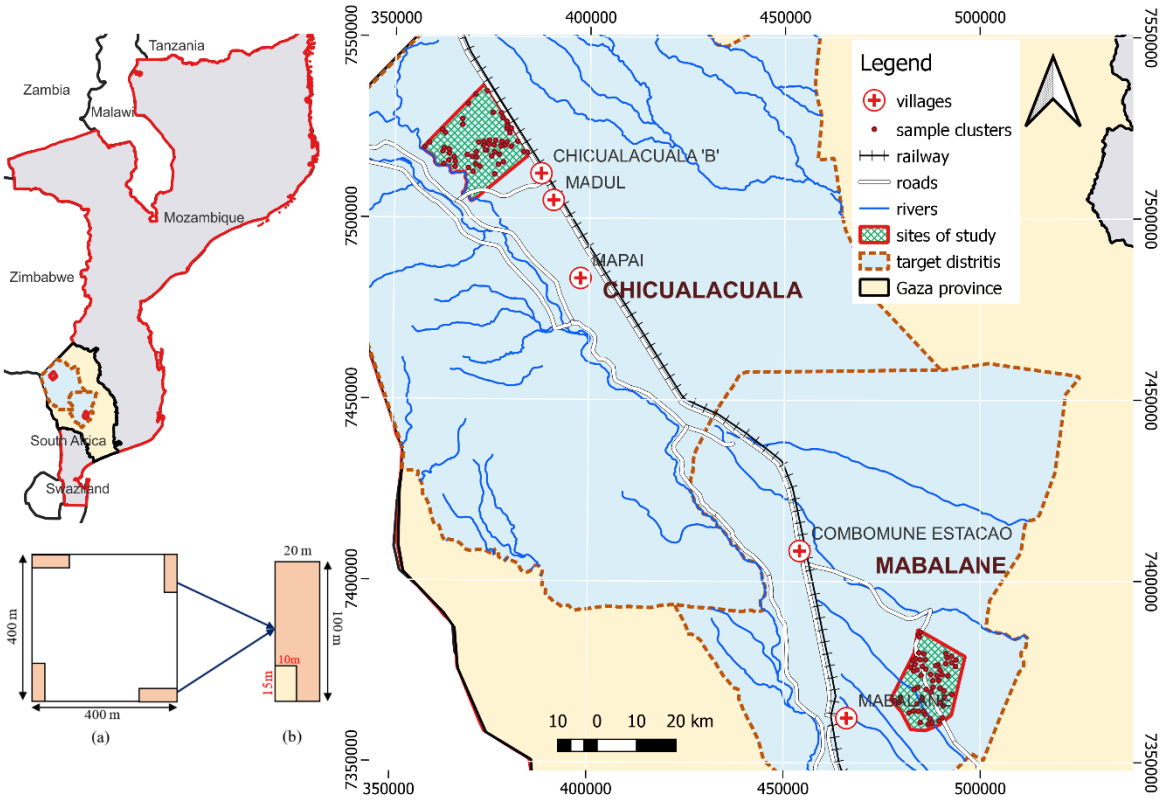
2. METODOLOGIA

2.1. Descrição da área de estudo

Moçambique localiza-se na costa Oriental Africana, na região da África Austral, entre as latitudes 10º a 26º Sul, com clima tropical seco a úmido, existindo algumas regiões semiáridas. A cobertura florestal do país é estimada em 31.693.872 hectares, sendo que a província de Gaza contribui com 3.096.817 hectares. As principais formações vegetais incluem Miombo, Mopane e Mangal (Ministério de Terra Ambiente e Desenvolvimento Rural – MITADER, 2018). O estudo foi realizado na província de Gaza, nas comunidades de Nwamandzele e Chihondzoene, nos distritos de Mabalane e Chicualacuala, respetivamente

(FIGURA 2.1). A região do estudo corresponde a um dos pontos mais secos do país. O clima é tropical semiárido, com precipitação que varia de 300 a 700 mm/ano e temperatura média anual de 20 a 26 °C (SEDANO et al., 2020; WOOLLEN et al., 2016).

Figura 2.1 – Localização dos sites de estudo (Chihondzoene, no distrito de Chicualacuala e Nwamandzele, no distrito de Mabalane) e layout do cluster (a) e da parcela de amostragem (b).



Fonte: Do autor (2023).

A comunidade de Nwamandzele (32.461,0 ha) possui 30.001,0 ha de floresta e Chihondzoene (43.100,0 ha) tem 37.983,0 ha de floresta. A cobertura florestal é na sua maioria composta por Mopane, intercalado com fragmentos de Mecrusse (*Androstachys johnsonii*, Prain) e Mata Mista de *Guibortia conjugada* (Bolle) J. Léonard*, Combetum spp., Acacia spp* e outras espécies. O Mopane e Mecrusse formam povoamentos quase monoespecíficos de *C. mopane* e *A. johnsonii,* respetivamente. A floresta constitui principal fonte de sustento da população local, fornecendo alimentos, material de construção, medicamentos, pastagem e combustível lenhoso. A prática de agricultura é geralmente feita nas zonas baixas. Em Mabalane, a floresta de Mopane foi alvo de exploração seletiva, nas últimas décadas, para produção de carvão vegetal (SEDANO et al., 2020; WOOLLEN et al., 2016), enquanto em

Chihondzoene, no período amostragem, a floresta apresentava maior nível de conservação, pois não havia sido explorada para carvão.

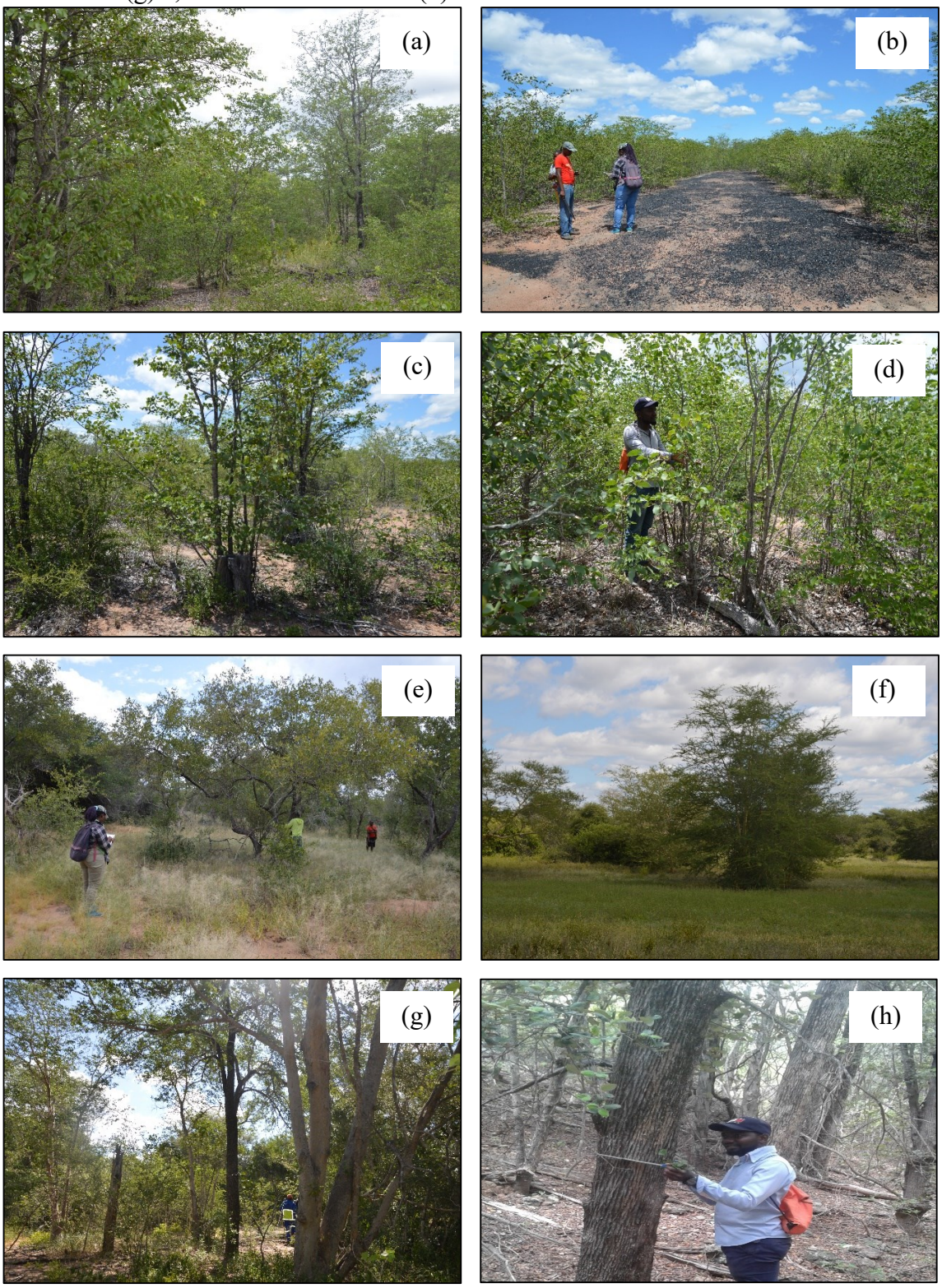
2.2.Coleta de dados

2.2.1. Amostragem de campo

Os dados foram provenientes do inventário florestal realizado entre os meses de fevereiro a abril de 2019. A coleta foi baseada na combinação entre amostragem aleatória estratificada e amostragem em conglomerados (*cluster sampling*). Em cada *site* de estudo, primeiramente foram definidos estratos com base nos tipos florestais, nomeadamente: (*i*) Mopane, (*ii*) Mecrusse e (*iii*) Mata Mista. A identificação dos tipos florestais foi baseada na estrutura fisionômica e composição de espécies. A vegetação de Mopane é dominada por *C. mopane*, o Mecrusse por *A. johnsonii* e a Mata mista por *Acacia spp, G. conjugata* e *Combetum spp* (FIGURA 2.2). Os mapas de vegetação foram resultado da classificação supervisionada de imagens satélites do sensor Sentinel-2A MSI, obtidas a uma data próxima do período de realização do inventário.

Em cada estrato florestal buscou-se a alocação dos *clusters* baseada na amostragem aleatória restrita, de modo a evitar sua sobreposição, adotando assim, uma distância mínima de 1,500 km entre os pontos de amostragem. Cada *cluster* foi composto por quatro parcelas de 100 × 20 m (0,2 ha) cada, distanciadas por 400 m uma da outra (FIGURA 2.1a). Em cada parcela foi estabelecida uma sub-parcela de 15 ×10 m (0,015 ha) para amostragem da regeneração estabelecida (FIGURA 1, b). As parcelas fora dos limites, em áreas não vegetadas e/ou em campos agrícolas foram excluídas da amostra. A amostra foi composta por 114 *clusters* (404 parcelas), sendo 59 (205 parcelas) em Mabalane e 55 (199 parcelas) em Chicualacuala. Na parcela principal foi mensurado o diâmetro a altura do peito (DAP), medido a 1,3 m em relação a superfície do solo, usando suta com precisão em mm, e a altura total usando vara graduada, para todos indivíduos com DAP igual ou superior a 10 cm. Na sub-parcela foram incluídos indivíduos com DAP entre 5 cm e 10 cm. Adicionalmente, foi feita a identificação de espécies arbóreas e arbustivas, pelos nomes científicos e locais, usando guias de identificação de espécies (BURROWA et al., 2019; VAN WYK; VAN WYK, 1997), com auxílio de um botânico e das populações indígenas para os nomes locais.

Figura 2. 2 - Floresta de Mopane (a); Vestígios de exploração de carvão vegetal no Mopane em Mabalane (b); Rebroto de *C. mopane* pós corte (c) Regeneração de *C. mopane* (d) Mata mista de *Combretum spp* (e), *Acacia spp.* (f) e *G. conjugata* (g) e; Floresta de Mecrusse (h).



Fonte: Do autor (2019)

2.2.2. Cálculo do Carbono acima do solo

A biomassa acima de solo (*Above Ground Biomass* –AGB) de cada indivíduo foi calculada com base na equação pan-tropical (CHAVE et al., 2014), que utiliza como *inputs,* o diâmetro a altura do peito (DAP) em cm, altura total (H) em m e densidade básica da madeira (ρ) em g.cm^{-3} (Equação 3), com os coeficientes β_0 = 0,0673 e β_0 = 0,976. Os valores da densidade básica das principais espécies foram obtidos no catálogo tecnológico de madeiras de Moçambique de Bunster (1995), e para as restantes foi usado o valor médio das espécies cujos valores da densidade são conhecidos. Esta equação é atualmente aceite pela comunidade científica para o cálculo de biomassa na região tropical. Ademais, as áreas abrangidas pela amostragem destrutiva para o desenvolvimento da equação incluem Moçambique.

$$AGB_{est} = \beta_0 \times (\rho \times DAP^2 \times H)^{\beta_1} \quad (1)$$

A quantidade de carbono por árvore foi obtida multiplicando o valor de AGB de cada árvore pelo fator de conversão de 0,475, que corresponde ao teor médio de carbono na biomassa vegetal (CHAVE et al, 2014; GUEDES; SITOE; OLSSON, 2018). A quantidade de carbono acima do solo (*Above Ground Carbon* –AGC) por parcela foi dada pelo somatório dos valores de todas árvores da parcela e para cada cluster pela média aritmética das parcelas que o compõe e extrapolado por unidade de área.

2.2.3. Variáveis bioclimáticas e de sensoriamento remoto

As variáveis bioclimáticas e de altitude foram obtidas do *WorldClim* versão 2 (worldclim.org), com resolução espacial de 1km^2. O *WorldClim* fornece informações derivadas da compilação de uma série histórica de dados climáticos e meteorológicos globais do período de 1950 a 2000 (FICK; HIJMANS, 2017). Esses dados têm sido amplamente usados para modelagem de diferentes variáveis na área ecológica incluindo a distribuição de espécies e teor de carbono (CARVALHO et al., 2019, 2017; CERASOLI; D'ALESSANDRO; BIONDI, 2022; MIRANDA et al., 2022). As variáveis de cobertura vegetal foram obtidas do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*), produto "*MOD44B.006 Terra Vegetation Continuous Fields Yearly Global 250m*" do ano de 2019. O produto fornece informações anuais de percentagem de cobertura de solo, com resolução espacial de 250 m que incluem: (*i*) percentagem de cobertura arbórea, (*ii*) percentagem de cobertura de vegetação não arbórea e (*iii*) porcentagem sem vegetação (DIMICELI et al., 2017). As variáveis do Radar foram obtidas a partir de imagens de satélite Sentinel-1 (S1). O S1 é equipado com sensores de Radar de abertura sintética (*Sinthetic Aperture Range - SAR*), que opera na banda C (entre 3,8 – 7,5 cm) com polarização VV, VH, HH e HV (ØSTERGAARD et al., 2011; TORRES et al., 2012). Foi

usada a coleção de imagens Sentinel –1A e Sentinel-1B do mês de julho de 2019 na polarização VV e VH, modo IW, nível de processamento 1 (*Ground Range Detected* - GRD), com o tamanho de pixel de 10 m × 10 m. A coleção foi agregada usando a mediana. E por fim, as variáveis do sensor óptico foram derivadas de imagens do Sentinel-2A MSI (S2), com 13 bandas espectrais de resoluções espaciais diferentes, sendo 10 m para bandas do visível, 20 m infravermelho e 60 m bandas de correção atmosférica (CHEN et al., 2018; HARMSE; GERBER; VAN NIEKERK, 2022). Em primeira instância, foi feita a simulação da modelagem com imagens tanto da época seca como chuvosa, e percebeu-se que as imagens da época melhor explicam a variação de dados de AGC, provavelmente devido a menor influência das gramíneas. Portanto, foram selecionadas imagens com cobertura de nuvens abaixo de 1%, correspondentes ao período seco mais próximo da data de inventário (3 de julho de 2019).

Os valores das bandas foram usados diretamente na modelagem e também transformados em índices espectrais. O cálculo dos índices foi feito no *Google Earth Engine* (GEE) usando o módulo *spectral* do pacote "*Awesome Spectral Indices*", uma lista códigos padronizada e pronta para uso, aplicada para o cálculo de índices espectrais no API do GEE (MONTERO, 2022), bastando fornecer informações das bandas, dos parâmetros e constantes necessários de acordo com os índices requeridos (XUE et al., 2017). As imagens do SENTINEL 1 e 2 já foram aplicadas em estudos de Uso e Cobertura de Terra e Modelagem biofísica em florestas, incluindo biomassa/carbono (CHEN et al., 2018, DANG et al., 2019; QIAN, et al., 2021; MACAVE et al., 2022; SINGH et al., 2022). Adotou-se o software QGIS versão 3.24.3 para extração dos dados bioclimáticos e a plataforma GEE para cálculo dos índices (Sentinel-2A) e extração dos valores de *raster* dos restantes sensores. Antes da extração dos valores no GEE, foi feita a reamostragem dos *pixels* usando o método *Neighborhood*, com raio de 2 *pixels* (20m - S1 e S2 e 500m - MODIS) e mediana como métrica redutora. A amostragem dos valores de *raster* foi feita para uma escala de 100m, que permite cobrir a parcela de amostragem para todos sensores.

2.3. Modelagem do carbono arbóreo acima do solo (AGC)

Na modelagem de AGC foram testadas no total 139 variáveis de naturezas distintas, sendo três fisionômicas, 3 geográficas, 5 de cobertura vegetal, 6 de sensor Radar, 19 bioclimáticas e 103 de sensor ópticos (TABELA 2.1). Informação detalhada das variáveis encontra-se na Tabela I.1 do Apêndice I. As variáveis fisionômicas (tipos florestais) e algumas geográficas (locais de estudo) foram aplicadas como variáveis *dummy* (0 ou 1).

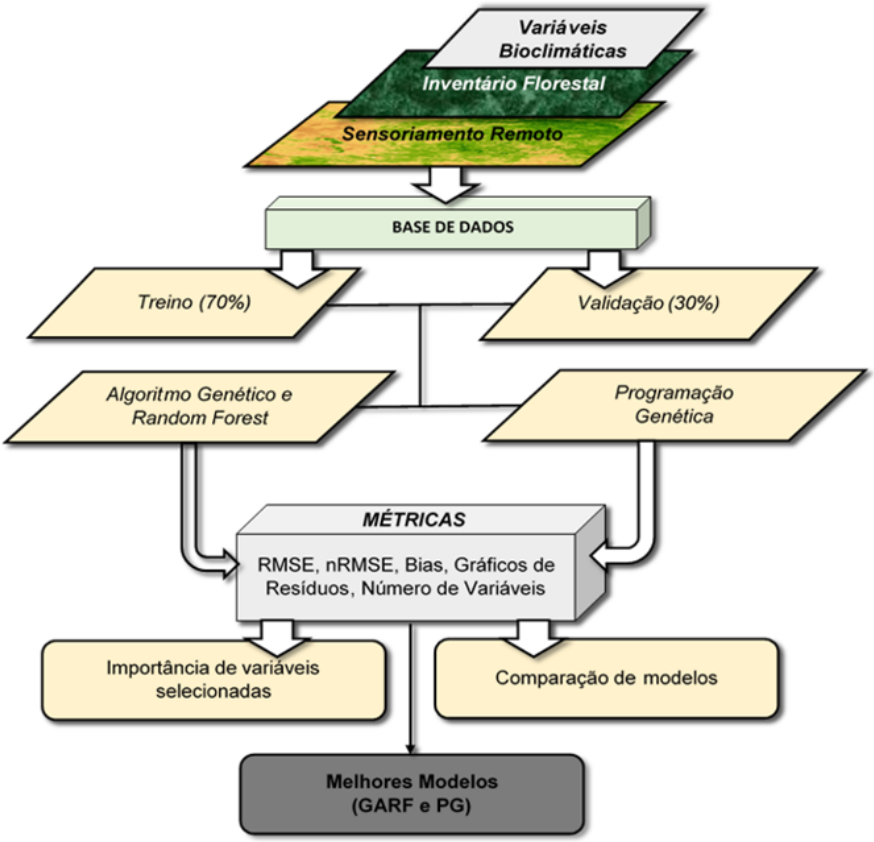
Tabela 2.1– Síntese das variáveis usadas na modelagem de AGC categorizadas em função da fonte de dados e da natureza das variáveis.

Natureza	Variáveis
Geográfica	Sites de estudo *(variável dummy)*, Altitude
Fisionômica	Tipos florestais *(variável dummy)*
C. vegetal	NTVg, NVg, NVg-sd, TCover, TCover-sd
RADAR	VH-A-D, VH-A, VH-D, VV-A-D, VV-A, VV-D
Bioclimática	bio1 a bio19 do WorldClim
Sensor óptico	**BANDAS**: B1 a B12, WVP, TCI-B, TCI-G, TCI-R **ÍNDICES**: AFRI1600, AFRI2100, ARVI, ATSAVI, BAIS2, BCC, BNDVI, BWDRVI, CIG, CIRE, CVI, DVI, EVI2, EVI, GARI, GBNDVI, GCC, GDVI, GEMI, GLI, GNDVI, GOSAVI, GRNDVI, GRVI, GSAVI, GVMI, IAVI, IPVI, IRECI, LSWI, MCARI1, CARI2, MCARI705, MCARIOSAVI705, MCARIOSAVI, MCARI, MGRVI, MNLI, MRBVI, MSAVI, MSR705, MSR, MTVI1, MTVI2, NBR, NDMI, NDPI, NDREI, NDVI705, NDVI, NDWI, NGRDI, NLI, OCVI, OSAVI, RDVI, REDSI, RENDVI, RGBVI, RGRI, RI, RVI, S2REP, S2WI, SARVI, SAVI, SIPI, SI, SR, SWM, SeLI, TCARI, TCI-B, TCI-G, TCI-R, TCI, TDVI, TGI, TRRVI, TTVI, TVI, TriVI, VARI, VI700, VIG, WDRVI, WDVI, WI1, WI2, WVP, mSR705

Fonte: Do autor (2023).

Foram testadas duas estratégias metodológicas: (*i*) método híbrido que combina o Algoritmo Genético com *Random Forest* (GARF) e (*ii*) Programação Genética (PG) via regressão simbólica. A Figura 2.3 mostra o fluxograma da metodologia usada.

Figura 2.3 – Esquema do procedimento metodológico para modelagem de AGC.



Fonte: Do autor (2023).

Inicialmente, testes preliminares foram aplicados no conjunto de dados para detecção/remoção de *outliers* usando o pacote "*outliers"* versão 0.15 (KOMSTA, 2022) onde removeu-se 1,2% da base de dados, definição do período de aquisição de imagens satélites (época seca ou chuvosa) e da estratégia de modelagem (nível parcela ou *cluster*). No entanto, as variações de AGC entre parcelas não foram facilmente explicadas pelas variáveis do sensoriamento remoto e bioclimáticas. Por outro lado, o estrato das gramíneas gerou ruídos nas imagens da época chuvosa, comprometendo sobremaneira o desempenho da modelagem. Portanto, foram usadas imagens da época seca e foi adotada a modelagem nível *cluster*, reduzindo os dados para uma escala de 114 unidades amostrais, obtidas através da média das parcelas que compõem cada *cluster*. Os dados foram aleatoriamente divididos em Treino (70%) e Validação (30%), para cada tipo florestal e site de estudo. Devido à natureza não paramétrica e estocástica dos métodos testados, adotou-se um conjunto de 30 repetições, extraindo o melhor modelo em cada método para critérios comparativos.

O primeiro método aplicado foi um algoritmo híbrido entre o Algoritmo Genético - AG e *Random Forest – RF*, denominado GARF (MIRANDA et al., 2022), que consiste na aplicação do AG para seleção de variáveis explicativas do modelo e RF para a modelagem/predição. O RF é um algoritmo de aprendizagem de máquinas baseado em um conjunto de árvores de decisão para resolver diferentes problemas classificação e regressão (BREIMAN, 2001). O RF usa amostras *bootstrap* para criar uma floresta não correlacionada de árvores de decisão, em que a predição é feita através do voto maioritário ou pela média das saídas de todas as árvores (BIAU, 2012; LIAW; WIENER, 2002; MARTINS SILVA et al., 2019). O RF tem se destacado pela simples parametrização, robustez, velocidade computacional, capacidade de resolver problemas complexos, geração de métricas de erro internas, importância de variáveis, detecção de *outliers*, entre outros (CUTLER; CUTLER; STEVENS, 2012; PHAM et al., 2020). Neste caso, a parametrização inicial do RF consistiu na definição do número de árvores (*ntrees* = 500) e do número de variáveis (*mtry* = 2), baseado no trabalho de Miranda et al. (2022). O algoritmo foi implementado no ambiente R (R CORE TEAM, 2022) usando o pacote *randomForest* (LIAW; WIENER, 2002).

As variáveis de entrada no modelo de RF foram selecionadas pelo AG. O AG é uma técnica de computação evolucionária que busca soluções para problemas complexos, imitando a teoria evolução de Darwin (DEB, 2001; GOLDBERG, 1989; HAUPT; HAUPT, 2004; PHAM et al., 2020; WIEGAND; PELL; COMAS, 2009). No AG a população evolui de uma geração para outra mediante a aplicação de operadores genéticos de seleção, *crossover* e mutação, de forma iterativa até alcançar o critério de parada (ARJASAKUSUMA; KUSUMA; PHINN,

2020; COLEY, 1999; MITCHELL, 1996; REEVES, 2003; ROTHLAUF, 2006). Neste estudo, a parametrização preliminar do AG incluiu a definição dos operadores de seleção (torneio), *crossover* (um ponto de corte) e mutação (bit aleatório), critério de parada (50 gerações) e o tamanho da população (100 indivíduos). A probabilidade de ocorrência de mutação foi de 10%, e taxa de troca aleatória de genes de 50% no cromossoma. A mutação visa manter a diversidade da população. Os indivíduos da população foram dimensionados para um vetor fixo de 139 posições (genes), correspondentes ao número de variáveis. Cada gene assume um valor de 0 ou 1, sendo que 0 desativa, e 1 ativa a variável para formar o indivíduo que vai servir de entrada no modelo de RF (MIRANDA, et al., 2022). A função *fitness* do AG foi definida no sentido de: (*i*) minimizar o erro médio quadrático (*out-of-bag* – OOB) da estimativa do AGC pelo RF e (*ii*) reduzir o número de variáveis preditoras (Equação 4). Neste caso, a função *fitness* foi composta pela soma da razão entre o erro OOB das variáveis habilitadas pelo AG e o erro OOB máximo possível (erro OOB do RF envolvendo todas as 139 variáveis) com a razão entre número de variáveis habilitadas pelo AG (n) e o número total de variáveis testadas no experimento (N).

$$fitness = \left(\frac{erro\,OOB}{erro\,OOB_{máx}} + \frac{n}{N} \right) \tag{2}$$

Por último, aplicou-se a Programação Genética (PG) que também utiliza o algoritmo genético na qual a população consiste em programas de computador, compostos por um conjunto de funções primitivas e terminais previamente fornecidos (KOZA, 2003, KOZA, 1992; POLI; LANGDON; MCPHEE, 2008; TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016). Portanto, quando esses programas de computadores são compostos por modelos empíricos, que buscam fornecer um bom ajuste, a um determinado conjunto de dados, a PG é designada como regressão simbólica (KOZA, 1992; MEHR e KAHYA, 2017). Diferentemente da regressão clássica, na qual a estrutura do modelo é conhecida, na regressão simbólica procura-se automaticamente construir o modelo e buscar os coeficientes numéricos. Neste caso, a regressão simbólica busca selecionar variáveis, operadores matemáticos e booleanos, funções matemáticas, assim como estimativa dos parâmetros para compor um modelo matemático (CABRAL et al., 2018; KOZA, 1992; TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016). Nesse sentido, os indivíduos são representados em forma de árvores de análise, criadas a partir da combinação de funções e terminais escolhidos em função da natureza e complexidade do problema. Para este método, foram feitos testes preliminares buscado a parametrização do algoritmo, tais como: o conjunto de funções, a função fitness, largura e profundidade máximas das árvores, taxa de mutação, tamanho da população e critério

de parada (número de gerações acima do qual não havia melhoria significativa da solução). O conjunto de funções primitivas foi compostos por operadores matemáticos (adição, subtração, multiplicação, divisão), funções matemática (exponencial e logarítmica) e operadores lógicos condicionais (*If, Then* e *Else*), estes últimos para permitir a criação de modelos mistos e/ou segmentados. Os terminais foram compostos por constantes e variáveis (139 variáveis testadas, TABELA 2.1). A criação de árvores foi probabilística, usando o conjunto de funções e terminais previamente definidas. Para cada árvore foi estabelecida uma profundidade máxima (*tree depth* = 12) e largura máxima (*tree length* = 30), conforme Poli, Langdon e Mcphee (2008). Adotou-se o software *HeuristicLab Optimizer 3.3.16.17186* do grupo "*Heuristic and Evolutionary Algorithms Laboratory (HEAL), University of Applied Sciences Upper Austria"* (FERNÁNDEZ-CARRILLO et al., 2022), baseado na linguagem C# e Microsoft .NET.

O algoritmo começou com uma população inicial de 2.000 programas de computador (modelos empíricos), gerados através de uma busca randômica, usando o conjunto de funções primitivas e terminais fornecidos. Cada indivíduo (modelo) na população constituía uma solução candidata ao problema. A evolução da população foi mediante o princípio de sobrevivência do mais apto e foram aplicados operadores de seleção (torneio), reprodução (1 indivíduo) por elitismo que consiste na cópia do parental com o melhor *fitness*, operador de *crossover* do tipo "*Subtree Swapping Crossover*" a uma probabilidade de 90%, que consiste na troca de uma fração de árvores entre os pais, e operador de mutação do tipo "*Multi Symbolic Expression Tree Manipulator"* a uma probabilidade de 10%, que consiste na troca aleatória de algumas funções ou terminais no modelo (POLI; LANGDON; MCPHEE, 2008; KOZA, 1992). Os operadores de seleção, reprodução, *crossover* e mutação foram repetidos até alcançar o critério de parada (100 gerações). A função *fitness* buscou minimizar o erro quadrático médio (*Mean Square Error* – MSE) das estimativas, conforme a Equação 3, onde: *i* é o número da instância; *n* é o número de observações do conjunto de dados (81 para treino e 33 para validação); *Yi* é o valor do carbono acima de solo ($MgC.ha^{-1}$) observado no *cluster i*; e *Yi* é o valor do carbono acima do solo ($MgC.ha^{-1}$) estimado pelo modelo no cluster *i*.

$$MSE = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{n} \quad (3)$$

2.4. Critérios de avaliação dos métodos

O desempenho dos métodos testados, para dados de treino e validação, foi avaliado com base nas seguintes métricas: Raiz do Erro Quadrático Médio (*Root Mean Square Error* –

RMSE, Fórmula 4), RMSE Normalizado (nRMSE, Fórmula 5), Erro Médio de Viés (*Mean Bias Error* – BEM, Fórmula 6) e Erro Médio Absoluto (*Mean Absolute Error* – MAE, Fórmula 7), assim como gráficos de dispersão e histograma de dispersão dos resíduos. Quanto mais próximo de zero os valores de RMSE, nRMSE e bias, melhor é o desempenho preditivo do modelo. Por outro lado, considerou-se também a frequência de seleção das variáveis durante todo o experimento. Neste caso, foram priorizados modelos com menor erro, menor número de variáveis, boa distribuição gráfica dos resíduos e que tenham selecionado variáveis mais frequentes durante o experimento.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{N}} \tag{4}$$

$$nRMSE = \frac{RMSE}{\bar{Y}} \tag{5}$$

$$MBE = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)}{N} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|\left(\hat{Y}_i - Y_i\right)\right|}{N} \tag{7}$$

Em que: i – número da instância; N – número total de observações do conjunto de dados (81 para treino e 33 para validação); Y_i – valor do AGC (MgC.ha^{-1}) observado no *cluster i*; $\hat{Y}\imath$ – valor do AGC (MgC.ha^{-1}) estimado pelo modelo no *cluster i;* $\overline{Y}$ – média aritmética do AGC (MgC.ha^{-1}) observado.

2.5. Análises pós-modelagem

A análise pós-modelagem para modelo de GARF consistiu na avaliação da importância relativa das variáveis através da análise da percentagem de melhoria do MSE (%IncMSE), obtida diretamente do pacote *randomForest* (LIAW; WIENER, 2002) para cada variável. Neste caso, o algoritmo RF foi executado 30 vezes, usando as variáveis do melhor modelo e a parametrização definida no treinamento (MIRANDA et al., 2022) para obtenção da média de importância (%IncMSE) de cada variável. Para o modelo de PG a análise pós-modelagem consistiu na simplificação e reajuste do modelo, determinação da complexidade e análise da sensitividade das variáveis (GHOSH; BEHERA; PARAMANIK, 2020), através da soma da complexidade das funções e terminais usados no modelo (ARYADOUST, 2015).

A PG via regressão simbólica no software *HeuristicLab* gera expressões em forma de árvores de sintaxe, que também podem ser visualizadas na sua forma matemática, porém a estrutura do modelo pode conter termos redundantes, os quais podem ser simplificados por operações algébricas. Neste caso, o melhor modelo foi simplificado, reajustado usando o método de mínimos quadrados ordinários (regressão linear) e submetido ao teste de multicolinearidade usando o pacote *car* do R (FOX; WEISBERG, 2019), com a remoção recursiva e permutada dos termos com Fator de inflexão de variância (*Variance inflation fator* – VIF) menor que cinco. Um dos objetivos da PG via regressão simbólica é identificar variáveis com maior poder explanatório sobre a variável de interesse. Portanto, a análise de sensitividade expressa o efeito da variação de uma determinada variável de entrada sobre a variável resposta (GHOSH; BEHERA; PARAMANIK, 2020; PANDEY; PANDEY, 2020). A sensitividade de cada variável foi calculada com base na derivada parcial do modelo gerado y = f(x), em função de cada variável independente (x), multiplicado pela razão entre a variância da variável em causa (σx) e a variável resposta (σy) conforme a Fórmula 8.

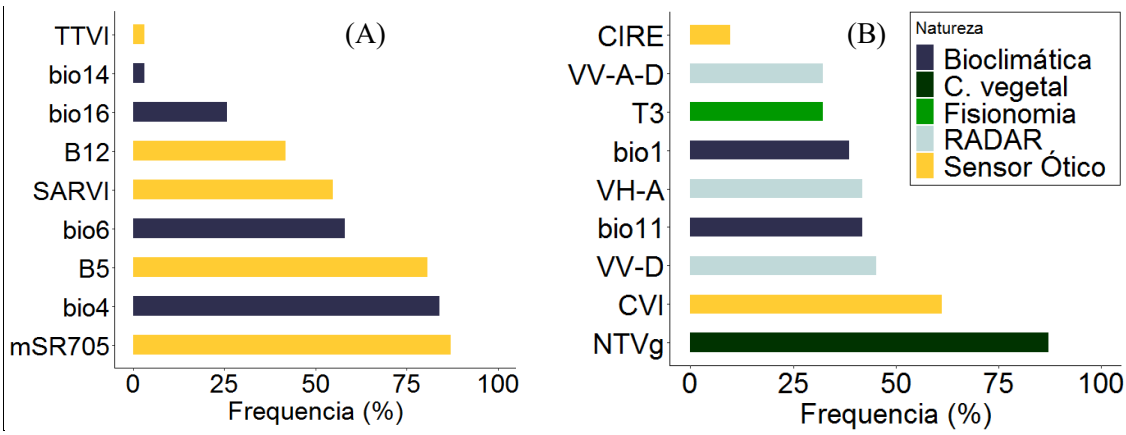
$$Sensitividade: \left|\overline{\frac{\partial y}{\partial x}}\right| \times \frac{\sigma x}{\sigma y} \quad \text{, para todo o conjunto de dados em análise} \tag{8}$$

A sensitividade indica a direção (positiva ou negativa) e a magnitude de correlação entre as variáveis independentes e dependente (GHOSH; BEHERA; PARAMANIK, 2020), e pode ser expressada pela percentagem e magnitude do positivo e do negativo. A percentagem e magnitude (fórmula 10) positiva/negativa são calculadas para o subconjunto de dados em que $\partial y / \partial x$ é positivo/negativo, respetivamente. A percentagem expressa a probabilidade na qual o aumento da variável de entrada resulta no aumento (positivo) ou redução (negativo) da variável de saída. A magnitude determina a quantidade de incremento ou redução da variável de saída devido a um aumento de uma unidade na variável de entrada. Quanto maior for a magnitude maior é a sensitividade da variável.

3. RESULTADOS

Na modelagem de carbono acima do solo (AGC) foram testadas 139 variáveis, provenientes de diferentes fontes. Adotou-se duas estratégias de modelagem, Algoritmo Genético combinado com *Random Forest* (GARF) e Programação Genética (PG) via regressão simbólica. Feita a modelagem, as variáveis com frequência de seleção abaixo do limite mínimo estabelecido (30%) foram excluídas das análises posteriores, com exceção daquelas que compõem os modelos selecionados (FIGURA 2.4). Apenas um grupo restrito de variáveis (9 para cada método), mostrou-se importante para explicar as variações do AGC na área de estudo para ambos métodos. As variáveis selecionadas com maior frequência (>30%) foram seis para GARF e oito para PG, sendo que o melhor modelo para ambos métodos foi composto por 6 variáveis, o que representa uma redução da base de dados em cerca de 95.6%. Estes resultados mostram o poder dos métodos testados em selecionar apenas aquelas variáveis que melhor explicam a variação do AGC.

Figura 2.4 – Frequência numérica das variáveis mais selecionadas e dos modelos selecionados de GARF (A) e PG (B).



LEGENDA: mSR705 – *Modified Simple Ratio (705 and 445 nm);* bio4 – sazonalidade da temperatura (desvio padrão ×100); B5 – *Red edge 1 (RE 1)*; bio6 – temperatura mínima do mês mais frio; *SARVI – Soil Adjusted and Atmospherically Resistant Vegetation Index*; B12 – *short wave infrared 2* (*SWIR 2*); bio16 – precipitação do trimestre mais úmido; bio14 – precipitação do mês mais seco; TTVI – *TTVI - Transformed Triangular Vegetation Index;* NTVg – *Non Tree Vegetation percent; CVI - Chlorophyll Vegetation Index*; VV-D – *Single co-polarization, vertical transmit/vertical receive*, *descending*; bio11 – temperatura média do trimestre mais frio; VH-A – *Dual-band cross-polarization, vertical transmit/horizontal receive, ascending*; bio1 – Temperatura Média Anual; T3 – tipo florestal 3 (mopane); VV-A-D – *co-polarization* VV *ascending and descending mean; CIRE – Chlorophyll Index Red Edge*.

Fonte: Do autor (2023).

No que concerne a composição dos melhores modelos, 50% das variáveis do melhor modelo de GARF, tiveram frequência de seleção acima de 80 %, nomeadamente, mSR705 (90.0%), bio4 (87.9%) e B5 (83.3%) e as restantes com frequência abaixo de 30%, nomeadamente, bio16, bio14 e TTVI com 26.7%, 3.3% e 3.3%, respetivamente. Para PG, quase

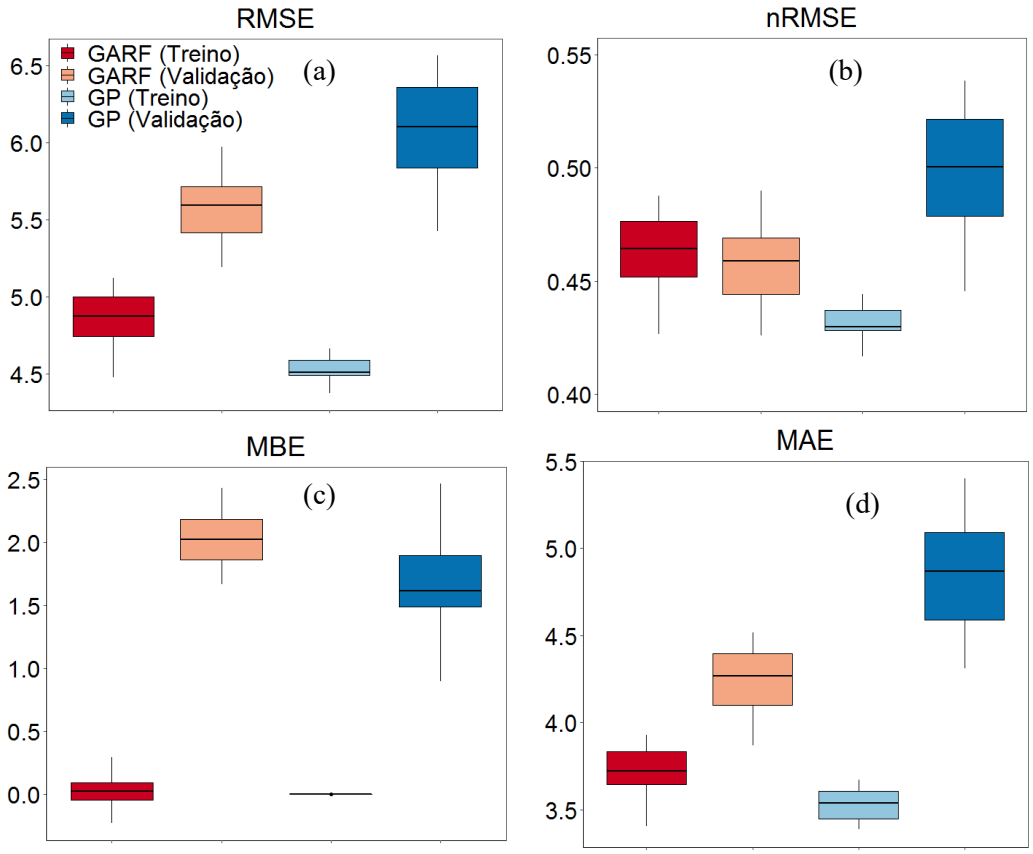
todas variáveis selecionadas obtiveram uma alta frequência de seleção, nomeadamente, NTVg (90.0%), CVI (63.3%), bio11(43.3%), T3 (33.3%) e VV-A-D (33.3%), com exceção da variável CIRE (10%), que teve frequência de seleção baixa. A inclusão de variáveis com baixa frequência de seleção nos melhores modelos está associada a natureza estocástica dos métodos testados, não obstante, estas variáveis podem estar correlacionadas com outras do mesmo tipo selecionadas com maior frequência.

A PG foi mais abrangente em termo da natureza das variáveis selecionadas em relação ao GARF. O GARF selecionou com mais frequência variáveis provenientes do sensor óptico (Sentinel-2 MSI), referente a bandas específicas (B5 e B12) e índices de vegetação (*mSR705, SARVI e TTVI*), e variáveis bioclimáticas (bio4, 6, 14 e 16) do World.Clim (FIGURA 2.3A). Em contrapartida, a PG selecionou com mais frequência variáveis referentes à cobertura vegetal (*NTVg*) proveniente do MODIS, variáveis do sensor óptico (Sentinel-2) referentes aos índices de vegetação (CVI e CIRE), variáveis do RADAR (sentinela-1 SAR) referentes a co-polarização VV (VV-A-D e VV-D) e polarização cruzada VH (VH-A), variáveis bioclimáticas (bio1 e 11), assim como de fisionomia (T3), nesta última gerando modelos com variável *dummy* (FIGURA 2.3B).

Por outro lado, notou-se um comportamento diferenciado no que se refere ao número de variáveis selecionadas para compor o modelo, assim como na correlação entre o AGC e as variáveis selecionadas por cada método (TABELA 2.2). Os modelos de GARF tenderam a selecionar um maior número de variáveis por modelo (6 a 17 variáveis, CV= 23.76%) e de baixa correlação com o carbono acima de solo ($r < \pm 0.2$). A PG foi mais estável no número de variáveis selecionadas (5 a 8 variáveis, CV= 14.18%) e tendeu a selecionar variáveis mais correlacionadas com o carbono acima de solo ($\pm 0.2 \leq r \leq \pm 0.35$).

Na Figura 2.5 apresenta-se o *boxplot* resumo das métricas de avaliação do desempenho dos modelos. Tanto o GARF quanto a PG foram consistentes nas estimativas do erro para o conjunto de treino. Portanto, pode se notar claramente que na base de treino a PG foi superior em relação ao GARF em todas as métricas e na base de validação em erro médio BEM (FIGURA 2.5 c). A média de MBE (em $MgC.ha^{-1}$) para GARF foi 0.019 (treino) e 2.021 (validação), comparado com 9.74×10^{-07} (treino) e 0.898 (validação) para PG, o que revela maior equilíbrio na distribuição dos resíduos e que os erros positivos são anulados pelos erros negativos.

Figura 2. 5 – Resumo das métricas de desempenho em RMSE (a), nRMSE (b), BEM (c) e MAE (d) dos modelos de GARF e PG em todas repetições (30) considerando dados de treino e de validação.



LEGENDA: RMSE – Raiz do Erro Quadrático Médio (*Root Mean Square Error*), nRMSE – RMSE Normalizado; MBE – Erro Médio de Viés (*Mean Bias Error*) e; MAE – Erro Médio Absoluto (*Mean Absolute Error*).

Fonte: Do autor (2023)

A associação das métricas de erro guiou a seleção do melhor modelo para cada método. A condição multiobjetivo do GARF teve como intuito atender a parcimônia dos modelos, minimizando o erro (erro das estimativas OOB do RF) e o número de variáveis para seleção interna do modelo (*best fitness*). Já para a PG, a condição do *fitness* foi mono-objetivo, e visava minimizar apenas o erro quadrático médio (MSE). Portanto, o critério de seleção entre os modelos (entre as repetições) envolveu o erro e gráfico de distribuição dos resíduos. A estrutura do modelo GARF não pode ser visualizada, podendo apenas conhecer as variáveis selecionadas (mSR705, bio4, B5 bio16, bio14 e TTVI). Em contrapartida, a PG fornece uma função visível, com variáveis e coeficientes definidos. O melhor modelo de PG apresentava aparentemente uma estrutura não linear (FIGURAS 3.1 e 3.2, ANEXO I). Porém, após a sua simplificação matemática ele apresentou uma forma linear (Expresso 9). Isto acontece porque na PG, via regressão simbólica, as expressões matemáticas são apresentadas em forma de árvores de sintaxe. No entanto, o *software* utilizado gera expressões com termos redundantes, que podem

ser simplificados matematicamente. Do ponto de vista teórico, uma simplificação matemática final é desejável.

$$C = \beta_0 + \beta_1 \frac{bi01 * VH}{MRBVI^2} + \beta_2 \frac{bio1 * bio8}{MRBVI^2} - \beta 3 \frac{NTVg}{MRBVI^2} + \beta 4 MRBVI^{-2} \quad (9)$$

Após a simplificação e reajuste, pelo método de mínimos quadrados ordinários e submetido ao teste de multicolinearidade, verificou-se que havia problema de multicolinearidade (VIF>5). Esse cenário revela que quando existe combinação de variáveis da mesma natureza, e que as mesmas se repetem nos termos do mesmo modelo, existe maior probabilidade de apresentar problema de multicolinearidade e *overiffting*. Desta forma, foi selecionado o segundo melhor modelo (FIGURAS 3.1 e 3.2, ANEXO I), cuja simplificação também resultou num modelo linear (Expressão 10). O modelo foi ajustado usando os mesmos critérios e submetido aos mesmos procedimentos resultando na remoção do segundo termo ($\beta_0' T_3$) devido a multicolinearidade. Após o ajuste todos os coeficientes de regressão foram significativos a 99% de probabilidade (β_0 = -146.3049, β_1 = -5.7122, β_2 = 4.6312, β_3 = 7.4755, β_4 = -0.3334 e β_5 = -91.9841). A equação incluiu as seguintes variáveis: T3 – Tipo florestal 3 (Mopane); CIRE – *Chlorophyll Index Red Edge*; bio11 – temperatura média do trimestre mais frio em °C; NTVg – *Non tree vegetation* (percentagem de cobertura de vegetação não arbórea) em %; CVI – *Chlorophyll Vegetation Index* (Índice de clorofila de vegetação otimizado e VV – co-polarization VV, média entre *ascending* e *descending*. Neste caso a PG teve a sua importância na geração da estrutura do modelo, combinação e seleção das variáveis que melhor explicam o AGC. A inclusão da variável *dummy* (T3) no modelo revela que a PG foi capaz de agrupar dois tipos florestais semelhantes e um distinto.

$$C = \beta_0 + \beta_0' T_3 + \beta_1 T_3 CIRE + \beta_2 CIRE^2 + \beta_3 bio11 + \beta_4 NTVg + \beta_5 \frac{CVI}{VV} \quad (10)$$

Na Tabela 2.3 estão apresentadas informações referentes às métricas dos modelos selecionados de GARF e PG para treino e validação. O erro médio das estimativas é maior para a base de validação que a base de treino. Na base de treino o desempenho de ambos métodos para RMSE, nRMSE e MBE foi similar, porém a PG foi superior quanto ao MBE com valor muito próximo zero (MBE: 2.73×10^{-17} – treino e 1.464 – validação). Os valores de MBE para ambos métodos mostram que os resíduos das estimativas são praticamente nulos, embora o GARF tenha uma ligeira tendência a subestimar.

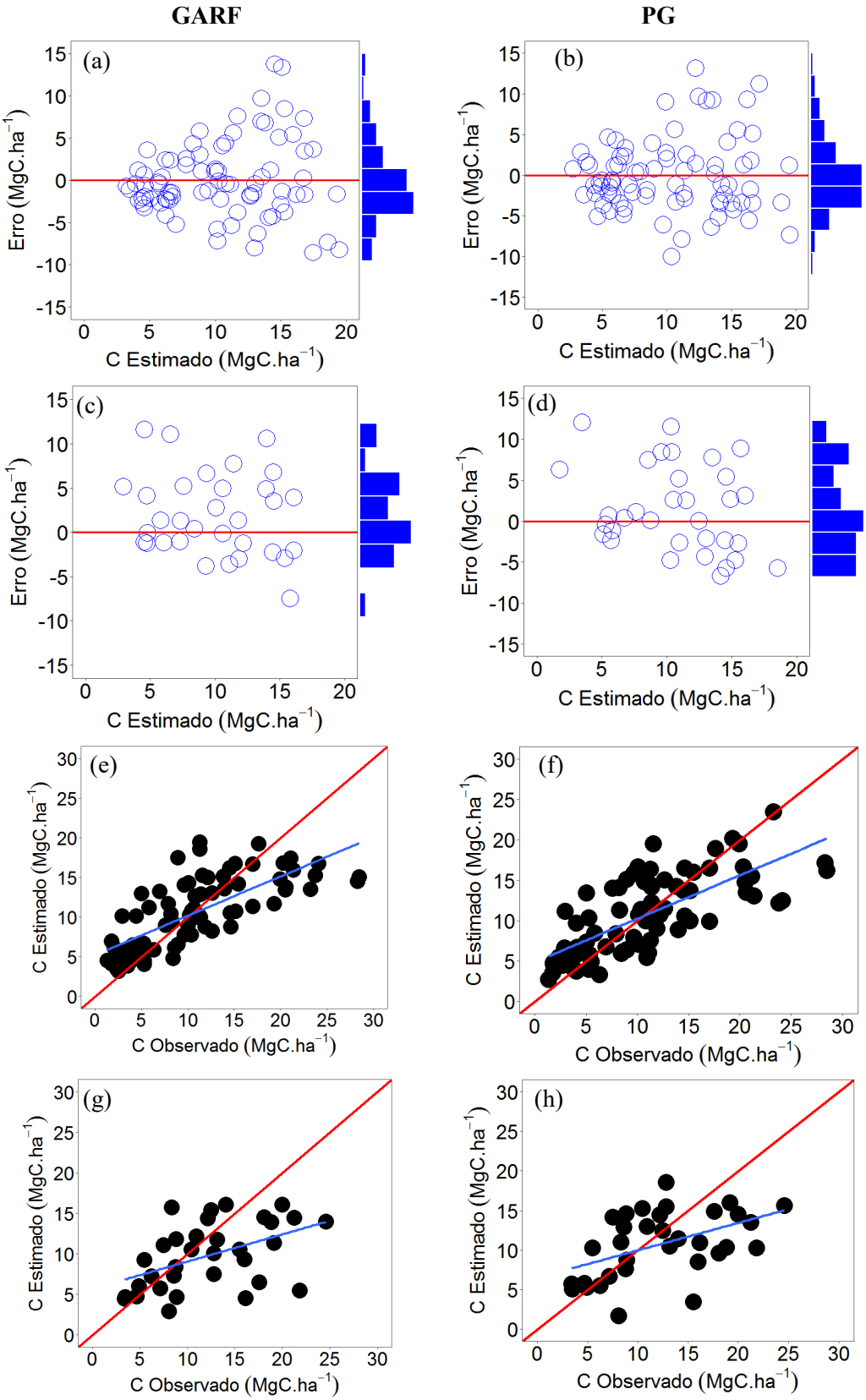
Tabela 2.2 – Métricas de avaliação do desempenho dos modelos de GARF e PG.

	Método	**RMSE ($MgC.ha^{-1}$)**	**nRMSE**	**MBE ($MgC.ha^{-1}$)**	**MAE ($MgC.ha^{-1}$)**
Treino	GARF	4.478	0.427	0.081	3.403
	PG	4.499	0.428	2.731×10^{-17}	3.451
Validação	GARF	5.685	0.466	2.401	4.268
	PG	5.386	0.442	1.464	4.318

Fonte: Do autor (2023)

As análises da dispersão gráfica dos resíduos, paras ambos modelos, mostraram uma distribuição equilibrada entre os valores (-7.5 a 7.5 $MgC.ha^{-1}$) na base de treino, embora o modelo de GARF obteve uma leve subestimativa acima de 17 $MgC.ha^{-1}$ (FIGURA 2.4a, b). Na base de validação, o desempenho em ambos modelos foi relativamente baixo, porém aceitável, com uma tendência geral de subestimativa, além de superestimar em valores de AGC abaixo de 5 $MgC.ha^{-1}$ e subestimar em valores a partir de 15 $MgC.ha^{-1}$ (FIGURA 2.4c, d). Os histogramas de distribuição de resíduos associados aos gráficos de dispersão aproximam-se mais a distribuição normal na base de treino que na base de validação. Portanto, de modo geral, a PG mostrou melhor distribuição em relação ao GARF. Quanto aos gráficos de tendência, pode-se observar uma maior aderência a linha de identidade, principalmente na base de treino. Portanto, é possível também notar claramente que o GARF tende a superestimar em valores menores e subestimar em valores maior em relação a PG (FIGURA 2.4e, f, g, h). Todavia, considerando que se trata de uma floresta natural e a natureza das variáveis explicativas, o comportamento gráfico dos modelos é aceitável, pois pode-se aferir que uma parte da variação da variável resposta é explicada por fatores externos.

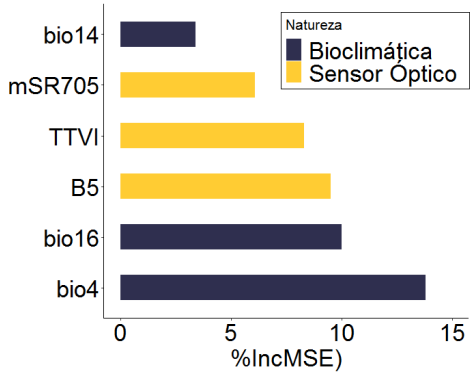
Figura 2.6 – Dispersão gráfica e histograma de distribuição dos resíduos (a, b, c, d) e relação entre AGC observado e estimado (e, f, g, h) para modelos de GARF e PG para a base de treino (a, b, e, f) e validação (c, d, g, h).



Fonte: Do autor (2023).

O RF tem a particularidade de fornecer informação sobre a importância relativa das variáveis, indicando a contribuição de cada variável na melhoria da solução. A ordem de importância de variáveis no modelo de GARF foi a seguinte: bio4, bio16, B5, TTVI, mSR705 e bio14, com percentagem de melhoria de MSE (%IncMSE) de 13.8, 13.4, 10.0, 9.5, 8.3, e 6.1, respetivamente (FIGURA 2.6). Pode-se verificar que a bio4 e B5 foram umas das variáveis mais selecionadas pelo modelo GARF e também fazem parte das variáveis mais importantes no modelo. Porém, nem sempre as variáveis mais selecionadas foram as mais importantes no modelo (exemplo mSR705), contudo a recorrente seleção (frequência de seleção: 90%) mostra sua pertinência para modelos de GARF. Por outro lado, uma das variáveis menos selecionadas (TTVI) teve relativamente maior importância que mSR705 e bio14, e esta última teve menor frequência (3.3%,) e menor importância (6.1), revelando a sua menor relevância para modelos de GARF, ou seja, a sua entrada no modelo deve-se a natureza estocástica do método. As variáveis bioclimáticas (bio4 e bio6) tiveram maior contribuição na melhoria de RMSE em relação as variáveis de sensores.

Figura 2.7 – Importância das variáveis pelo modelo GARF.



Fonte: Do autor (2023)

O modelo gerado pela PG e simplificado inclui sinais de adição (2), subtração (3) e divisão (1), função matemática do tipo exponencial (1), constantes (6) e variáveis (6). O nível de complexidade matemática destes elementos é o seguinte: (*i*) adição, subtração constantes e variáveis – complexidade 1; (*ii*) divisão – complexidade 2 e; (*iii*) exponencial – complexidade 4. No entanto a complexidade do modelo resultou da soma das complexidades (23) de cada usado. No final, buscou-se perceber a influência de cada variável dependente na variável resposta, através da análise de sensitividade das variáveis. A sensibilidade indica a direção e a magnitude de correlação entre as variáveis entrada e a variável de saída. A variável independente com maior impacto na variável resposta foi a CVI (sensitividade: 4.481) e a menos influente foi a NTVg (sensitividade: 0.014), tanto para treino quanto para validação.

Variáveis CIRE, bio11, CVI, VV tem magnitude positiva na variável de interesse e T3 e NTVg tem magnitude negativa, sendo que essas últimas variáveis estão relacionadas com a fisionomia e cobertura vegetal. Alta percentagem de cobertura vegetal não arbórea está relacionada com baixos estoques de carbono. O CIRE e CVI são índices de vegetação que expressam o conteúdo de clorofila na vegetação, portanto, a sua magnitude mostra que maior quantidade de clorofila na vegetação está relacionada com maior biomassa. Porém, o CIRE foi associado à variável T3 (com magnitude negativa), para separar os tipos florestais. Esse padrão pode estar vinculado a maior quantidade de vegetação verde na regeneração, principalmente em áreas exploradas ou no Mopane de baixa estatura, associada às características de solo e perturbação antrópica. Baixa temperatura é um fator limitante ao crescimento da vegetação, razão pela qual a variável bio11 (temperatura média do trimestre mais frio) tem magnitude positiva, pois o aumento dessa variável cria condições para o crescimento vegetal na época fria. A variável do RADAR (VV) está relacionada com a estrutura da copa das árvores, e estas estão positivamente correlacionadas com o acúmulo de biomassa e carbono.

Tabela 2.3 – Análise de sensitividade e magnitude das variáveis de entrada no modelo gerado pela PG.

Base de dados	Variáveis	Sensitividade	% Positivo	Magnitude Positivo	% Negativo	Magnitude Negativo
Treino	T3	0.261	0.00	0.000	100	0.261
	CIRE	0.279	93.83	4.078	6.17	0.232
	bio11	0.196	100	0.196	0.00	0.000
	NTVg	0.014	0.00	0.000	100	0.014
	CVI	4.481	100	4.481	0.00	0.000
	VV	0.378	100	0.378	0.00	0.000
Validação	T3	0.237	0.00	0.000	100	0.237
	CIRE	0.310	100	0.310	0.00	0.000
	bio11	0.217	100	0.217	0.00	0.000
	NTVg	0.015	0.00	0.000	100	0.015
	CVI	3.396	100	3.396	0.00	0.000
	VV	0.402	100	0.402	0.00	0.000

Fonte: Do autor (2023).

4. DISCUSSÃO

A seleção de variáveis preditoras é um passo crítico na modelagem de biomassa/carbono, principalmente em florestas tropicais secas, caracterizadas por alta heterogeneidade. Os métodos GARF e PG foram eficientes na seleção de variáveis, reduzindo o tamanho da base de dados em cerca de 95%, economizando desta forma o esforço computacional na modelagem e ou predição do AGC. Por outro lado, demostraram que o tipo de variáveis selecionadas varia em função do método. O GARF limitou-se na seleção de variáveis do sensor óptico (Sentinel-2 MSI) e bioclimáticas, enquanto a PG abrangeu todas as naturezas (FIGURA 2.4). É importante ressaltar que as variáveis do sensor óptico selecionados por GARF (B12, B5, mSR705, SARVI e TTVI) e pelo PG (CIRE e CVI) são índices de clorofila e/ou estão relacionados com as bandas do *Red edge*, com exceção de B12 e SARVI, e são diferentes dos índices de vegetação comumente usados (ex. SAVI e NDVI). O *Red edge* localiza-se na faixa entre Vermelho e Infravermelho Próximo (NIR), onde a assinatura espectral da vegetação altera bruscamente, tornando-o sensitivo a pequenas mudanças na estrutura da copa ou no conteúdo de clorofila (IMRAN et al., 2020; SINGH et al., 2022). Este resultado corrobora com o observado por Jiang et al. (2022) ao combinar dados do Sentinel-2 e do ICESat-2 para mapear a biomassa acima do solo (*Above Ground Biomass* – AGB) em florestas naturais na China, usando *Extreme Learning Machine* (ELM). Este cenário destaca a importância destas bandas na modelagem de carbono acima do solo (*Above Ground Carbon* – AGC) e a vantagem do Sentinel-2 MSI em relação a outros sensores ópticos. As variáveis bioclimáticas selecionadas estão relacionados com a temperatura (bio1, 4, 6 e 11) e precipitação (bio14 e 16). Essas variáveis constituem fatores limitantes ao crescimento vegetal, sendo que um aumento no valor dessas variáveis é favorável a acumulação de carbono e vice-versa. A amplitude de temperatura (bio7), precipitação anual (bio12) e radiação solar foram destacados como fatores que mais influenciam a distribuição geográfica da vegetação de Mopane na África Austral (NGAREGA, MASOCHA; SCHNEIDER, 2021). O trabalho destes autores foi desenvolvido numa escala regional, o que pode originar diferenças nas variáveis selecionadas, em relação ao presente estudo que foi realizado em escala local.

Além das variáveis bioclimáticas e de sensor óptico, os modelos de PG também incluíram variáveis do Radar (Sentinel-1 SAR) referentes aos coeficientes de retro espalhamento (VV e VH). Esse padrão foi comprovado por autores como Chen et al. (2018), Gara et al. (2023); Ghosh, Behera e Paramanik (2020), Macave et al. (2022) e Tavasoli e Arefi (2021) que demostraram essa relação positiva com altura da copa e consequentemente com a biomassa/carbono. Outro aspecto relevante em relação a PG, foi a seleção da variável de

cobertura florestal (*Non tree vegetation percent - NTVg*), proveniente de MODIS e do tipo florestal (T3, variável *dummy*). Estas variáveis estão relacionadas à fisionomia e à distribuição espacial da vegetação, sendo que ambas apresentam uma sensitividade negativa em relação ao AGC (TABELA 2.4). Um aumento nos valores destas variáveis resulta na redução do AGC. Obviamente, uma maior percentagem de vegetação não arbórea (*NTVg*) está associada a menor estoque de carbono. A ativação desta variável pode estar relacionada com os estratos vegetais, Mata Mista e Mopane (T3), que possuem extensas áreas de floresta aberta e com vegetação arbustiva. O estrato vegetacional de Mopane também foi ativado no modelo de PG como variável *dummy* (T3) e está associado ao índice de clorofila CIRE. Esta associação pode também estar relacionada com alta taxa de regeneração observada, principalmente em Nwamandzele (Mabalane), onde houve exploração seletiva para produção do carvão vegetal, assim como pode estar associada a áreas com Mopane de baixa estatura (arbustos menores que 2 m) encontradas em ambos sites de estudo. Woollen et al., (2016) verificaram menor estoque de biomassa (AGB = 7.31 $Mg.ha^{-1}$) no Mopane arbustivo em Mabalane. As áreas com alta regeneração e com Mopane arbustivo podem apresentar valores altos de índice de clorofila (CIRE), porém, com menor valor de AGC. Neste caso, a variável T3 (*dummy*), com sensitividade negativa, foi combinada com CIRE para reduzir e equilibrar os estoques de AGC. Ghosh; Behera e Paramanik (2020) também relataram a sensitividade negativa da fração de cobertura vegetal, na modelagem da altura de copa em manguezais, sendo que a altura da copa está positivamente correlacionada ao acúmulo de AGC. Por outro lado, a ativação da variável de cobertura (NTVg) e do tipo florestal (T3) pode também estar relacionada com a influência da percentagem de cobertura vegetal e dos tipos de vegetação nos estoques de carbono. Vários autores relataram diferenças significativas de AGC entre diferentes percentagens de cobertura e/ou tipos florestais e sublinharam a importância de estratificação na melhoria da precisão das estimativas (CARREIRAS; MELO; VASCONCELOS, 2013; GHOSH; BEHERA; PARAMANIK, 2020; MACAVE et al., 2022; QIAN et al., 2021; SILVEIRA et al., 2019). Gara et al. (2023) testaram combinação de dados de Landsat e ALOS PALSAR para estimar AGC em floresta de Mopane, no Norte de Zimbabwe, usando ANN e verificaram que o nRMSE reduziu de 16% (só ALOS PALSAR) e 14% (só Landsat) para 12% (combinação). De modo geral, a combinação de dados de diferentes fontes gera melhores estimativas em relação ao uso de dados provenientes de uma única fonte (CARVALHO et al., 2022; GHOSH; BEHERA; PARAMANIK, 2020; MACAVE et al., 2022; TAVASOLI; AREFI, 2021).

Na literatura, nota-se que na maioria dos trabalhos de modelagem via RF com redução de variáveis, a seleção foi feita pelo método de *stepwise*, com adição ou remoção recursiva,

baseada na importância relativa das variáveis dada pelo RF (DANG et al., 2019; JIANG et al., 2022; LI et al., 2019; SILVEIRA et al., 2019), sendo que o resultado é influenciado pela ordem na qual as variáveis são adicionadas no processo, além de ser oneroso e exigir muito esforço. Neste estudo, o método GARF mostrou-se ser eficiente na seleção de variáveis e na modelagem (TABELA 2.3). A abordagem híbrida do método de GARF visa melhorar o desempenho da predição do RF, reduzindo o RMSE ao mesmo tempo que garante a redução das variáveis. O Algoritmo Genético (AG) guia a busca de soluções ótimas, entre várias possibilidades testadas a cada iteração, selecionando variáveis com maior potencial na melhoria da solução do RF. A potencialidade da estratégia GARF para seleção de variáveis e modelagem de biomassa/carbono já foi comprovada por Carvalho et al. (2022) e Tavasoli e Arefi (2021). Carvalho et al. (2022) aplicaram modelo de GARF para modelar AGC em floresta tropical na Bacia do Rio Grande (MG, Brasil), onde obtiveram melhor desempenho em relação a remoção recursiva de variáveis. Com o GARF, os autores reduziram o tamanho da base de dados também em cerca de 95%, gerando estimativas com precisão aceitável (RMSE = 17.75 $MgC.ha^{-1}$). Tavasoli e Arefi (2021) também reportaram ganhos na predição de AGB usando o método de GARF, em termos de redução de número de variáveis, precisão e desempenho da modelagem. Outra aplicação do método GARF no setor florestal foi na hipsometria, feita por Miranda et al. (2022) que também comprovaram a sua eficiência na seleção de variáveis e modelagem. A PG já foi aplicada em diferentes áreas de conhecimento (LONDHE et al., 2022; MEHR e KAHYA, 2017; ZHOU et al., 2023), porém a sua aplicabilidade na modelagem florestal, incluindo biomassa/carbono, ainda não foi muito estudada, principalmente quando se trata da regressão simbólica. A regressão simbólica, procura selecionar variáveis, operadores matemáticos e booleanos, funções matemáticas assim como estimativa dos parâmetros para o modelo construído. Do conjunto de funções estabelecidas na PG (regressão simbólica), foram selecionados os operadores de adição, multiplicação, subtração, divisão e exponencial, e os terminais foram compostos por seis variáveis e seis constantes. Foi gerado um modelo com estrutura linear, complexidade 23, considerando o critério usado por Aryadoust (2015). O presente estudo demostrou a potencialidade da PG na modelagem de AGC (TABELA 2.3). A PG fornece uma função visível e facilmente interpretável pelos usuários, com variáveis e coeficientes definidos. Essa particularidade faz da PG um método prático e replicável para outras áreas. O Ghosh; Behera e Paramanik (2020) usando dados de Sentinel-1 SAR e Sentinel-2 MSI testaram o desempenho de RF e PG (regressão simbólica) na predição da altura da copa de manguezais. Os autores constataram que a PG gerou melhores resultados (RMSE=1.48 m; R^2=0.62) em relação a RF (RMSE=1.57 m; R^2=0.6) e o modelo gerado pela PG foi linear com

complexidade 32. Após a modelagem os autores também analisaram a sensitividade das variáveis onde a Fração de Cobertura Vegetal (FVC) teve a maior sensitividade (1.22, negativa), seguida de LAI (1.108, positiva), DEM (0.34, positiva), coerência (0.57, negativa) e VH (0.02177, positiva). Outro estudo envolvendo a regressão simbólica no setor floresta foi do Fernández-Carrillo et al. (2022) que testaram o desempenho da PG, *Gaussian Regression Process (PGR), Category Boosting (CatBoost) e Artificial Neural Networks (ANN)* em funções de afilamento de *Tectona grandis*, comparando com os modelos convencionais de Fang 2000 e Kozak 2004. Os autores verificaram que os modelos de Kozak 2004 e ANN foram superiores, não obstante, a PG foi superior ao modelo de Fang 2000 e mostrou-se potencial na modelagem da forma das árvores. Cabral et al., (2018) testaram a potencialidade da PG na classificação de áreas queimadas usando imagens satélites em savanas tropicais (estudo de caso de Brasil, Guiné-Bissau e Congo Democrático), em comparação com as metodologias clássicas (Máxima verossimilhança - MaxVer e *Classification and Regression Trees* - CART) e os resultados mostraram que a PG alcançou melhor acurácia na maioria dos casos.

A comparação dos resultados do presente estudo com outros realizados em áreas similares é limitada devido à menor disponibilidade de estudos de biomassa/carbono conduzidos na floresta de Mopane em Moçambique e regiões vizinhas. Os valores médios de AGC (em $MgC.ha^{-1}$) observados no campo e estimados por GARF e PG para cada tipo florestal foram: Mata Mista (11.701 – campo, 10.508 – GARF e 11.265 – PG), Floresta de Mecrusse ou *Androstachys johnsonii* (14.733 – campo, 14.343 – GARF e 14.440 – PG) e Floresta de Mopane (7.642 –campo, 7.848 – GARF e 7.196 – PG). O modelo de PG apresentou valores mais próximos aos valores observados, quando comparado com GARF, comprovando a sua superioridade. No estudo conduzido por Woollen et al., (2016), na floresta de Mopane em Mabalane (Moçambique), foram reportados valores médios de AGB, em $Mg.ha^{-1}$, por tipo florestal de 31.7±2.5 (Mecrusse), 11.8±1.6 (Mopane), Mopane arbustivo (7.31±1.31) e 5.4±1.38 (Mista). Gara et al. (2023) ao estimar AGC na floresta de Mopane no Norte de Zimbabwe, usando imagens de Landsat e ALOS PALSAR aplicando ANN, obtiveram valores de AGC que variam de 17 a 32 $MgC.ha^{-1}$ com nRMSE de 0.12 (12%). Macave et al. (2022) usaram Sentinel-1 e Sentinel-2 e ALOS PALSAR, usando regressão para estimar ABG na floresta de Miombo na Reserva Nacional de Niassa (Moçambique) usando regressão linear múltipla. Os valores de AGB estimados variaram de 0.6 a 200 Mg/ah, com média de 63+/- 20.3 $Mg.ha^{-1}$ e nRMSE de 20.46%. A precisão das estimativas do presente estudo tanto para GARF (nRMSE: 42.7% treino e 46.6% - validação) como para PG (nRMSE: 42.8% - treino e 44.2% - validação) diferem dos resultados obtidos por Gara et al., (2023) e MACAVE et al., (2022),

com nRMSE de 12% e 20.3%, respetivamente. Porém, estes estudos foram conduzidos dentro de áreas protegidas, onde a influência antrópica é menor, o que pode ter contribuído para melhor precisão das estimativas. Adicionalmente, Pelletier et al. (2017) demostraram que as áreas protegidas na Zâmbia são importantes para acumulação de biomassa/carbono. Por outro lado, se os mesmos modelos forem aplicados em áreas perturbadas, como é o caso do presente estudo, podem não apresentar erros maiores, pois não formam desenvolvidos para tal condição. Portanto o presente estudo constitui uma contribuição para estimativa da AGC em áreas perturbadas. Em ambos sites de estudo foram notados vestígios de perturbação antrópica por agricultura, pastoreio e extrativismo, embora com maior intensidade em Mabalane, devido a exploração dos indivíduos de maior diâmetro para produção de carvão vegetal. Estes e outros fatores propiciam também a ocorrência de altas taxas de regeneração e de gramíneas, aumentando o risco de ocorrência de incêndios (WOLLEN et al., 2016). O fogo causa mudanças significativas na estrutura da floresta de Mopane e nos atributos morfológicos (DAP, H, diâmetro da copa, e número de fustes). Áreas de Mopane queimadas com frequência apresentam indivíduos com altura menor e mais ramificados (KENNEDY; POTGIETER, 2003; TAMENE, 2016; STEVENS, 2021).

A precisão obtida no presente estudo é aceitável considerado a heterogeneidade da floresta em termos de estoques de AGC e a perturbação antrópica. Comparando os resultados obtidos com outros estudos realizados em condições similares (heterogeneidade), verifica-se semelhança no comportamento do erro, como por exemplo no estudo de Li et al. (2019) na China que obtiveram nRMSE variando de 47% a 62% sem a estratificação e 31% a 61% com estratificação por tipo de floresta. Silveira et al. (2019), estimando AGB em savanas arbóreas tropicais (Brasil) obtiveram valores de MAE percentual variando de 41.50 a 31.51%. Carvalho et al. (2022) usando o método de GARF para estimativa de AGC em floresta nativa no Brasil observou nRMSE de 35.56 % para treino e 37% para validação.

5. CONCLUSÃO

O método híbrido entre Algoritmo Genético - AG e *Random Forest* - RF (GARF) e a Programação Genética (PG) apresentam potencial para modelagem do carbono acima de solo (AGC), como alternativa aos métodos clássicos, e robustez na seleção de variáveis explicativas à variação dos estoques de carbono. O presente estudo demostrou que é possível predizer e monitorar estoques de carbono em florestas naturais tropicais, usando variáveis de fácil obtenção e de baixo custo. A combinação entre dados de diferentes fontes ajuda a captar informações, ampliando a capacidade preditiva do método. Contudo, devido à natureza estocástica dos métodos, detectou-se um padrão diverso de seleção das variáveis, embora isso não tenha prejudicado o resultado final das predições. O método GARF adere-se mais a variáveis bioclimáticas e de sensores ópticos (sentinel-2), enquanto a PG via regressão simbólica, combina variáveis independente de sua natureza, apresentando um grande potencial na geração de modelos de efeito misto a segmentados. A exatidão das estimativas geradas pelos métodos é aceitável, principalmente considerando a heterogeneidade da floresta, dado que o MBE é muito próximo de zero. Embora prevaleça o problema da superestimativa em valores menores e subestimativas em valores maiores de AGC, a ativação da variável *dummy* (tipo florestal) no modelo de PG contribuiu para redução das subestimativas em valores maiores. Os resultados revelaram, para ambos métodos, a importância das bandas do *Red edge e SWIR* do Sentinel-2 na modelagem biofísica, devido a sua sensitividade a variação de conteúdo de clorofila na vegetação, e das variáveis bioclimáticas relacionadas à temperatura. Para PG, a percentagem de cobertura vegetal (produtos de MODIS) e variáveis de RADAR (Sentinel-1) tem potencial na predição de AGC por fornecerem informações de ocupação do solo e relacionada com a copa das árvores. A PG via regressão simbólica é mais prática em relação ao método GARF, por fornecer um modelo com estrutura visível e facilmente replicável a outras áreas. Contudo, os modelos gerados precisam de simplificação e de análise de um especialista antes da sua aplicação, complementando o uso da inteligência artificial no processo autônomo de geração de funções.

REFERENCIAS

ARJASAKUSUMA, S.; KUSUMA, S. S.; PHINN, S. Evaluating Variable Selection and Machine Learning Algorithms for Estimating Forest Heights by Combining Lidar and Hyperspectral Data. **International Journal o Gero-Information**. v. 9, n. 9, p. 507, 2020.

BEHERAA S. K., et al. Aboveground biomass and carbon stock assessment in Indian tropical deciduous forest and relationship with stand structural attributes. **Ecological Engineering**. v. 99, p. 513–524, 2016.

BIAU, G. Analysis of a Random Forests Model. **Journal of Machine Learning Research**. v. 13, p. 1063 – 1095, 2012.

BONAN, G.B. Forests And Climate Change: Forcings Feedbacks, and The Climate Benefits of Forests. **Science.** v. 320, p. 1444–1449, 2008.

BREIMAN, L. Random Forests. **Machine Learning**. v. 45, p. 5 – 31, 2001.

BUNSTER, J. **52 Madeiras de Moçambique: Catálogo Tecnológico.** Universidade Eduardo Mondlane. Faculdade de Agronomia e Engenharia Florestal. Departamento de Engenharia Florestal, 1995.

CABRAL, A. I. R. et al. Burned area estimations derived from Landsat ETM+ and OLI data: Comparing Genetic Programming with Maximum Likelihood and Classification and Regression Trees. **ISPRS Journal of Photogrammetry and Remote Sensing.** V. 142, p. 94-105, 2018.

CARREIRAS, J. M. B.; MELO, J. B.; VASCONCELOS, M. J. Estimating the Above-Ground Biomass in Miombo Savanna Woodlands (Mozambique, East Africa) Using L-Band Synthetic Aperture Radar Data. **Remote Sensing**. v. 5, p. 1524-1548, 2013.

CARVALHO, M. C. et al. Algoritmos de aprendizagem de máquina na modelagem da distribuição potencial de habitats de espécies arbóreas. **Nativa**. v. 7, n. 5, p. 600-606, 2019.

CARVALHO, M. C. et al. Modeling Ecological Niche of Tree Species in Brazilian Tropical Area. **Cerne**. v. 23, n. 2, p. 229-240, 2017.

CARVALHO, M.C. et al. Data mining applied to feature selection methods for aboveground carbon stock modelling. **Pesquisa Agropecuária Brasileira**. v. 57, 2022.

CERASOLI, F.; D'ALESSANDRO, P. BIONDI, M. Worldclim 2.1 versus Worldclim 1.4: Climatic niche and grid resolution affect between-version mismatches in Habitat Suitability Models predictions across Europe. **Ecology an Evolution**. v. 12, n. 2, p. 1-20, 2022.

CHAVE, J. et al. Improved allometric models to estimate the aboveground biomass of tropical trees. **Global Change Biology. v.** 20, n. 10, p. 3177-3190, 2014.

CHEN, L.; et al. Estimation of Forest Above-Ground Biomass by Geographically Weighted Regression and Machine Learning with Sentinel Imagery. **Forests**. v. 9, p. 582, 2018.

CIAIS, P., C. et al. **Carbon and Other Biogeochemical Cycles**. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.). Cambridge: Cambridge University; New York, 2013.

COLEY, D. **An Introduction to Genetic Algorithms for Scientists and Engineers**. Singapore: World Scientific. 223p, 1999.

CUTLER, A., CUTLER, D.R., STEVENS, J.R. **Random Forests**. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, Boston, MA. pp. 157-175, 2012.

DANG, S. N. et al. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. **Ecological Informatics**. v. 50, p. 24-32, 2019.

DE SOUSA, V. J. Dependence, pressure and recovery of forest resources in Limpopo National. **Cadernos de Geografia**. n. 44. pp. 21-35, 2021

DEB, K. **Multi-Objective Optimization using Evolutionary Algorithms**. Chichester: John Wiley. 497p, 2001.

DIMICELI, C. M., M. L. et al. **Annual global automated MODIS vegetation continuous fields (MOD44B) at 250 m spatial resolution for data years beginning day 65, 2000–2014, collection 5 percent tree cover, version 6**. University of Maryland, College Park, MD, USA, 2017.

FERNÁNDEZ-CARRILLO, V. H. et al. Do AI Models Improve Taper Estimation? A Comparative Approach for Teak. Forests. 2022, 13, 1465.

FICK, S. E.; HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**. v. 37, n. 12, p. 4302-4315, 2017.

FISCHER, R. The Long-Term Consequences of Forest Fires on the Carbon Fluxes of a Tropical Forest in Africa. **Applied Sciences**. Basel: MPDI. v. 11, n. 4696, 2021.

FOX J.; WEISBERG, S. **An R Companion to Applied Regression**. 3ª Ed. Sage, Thousand Oaks CA, 2019.

GARA, T. W. et al. Integrating RADAR and optical imagery improve the modelling of carbon stocks in a mopane-dominated African savannah dry forest. **African Journal of Ecology**. p.1–10, 2023.

GHOSH, S.M.; BEHERA, M.D.; PARAMANIK, S. Canopy Height Estimation Using Sentinel Series Images through Machine Learning Models in a Mangrove Forest. **Remote Sensing**. v. 12, 2020.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. New York: Addison-Wesley. 412p, 1989.

GOU, Y.; RYAN, C.M.; REICHE, J. Large Area Aboveground Biomass and Carbon Stock Mapping in Woodlands in Mozambique with L-Band Radar: Improving Accuracy by Accounting for Soil Moisture Effects Using theWater Cloud Model. **Remote Sensing**. v. 14, p. 404, 2022.

GUEDES, B. S.; SITOE A. A.; OLSSON, B. A. Allometric models for managing lowland miombo woodlands of the Beira corridor in Mozambique. **Global Ecology and Conservation**. v. 13, p. e00374, 2018.

HARMSE, C.J.; GERBER, H.; VAN NIEKERK, A. Evaluating Several Vegetation Indices Derived from Sentinel-2 Imagery for Quantifying Localized Overgrazing in a Semi-Arid Region of South Africa. **Remote Sensing**. v. 14, p. 1720, 2022.

HAUPT, R. L.; HAUPT, S. H. **Practical Genetic Algorithms**. 2nd ed. New Jersey: John Wiley. 253p, 2004.

HOLZMAN, B. A. **Tropical forest biomes.** In, WOODWARD, S. L. (Ed.). Greenwood guides to biomes of the world. London: Greenwood. 242p, 2008.

IMRAN, H.A. et al. VIS-NIR, Red-Edge and NIR-Shoulder Based Normalized Vegetation Indices Response to Co-Varying Leaf and Canopy Structural Traits in Heterogeneous Grasslands. **Remote Sensing**. v. 12, n. 14, 2020.

JIANG, F. et al. Improving aboveground biomass estimation of natural forests on the Tibetan Plateau using spaceborne LiDAR and machine learning algorithms. **Ecological Indicators**. v. 143, 2022.

KENNEDY, A.D.; POTGIETER, A.L.F. Fire season affects size and architecture of Colophospermum mopane in southern African savannas. Plant Ecology. v. 167, p. 179–192, 2003.

KOZA, J. R. **Genetic Programming: Automatic Synthesis of Topologies and Numerical Parameters**. In: GLOVER, J.; KOCHENBERGER, G; A (Ed). Handbook of Metaheuristics. New York: Kluwer Academic. 556p, 2003.

KOZA, J. R. **Genetic programming: on the programming of computers by means of natural selection**. Namco, 1992.

LI, Y. et al. Influence of Variable Selection and Forest Type on Forest Aboveground Biomass Estimation Using Machine Learning Algorithms. **Forests**. v. 10, n. 12, 2019.

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

LISBOA, S.N. et al. Biomass allometric equation and expansion factor for a mountain moist evergreen forest in Mozambique. **Carbon Balance Manage**. v. 13, 2018.

LONDHE, S. et al. Tree Based Approaches for Predicting Concrete Carbonation Coefficient. ***Applied Sciences***. v. 12, n. 8, p. 3874, 2022

LUKASZ KOMSTA. **outliers: Tests for Outliers**. R package version 0.15, 2022. https://CRAN.R-project.org/package=outliers

MACAVE, O. A. et al. Modelling Aboveground Biomass of Miombo Woodlands in Niassa Special Reserve, Northern Mozambique. **Forests**. V. 13, p. 311, 2022.

MAGALHÃES, T. M.; SEIFERT, T. Estimation of Tree Biomass, Carbon Stocks, and Error Propagation in Mecrusse Woodlands. **Open Journal of Forestry**. v. 5, p. 471-488, 2015.

MAGNUSSEN, S., REED, D. **Modelling for Estimation and Monitoring**. Chapter in National Forest Assessment Knowledge Reference. FAO, Rome, Italy, 2004.

MAKHADO, R. A. et al. Factors influencing the adaptation and distribution of *Colophospermum mopane* in southern Africa's mopane savannas – A review. **Bothalia**. v. 44, n. 152, 2014.

MALHI, Y. The productivity, metabolism and carbon cycle of tropical forest vegetation. **Journal of Ecology**. v.100, p.65–75, 2012

MAQUIA, I. et al. Diversification of African Tree Legumes in Miombo–Mopane Woodlands. **Plants**. n. 8, p. 182, 2019.

MARTINS SILVA, J. P. et al. Computational techniques applied to volume and biomass estimation of trees in Brazilian savanna. **Journal of Environmental Management**. v. 249, p 1 -12, 2019.

MATE, R.; JOHANSSON, T.; SITOE, A. Biomass Equations for Tropical Forest Tree Species in Mozambique. **Forests**. v. 5, n. 3, p. 535-556, 2014.

MEHR, A. D.; KAHYA, E. A Pareto-optimal moving average multigene genetic programming model for daily streamflow prediction. **Journal of Hydrology**. v. 549, p. 603-615, 2017.

MINISTÉRIO DE TERRA, AMBIENTE E DESENVOLVIMENTO RURAL – MITADER. **Inventário Florestal Nacional**. Relatório Final. Direcção Nacional de Florestas. Maputo, 2018.

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

MITCHELL, M. **An Introduction to Genetic Algorithms**. Cambridge: Massachusetts Institute of Technology. 158p, 1996.

MONTERO, D., AYBAR, C., MAHECHA, M. D., WIENEKE, S. (2022). spectral: Awesome Spectral Indices deployed via the Google Earth Engine JavaScript API. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**. Volume XLVIII-4/W1-2022. Free and Open Source Software for Geospatial (FOSS4G) 2022 Academic Track. Florence, Italy. 22-28 August, 2022.

NGAREGA, B. K.; MASOCHA, V. F.; SCHNEIDER, H. Forecasting the effects of bioclimatic characteristics and climate change on the potential distribution of *Colophospermum mopane* in southern Africa using Maximum Entropy (Maxent). **Ecological Informatics**. v. 65, 2021.

ØSTERGAARD, P. et al. C-band SAR for the GMES Sentinel-1 mission. **2011 8th European Radar Conference**. Manchester, UK, pp. 234-240, 2011.

PANDEY, P.K.; PANDEY, V. Development of reference evapotranspiration equations using an artificial intelligence-based function discovery method under the humid climate of Northeast India. **Computers and Electronics in Agriculture**. v. 179, 2020.

PELLETIER, J. et al. Human and natural controls of the variation in aboveground tree biomass in African dry tropical forests. **Ecological Applications**. v. 27, n. 5, pp. 1578–1593, 2017.

PHAM, T. D. et al. Comparison of Machine Learning Methods for Estimating Mangrove Above-Ground Biomass Using Multiple Source Remote Sensing Data in the Red River Delta Biosphere Reserve, Vietnam. **Remote Sensing**. v. 12, n. 1334, p. 1- 24, 2020.

POLI, R.; LANGDON, W. B.; MCPHEE, N. F. **A field guide to genetic programming**. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, (With contributions by J. R. Koza). GPBiB, 2008.

QIAN, C. et al. Estimation of Forest Aboveground Biomass in Karst Areas Using Multi-Source Remote Sensing Data and the K-DBN Algorithm. **Remote Sensing**. v. 13, p. 5030, 2021.

R CORE TEAM. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

REEVES, C. **Genetic Algorithms**. In: GLOVER, J.; KOCHENBERGER, G; A (Ed). Handbook of Metaheuristics. New York: Kluwer Academic. 556p, 2003.

ROTHLAUF, F. **Representations for Genetic and Evolutionary Algorithms**. 2nd ed. New York: Springer. 325p, 2006.

RYAN, C. M.; WILLIAMS, M.; GRACE, J. Above- and Belowground Carbon Stocks in a Miombo Woodland Landscape of Mozambique. **Biotropica**. v. 43, n. 4, p. 423–432, 2011.

SEDANO, F. et al. Monitoring Forest Degradation from Charcoal Production with Historical Landsat Imagery. A Case Study in Southern Mozambique. **Environmental Research Letters**. v. 15, n.1, 2020.

SEDANO, F. et al. The Impact of Charcoal Production on Forest Degradation: A Case Study in Tete, Mozambique. **Environmental Research Letters**. v. 11, n. 9, 2016.

SILVEIRA, E. M. O. et al. Pre-stratified modelling plus residuals kriging reduces the uncertainty of aboveground biomass estimation and spatial distribution in heterogeneous savannas and forest environments. **Forest Ecology and Management**. v. 445, p. 96-109, 2019.

SINGH C. et al. Remote sensing-based biomass estimation of dry deciduous tropical forest using machine learning and ensemble analysis. **Journal of Environmental Management.** v. 308, 2022.

SITOE, A. A.; MANDLATE, L. J. C.; GUEDES, B. S. Biomass and Carbon Stocks of Sofala Bay Mangrove Forests. **Forests**. v. 5, p. 1967-1981, 2014.

STEVENS, N. What shapes the range edge of a dominant African savanna tree, *Colophospermum mopane*? A demographic approach. **Ecology and Evolution**. v. 11, p. 3726–3736, 2021.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; KERRY, R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. **Geoderma**. v. 266, p. 98-110, 2016.

TAMENE, L. et al. Spatial Variation in Tree Density and Estimated Aboveground Carbon Stocks in Southern Africa. **Forests**. v. 7, n. 57, 2016.

TAVASOLI, N.; AREFI, H. Comparison of Capability of SAR and Optical Data in Mapping Forest above Ground Biomass Based on Machine Learning. **Environ. Sci. Proc**. v. 5, n. 13, 2021.

TEXEIRA, K. A. et al. Carbon dynamics of mature and regrowth tropical forests derived from a pantropical database (TropForC-db). **Global Change Biology**. v. 22, p. 1690–1709, 2016.

TORRES, R. et al. GMES Sentinel-1 Mission. **Remote sensing of environment**. v.120, p. 9-24, 2012.

WEN, C. et al. An Object-Based Genetic Programming Approach for Cropland Field Extraction. **Remote Sensing**. v. 14, 1275, 2022.

WIEGAND, P.; PELL, R. COMAS, C. Simultaneous variable selection and outlier detection using a robust genetic algorithm. **Chemometrics and Intelligent Laboratory Systems.** v. 98, p. 108 – 114, 2009.

WOOLLEN, E. et al. Charcoal production in the Mopane woodlands of Mozambique: what are the trade-offs with other ecosystem services? **Philosophical. Transactions. R. Soc. B**. 2016.

XUE, J.; SU, B. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. **Journal of Sensors**. v. 2017, 2017.

YESUF, G.; BROWN, K. A.; WALFOR, N. Assessing regional-scale variability in deforestation and forest degradation rates in a tropical biodiversity hotspot. **Remote Sensing in Ecology and Conservation**. London: John Wiley. v. 5, n. 4, p. 346-359, 2019.

ZHOU, Z.; YANG, Y.; ZHANG, G.; XU, L.; WANG, M. EBM3GP: A novel evolutionary bi-objective genetic programming for dimensionality reduction in classification of hyperspectral data. **Infrared Physics & Technology**. v. 129, 2023.

VAN WYK, B.; VAN WYK. P. **Field Guide to Trees of Southern Africa**. Struik Publishers. Cape Town. 1997. 520p.

BURROWA, J. F. et al. **Trees and Shrubs of Mozambique**. Publisnhint Print Matter. Cape Town. 2019. 1114p.

ARYADOUS, V. Application of evolutionary algorithm-based symbolic regression to language assessment: Toward nonlinear modeling**. Psychological Test and Assessment Modeling**. v. 57, p. 301-337, 2015.

APÊNDICE I

Tabela I. 1– Descrição das variáveis testadas em modelos de GARF e PG.

(Continua)

Tipo de variável	Nome da Variável
Geográfica (Inventário)	**Sites de estudo *(variável dummy)*:** L1 – local 1; L2 – LOCAL 2
Biofísica (World.Clim)	Altitude
Fisionômica	**Tipos florestais *(variável dummy)*:** T1 – Mista (*Guibourtia conjugata, Cumbretum spp e Acacia spp*); T2 – Mecrusse; T3 – Mopane;
Bioclimáticas (World.Clim)	bio1- Temperatura Média Anual; bio2 - Amplitude Diurno Médio (Média mensal (temp-máx – temp-mín.)); bio3- Isotermalidade (BIO2/BIO7) (×100); bio4 - Sazonalidade da Temperatura (desvio padrão ×100); bio5 - Temperatura máxima do mês mais quente; bio6- Temperatura mínima do mês mais frio; bio7 - Amplitude Anual de Temperatura (BIO5-BIO6); bio8 - Temperatura média do trimestre mais úmido; bio9 - Temperatura média do trimestre mais seco; bio10- Temperatura Média do Trimestre Mais Quente; bio11- Temperatura média do trimestre mais frio; bio12 - Precipitação anual; bio13 - Precipitação do mês mais úmido; bio14 - Precipitação do Mês Mais Seco; bio15 - Sazonalidade de Precipitação (Coeficiente de Variação); bio16 - Precipitação do trimestre mais úmido; bio17 - Precipitação do trimestre mais seco; bio18 - Precipitação do trimestre mais Quente; bio19 - Precipitação do trimestre mais frio.
Sensor de RADAR (Sentinel-1)	VV-A – *Single co-polarization, vertical transmit/vertical receive*, *ascending;* VV-D – *Single co-polarization, vertical transmit/vertical receive*, *descending;* VV-A-D – média entre VV-A e VV-D; VH-A –*Dual-band cross-polarization, vertical transmit/horizontal receive, ascending*; VH-D –*Dual-band cross-polarization, vertical transmit/horizontal receive, descending*; VH-A-D – média entre VH-A e VH-D
Cobertura Vegetal (MODIS)	NTVg – *Non Tree Vegetation percent;* NVg – *Non Vegetated percent;* NVg-sd – *Non Vegetated percent standard deviation;* TCover – *Tree Cover percent;* TCover-sd – *Tree Cover percent standard deviation*
Sensor Óptico (Sentinel-2)	**Bandas:** B1 – *Aerossol*; B2 – *Blue*; B3 – *Green*; B4 – *Red* ; B5 – *Red Edge 1*; B6 – *Red Edge 2;* B7 – *Red Edge 3*; B8 – *NIR 1*; B8A – *NIR 2*; B11 – *SWIR 1*; B12 – *SWIR 2*, WVP - *Water vapor* (B9); TCI-B – *True color image Blue*, TCI-G – *True color image Green*, TCI-R – *True color image Red*

Tabela I. 2– Descrição das variáveis testadas em modelos de GARF e PG.

(Conclusão)

Sensor Óptico (Sentinel-2)	**Índices espectrais:** *AFRI1600 - Aerosol Free Vegetation Index (1600 nm); AFRI2100 - Aerosol Free Vegetation Index (2100 nm); ARVI - Atmospherically Resistant Vegetation Index; ATSAVI - Adjusted Transformed Soil-Adjusted Vegetation Index; BCC - Blue Chromatic Coordinate; BNDVI - Blue Normalized Difference Vegetation Index; BWDRVI - Blue Wide Dynamic Range Vegetation Index; CIG - Chlorophyll Index Green; CIRE - Chlorophyll Index Red Edge; CVI - Chlorophyll Vegetation Index; DVI - Difference Vegetation Index; EVI - Enhanced Vegetation Index; EVI2 - Two-Band Enhanced Vegetation Index; GARI - Green Atmospherically Resistant Vegetation Index; GBNDVI - Green-Blue Normalized Difference Vegetation Index; GCC - Green Chromatic Coordinate; GDVI - Generalized Difference Vegetation Index; GEMI - Global Environment Monitoring Index; GLI - Green Leaf Index; GNDVI - Green Normalized Difference Vegetation Index; GOSAVI - Green Optimized Soil Adjusted Vegetation Index; GRNDVI - Green-Red Normalized Difference Vegetation Index; GRVI - Green Ratio Vegetation Index; GSAVI - Green Soil Adjusted Vegetation Index; GVMI - Global Vegetation Moisture Index; IAVI - New Atmospherically Resistant Vegetation Index; IPVI - Infrared Percentage Vegetation Index; IRECI - Inverted Red-Edge Chlorophyll Index; MCARI - Modified Chlorophyll Absorption in Reflectance Index; MCARI1 - Modified Chlorophyll Absorption in Reflectance Index 1; MCARI2 - Modified Chlorophyll Absorption in Reflectance Index 2; MCARI705 - Modified Chlorophyll Absorption in Reflectance Index (705 and 750 nm); MCARIOSAVI - MCARI/OSAVI Ratio; MCARIOSAVI705 - MCARI/OSAVI Ratio (705 and 750 nm); MGRVI - Modified Green Red Vegetation Index; MNLI - Modified Non-Linear Vegetation Index; MRBVI - Modified Red Blue Vegetation Index; MSAVI - Modified Soil-Adjusted Vegetation Index; MSR - Modified Simple Ratio; MSR705 - Modified Simple Ratio (705 and 750 nm); MTVI1 - Modified Triangular Vegetation Index 1; MTVI2 - Modified Triangular Vegetation Index 2; NDMI - Normalized Difference Moisture Index; NDPI - Normalized Difference Phenology Index; NDREI - Normalized Difference Red Edge Index; NDVI - Normalized Difference Vegetation Index; NDVI705 - Normalized Difference Vegetation Index (705 and 750 nm); NGRDI - Normalized Green Red Difference Index; NLI - Non-Linear Vegetation Index; OCVI - Optimized Chlorophyll Vegetation Index; OSAVI - Optimized Soil-Adjusted Vegetation Index; RDVI - Renormalized Difference Vegetation Index; RENDVI - Red Edge Normalized Difference Vegetation Index; RGBVI - Red Green Blue Vegetation Index; RGRI - Red-Green Ratio Index; RI - Red-Green Ratio Index; RVI - Ratio Vegetation Index; S2REP - Sentinel-2 Red-Edge Position; SARVI - Soil Adjusted and Atmospherically Resistant Vegetation Index; SAVI - Soil-Adjusted Vegetation Index; SI - Shadow Index; SIPI - Structure Insensitive Pigment Index; SR - Simple Ratio; SeLI - Sentinel-2 LAI Green Index; TCARI - Transformed Chlorophyll Absorption in Reflectance Index; TCI - Triangular Chlorophyll Index; TDVI - Transformed Difference Vegetation Index; TGI - Triangular Greenness Index; TRRVI - Transformed Red Range Vegetation Index; TTVI - Transformed Triangular Vegetation Index; TVI - Transformed Vegetation Index; TriVI - Triangular Vegetation Index; VARI - Visible Atmospherically Resistant Index; VI700 - Vegetation Index (700 nm); VIG - Vegetation Index Green; WDRVI - Wide Dynamic Range Vegetation Index; WDVI - Weighted Difference Vegetation Index; mSR705 - Modified Simple Ratio (705 and 445 nm); LSWI - Land Surface Water Index; NDWI - Normalized Difference Water Index; S2WI - Sentinel-2 Water Index; SWM - Sentinel Water Mask; WI1 - Water Index 1; WI2 - Water Index 2; BAIS2 - Burned Area Index for Sentinel 2; NBR - Normalized Burn Ratio;*

ANEXO I

Figura 3.1 – Primeiro modelo selecionado e posteriormente descartado devido a multicolinearidade na forma matemática, sem simplificação.

$$C = \left(\frac{\left(\left(\frac{c_0 \cdot \text{VH\_ASC\_DESC}}{c_1 \cdot \text{OCVI}} - c_2 \cdot \text{bio\_8} \right) \cdot c_3 \cdot \text{bio\_1} - \left(\frac{c_4 \cdot \text{VH\_ASC}}{c_5 \cdot \text{VH\_ASC}} + c_6 \cdot \text{NonTree\_Vegetation} \right) \right)}{c_7 \cdot \text{MRBVI}}{c_8 \cdot \text{MRBVI}} \cdot c_9 + c_{10} \right)$$

$$\begin{aligned} c_0 &= 1.0601 \\ c_1 &= 1.224 \\ c_2 &= -0.25964 \\ c_3 &= 2.0661 \\ c_4 &= 1.0501 \\ c_5 &= 1.0501 \\ c_6 &= 1.841 \\ c_7 &= -0.25132 \\ c_8 &= -0.25132 \\ c_9 &= 0.0032109 \\ c_{10} &= 4.6107 \end{aligned}$$

Fonte: Do autor (2023)

Figura 3.2. - Primeiro modelo selecionado e posteriormente descartado devido a multicolinearidade em forma de árvore de sintaxe gerada no HeuristicLab, sem simplificação.

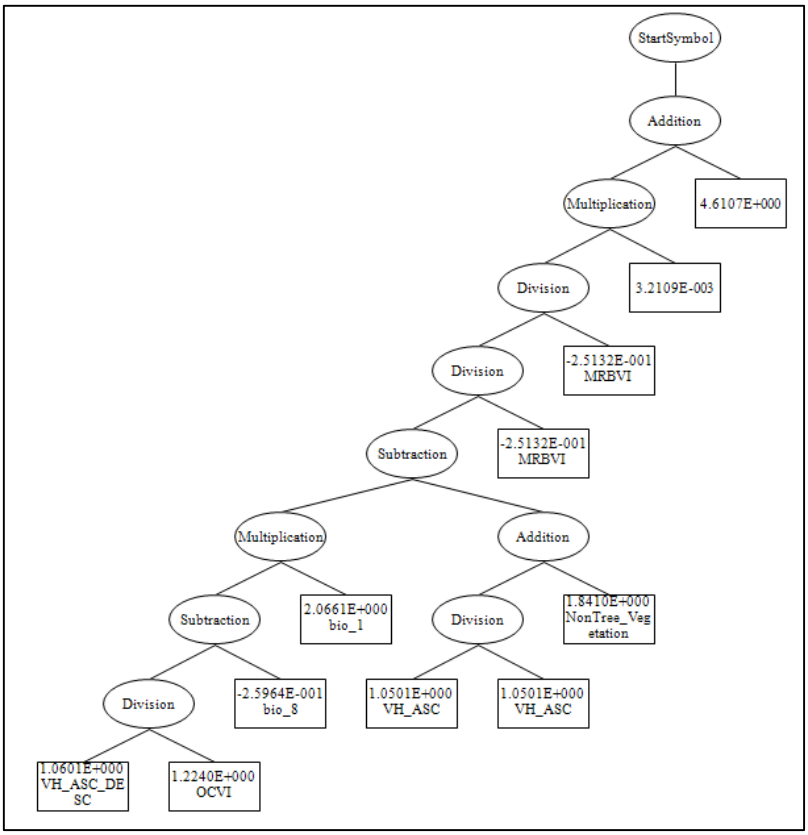


Figura 3.1 – Segundo modelo selecionado na forma matemática, sem simplificação.

$$C = \left(\left(\frac{c_0 \cdot \text{CVI}}{c_1 \cdot \text{VV\_ASC\_DESC}} + \left(c_2 \cdot \text{bio\_11} + \left(\left(\left(\left(\left(c_3 \cdot \text{CIRE} - (c_4 \cdot \text{T3})^2\right)\right)^2 - c_5 \cdot \text{T3}\right) - c_6 \cdot \text{NonTree\_Vegetation}\right) + c_7 \cdot \text{bio\_11}\right)\right)\right) \cdot c_8 + c_9\right)$$

$$\begin{aligned}
c_0 &= 1.845 \\
c_1 &= -0.038307 \\
c_2 &= 2.469 \\
c_3 &= 1.6758 \\
c_4 &= 1.1762 \\
c_5 &= 1.1762 \\
c_6 &= 0.19508 \\
c_7 &= 2.469 \\
c_8 &= 1.6839 \\
c_9 &= -159.75
\end{aligned}$$

Figura 3.2. - Segundo modelo selecionado em forma de árvore de sintaxe gerada no HeuristicLab, sem simplificação.