



EVANDRO NUNES MIRANDA

**MANEJO FLORESTAL ORBITAL: USO DE IMAGENS DE
SATÉLITE E *MACHINE LEARNING* PARA OTIMIZAR E
AUXILIAR A TOMADA DE DECISÃO**

LAVRAS – MG

2024

EVANDRO NUNES MIRANDA

**MANEJO FLORESTAL ORBITAL: USO DE IMAGENS DE SATÉLITE E *MACHINE*
LEARNING PARA OTIMIZAR E AUXILIAR A TOMADA DE DECISÃO**

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

Prof. Dr. Lucas Rezende Gomide
Orientador

Dr. Henrique Ferraço Scolforo
Coorientador

**LAVRAS – MG
2024**

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

Miranda, Evandro Nunes.

Manejo Florestal Orbital: uso de imagens de satélite e machine
learning para otimizar e auxiliar a tomada de decisão / Evandro
Nunes Miranda. - 2024.

107 p.

Orientador(a): Lucas Rezende Gomide.

Coorientador(a): Henrique Ferraço Scolforo.

Tese (doutorado) - Universidade Federal de Lavras, 2024.

Bibliografia.

1. Random Forest. 2. XGboost. 3. Sensoriamento remoto. I.
Gomide, Lucas Rezende. II. Scolforo, Henrique Ferraço. III. Título.

EVANDRO NUNES MIRANDA

**MANEJO FLORESTAL ORBITAL: USO DE IMAGENS DE SATÉLITE E MACHINE
LEARNING PARA OTIMIZAR E AUXILIAR A TOMADA DE DECISÃO**

**ORBITAL FOREST MANAGEMENT: USING SATELLITE IMAGERY AND
MACHINE LEARNING TO OPTIMIZE AND HELP DECISION MAKING**

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

APROVADA em 23 de Agosto de 2024.

Prof. Dr. Lucas Rezende Gomide	UFLA
Prof. Dr. Otávio Camargo Campoe	UFLA
Dr. Kalill José Viana da Páscoa	UFLA
Dr. Henrique Ferraço Scolforo	SUZANO
Dr. Cássio Augusto Ussi Monti	SUZANO

Prof. Dr. Lucas Rezende Gomide

Orientador

Dr. Henrique Ferraço Scolforo

Coorientador

LAVRAS – MG

2024

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por sempre me dar forças e me guiar no caminho correto, colocando pessoas em minha vida que me ajudaram a construir o homem que sou, pessoas que me inspiram, ajudam e desafiam a sempre alcançar o melhor.

Agradeço à Universidade Federal de Lavras (UFLA), por toda a trajetória construída durante minha vida acadêmica, desde a graduação, passando pelo mestrado, até a realização do doutorado. Em especial, agradeço ao programa de Pós-Graduação em Engenharia Florestal.

Meus pais, senhora Anatalete Alves Miranda e senhor Valdete Nunes de Miranda, merecem um agradecimento especial. Eles me educaram e formaram o homem que sou. Sem o amor e suporte deles, nada disso seria possível.

À minha parceira de vida, Keliane, agradeço pelo amor, carinho, suporte e companheirismo. Ela sempre acreditou em meus sonhos e me apoiou nas melhores decisões para que eu pudesse evoluir. Agradeço também aos seus pais, senhora Elma e senhor Pedroso, que me acolheram como filho e sempre me incentivaram.

Agradeço aos meus irmãos, Mácio e Natália, pela fraternidade e carinho.

Ao meu orientador, Prof. Dr. Lucas Rezende Gomide, sou grato pelos diálogos abertos, conselhos de vida, orientações valiosas e incentivo constante para que eu sempre melhorasse.

Ao meu coorientador, Dr. Henrique Ferraço Scolforo, agradeço a oportunidade de crescimento profissional, pelas orientações e amizade ao longo desses anos, além de todo o suporte na coorientação deste trabalho.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro, Código de Financiamento 001, que possibilitou a concessão da bolsa de estudos.

Agradeço a todos os professores que contribuíram para minha formação e me incentivaram a seguir o caminho do conhecimento.

Aos amigos de laboratório do LEMAF, que estiveram ao meu lado ao longo destes anos, e aos amigos da empresa, que me incentivaram e apoiaram durante a construção deste trabalho.

Agradeço à Suzano pela oportunidade profissional e pela possibilidade de conclusão do doutorado.

Por fim, gostaria de expressar minha profunda gratidão a todas as pessoas que me apoiaram de alguma forma ao longo destes anos de doutorado, ou que de alguma forma, fizeram parte da minha vida e contribuíram para a realização deste trabalho.

RESUMO GERAL

O monitoramento florestal com o uso de imagens é uma prática consolidada na indústria florestal, seja por meio de drones ou sensoriamento remoto. No entanto, há uma necessidade contínua de aprimorar tecnologias e metodologias para alcançar resultados mais precisos e eficazes. Imagens de satélite oferecem uma fonte rica de informações, com resolução espacial e temporal que permite o monitoramento constante das florestas, com precisão confiável. Esta tese tem como objetivo utilizar dados de sensoriamento remoto de forma econômica e periódica, integrando essas informações com as condições das florestas. Isso inclui a detecção automatizada e periódica de anomalias florestais, a identificação da competição de mato com florestas de eucalipto em plantios jovens e a previsão do potencial de crescimento e produtividade futura. No primeiro artigo, o objetivo foi construir um modelo autorregressivo que utiliza informações de índice de vegetação, com variáveis cadastrais, como zonas climáticas e material genético, para a detecção de anomalias em plantios florestais. Modelos de *machine learning* (ML), como *random forest* (RF) e *extreme gradient boosting* (XGBoost), foram usados para alcançar esse objetivo, com os melhores resultados obtidos pelo XGBoost. A metodologia proposta mostrou-se altamente aderente e confiável para a detecção de anomalias. No segundo artigo, o objetivo foi detectar a competição de mato com plantios jovens de eucalipto usando imagens de Sentinel-2. Utilizou-se uma metodologia de seleção de variáveis com algoritmo genético e classificação da presença ou ausência de competição com RF, denominado GARF. O método foi eficiente e selecionou um subconjunto ótimo de variáveis que permitem a identificação dessa competição usando imagens de satélite. No terceiro artigo, o foco foi estimar a altura dominante de florestas de eucalipto aos 2 anos de idade, utilizando informações satelitais e climáticas coletadas ao completar um ano de idade dos plantios. Essa estimativa antecipada da qualidade local permite gerar um inventário florestal precoce. Utilizou-se o GARF para seleção de variáveis e regressão, permitindo a inclusão de variáveis climáticas e influências biológicas no modelo. Essas informações serão integradas a um banco de dados para apoiar a tomada de decisões, oferecendo uma base sólida para uma gestão florestal mais eficiente e sustentável. A incorporação de inteligência computacional, como ML, potencializa ainda mais a capacidade de monitoramento constante das florestas e a detecção de anomalias.

Palavras chaves: *Random Forest*; *XGBoost*; Sensoriamento remoto; Detecção de anomalias; Crescimento e produtividade.

GENERAL ABSTRACT

Forest monitoring using images is a well-established practice in the forestry industry, utilizing both drones and remote sensing. However, there is a continuous need to enhance technologies and methodologies to achieve more accurate and effective results. Satellite images provide a rich source of information with spatial and temporal resolution, enabling constant monitoring of forests with reliable precision. This thesis aims to use remote sensing data economically and periodically, integrating this information with forest conditions. This includes the automated and periodic detection of forest anomalies, the identification of weed competition with young eucalyptus plantations, and predicting future growth and productivity potential. In the first article, the objective was to develop an autoregressive model using vegetation index information combined with cadastral variables such as climate zones and genetic material to detect anomalies in forest plantations. Machine learning (ML) models, including random forest (RF) and extreme gradient boosting (XGBoost), were employed to achieve this goal, with XGBoost delivering the best results. The proposed methodology proved to be highly effective and reliable for anomaly detection. The second article focused on detecting weed competition with young eucalyptus plantations using Sentinel-2 images. A methodology for variable selection using a genetic algorithm and classification of the presence or absence of competition with RF, known as GARF, was implemented. This method efficiently selected an optimal subset of variables, enabling the identification of weed competition using satellite images. The third article aimed to estimate the dominant height of eucalyptus forests at two years of age, utilizing satellite and climate information collected when the plantations were one year old. This early estimate of local quality facilitates the generation of a preliminary forest inventory. GARF was employed for variable selection and regression, incorporating climatic variables and biological influences into the model. This information will be integrated into a database to support decision-making, offering a solid basis for more efficient and sustainable forest management. The incorporation of computational intelligence, such as ML, further enhances the ability to constantly monitor forests and detect anomalies.

Keywords: Random Forest; XGBoost; Anomaly detection; Growth and yield.

INDICADORES DE IMPACTO

Este trabalho teve como objetivo buscar soluções sustentáveis, tecnológicas, econômicas e sociais que impactam diretamente a indústria florestal. Diante das mudanças climáticas e das pressões para tornar a produção mais sustentável, ele visa garantir a eficiência das operações, atividades e planejamento da gestão florestal. Intitulado “Manejo Florestal Orbital: uso de imagens de satélite e machine learning para otimizar e auxiliar a tomada de decisão”, o estudo propõe o uso de tecnologias avançadas, como sensoriamento remoto e modelos de machine learning, para modernizar o monitoramento e a gestão florestal. Tais tecnologias permitem o acompanhamento contínuo e abrangente de plantios florestais, com maior precisão, confiabilidade e periodicidade, além de reduzir custos operacionais, especialmente em regiões remotas e de difícil acesso, como Maranhão e Mato Grosso do Sul, áreas caracterizadas por baixa densidade populacional e vastos territórios. No âmbito social, o trabalho contribui significativamente ao democratizar o acesso a informações de alta qualidade sobre os recursos florestais, o que beneficia comunidades locais, governos e empresas ao promover uma gestão territorial mais inclusiva e sustentável. Ele também estimula a geração de empregos qualificados na interface entre tecnologia e gestão florestal, incentivando a capacitação de profissionais e a valorização de talentos locais. Em áreas onde a escassez de mão de obra especializada é um desafio, a adoção de tecnologias automatizadas representa uma solução transformadora que otimiza recursos humanos e financeiros, ao mesmo tempo que amplia as possibilidades de monitoramento em larga escala. Do ponto de vista cultural, o trabalho contribui para a conscientização sobre o papel das florestas na mitigação das mudanças climáticas e na conservação ambiental, reforçando a importância de práticas sustentáveis na preservação de recursos naturais que sustentam populações rurais e indígenas. Ao destacar a relevância das florestas para o equilíbrio ecológico, o projeto também fomenta o engajamento social e educacional em prol de uma cultura de preservação e uso responsável dos recursos naturais. A implementação das tecnologias inovadoras propostas não apenas incentiva a pesquisa e o desenvolvimento na área florestal, mas também promove a transição para uma economia de baixo carbono, alinhando-se aos Objetivos de Desenvolvimento Sustentável (ODS), especialmente os relacionados à indústria, inovação e infraestrutura (ODS 9), ação contra a mudança global do clima (ODS 13) e vida terrestre (ODS 15). Ao integrar dados ambientais, climáticos e biológicos em plataformas digitais, o trabalho oferece suporte técnico robusto para a tomada de decisões estratégicas, contribuindo para a eficiência e sustentabilidade do setor. Por fim, os impactos diretos e indiretos deste trabalho demonstram o potencial transformador da tecnologia em escalas locais e globais, promovendo inovação, inclusão e sustentabilidade para as futuras gerações.

IMPACT INDICATORS

This study aimed to find sustainable, technological, economic and social solutions that directly impact the forestry industry. In the face of climate change and pressures to make production more sustainable, it aims to ensure the efficiency of forest management operations, activities and planning. Entitled “Orbital Forest Management: use of satellite images and machine learning to optimize and assist decision-making”, the study proposes the use of advanced technologies, such as remote sensing and machine learning models, to modernize forest monitoring and management. Such technologies allow continuous and comprehensive monitoring of forest plantations, with greater precision, reliability and frequency, in addition to reducing operating costs, especially in remote and difficult-to-access regions, such as Maranhão and Mato Grosso do Sul, areas characterized by low population density and vast territories. In the social sphere, the study contributes significantly by democratizing access to high-quality information on forest resources, which benefits local communities, governments and companies by promoting more inclusive and sustainable territorial management. It also stimulates the generation of skilled jobs at the interface between technology and forest management, encouraging the training of professionals and the appreciation of local talent. In areas where the shortage of skilled labor is a challenge, the adoption of automated technologies represents a transformative solution that optimizes human and financial resources, while expanding the possibilities for large-scale monitoring. From a cultural perspective, the work contributes to raising awareness about the role of forests in mitigating climate change and environmental conservation, reinforcing the importance of sustainable practices in preserving natural resources that sustain rural and indigenous populations. By highlighting the importance of forests for ecological balance, the project also fosters social and educational engagement in favor of a culture of preservation and responsible use of natural resources. The implementation of the proposed innovative technologies not only encourages research and development in the forestry sector, but also promotes the transition to a low-carbon economy, aligning with the Sustainable Development Goals (SDGs), especially those related to industry, innovation and infrastructure (SDG 9), climate action (SDG 13) and life on land (SDG 15). By integrating environmental, climate and biological data into digital platforms, the work offers robust technical support for strategic decision-making, contributing to the efficiency and sustainability of the sector. Finally, the direct and indirect impacts of this work demonstrate the transformative potential of technology at local and global scales, promoting innovation, inclusion and sustainability for future generations.

LISTA DE FIGURAS

PRIMEIRA PARTE

Figura 1 - Fluxograma do funcionamento do Random Forest para problemas de regressão e classificação.....	20
Figura 2 - Fluxograma de regressão do algoritmo de machine learning Extreme Gradient Boosting (XGBoost).....	21

SEGUNDA PARTE

ARTIGO 1

Figura 1 - Mapa de localização da área de estudo.....	32
Figura 2 - Fluxograma ilustrativo das etapas para o pré-processamento das imagens Sentinel 2 para compor o banco de dados de treinamento.....	35
Figura 3 - Fluxograma da metodologia para detecção de anomalia.....	36
Figura 4 - Valores médios de importância das variáveis explicativas dos modelos RF e XGBoost para prever o valor de NDVI.....	41
Figura 5 - Dinâmica das mudanças e detecção de anomalias para um talhão com RF e XGBoost.....	42
Figura 6- Mapa com mudanças e detecção de anomalias para um talhão com XGBoost e XGBoost com Quantiles.....	43
Figura 7 - Matriz de confusão com os dados de validação para XGBoost e XGBoost com quantiles.....	44

ARTIGO 2

Figura 1 - (a) Exemplo de caminhamento dos transectos de campo em um povoamento, e avaliação da matocompetição. (b) Localização dos quatro locais de estudo.....	58
Figura 2 - Fluxograma da metodologia aplicada para a predição da ocorrência espacial de matocompetição em plantios comerciais.....	62
Figura 3 -Gráfico de violino para análise exploratória dos dados.....	64
Figura 4 - Panorama do comportamento do grau de matocompetição que ocorre nos plantios florestais de eucalipto.....	65
Figura 5 - Valores médios de importância das variáveis explicativas do modelo para classificar a ocorrência de matocompetição nas florestas jovens de eucalipto (A). Gráficos de redução do número de variáveis e valor da função de aptidão do GARF em função do número de gerações.....	66
Figura 6 – a) Matriz de confusão com os dados de validação; b) Exemplo de saída da classificação do mato, comparado com a imagem de satélite e a foto de campo; c) Mapas com a classificação de MC comparadas com imagens em cores verdadeiras, separando pelo nível de intensidade.....	68

ARTIGO 3

Figura 1 - Mapa de localização, clima e de solo da área de estudo.....	85
Figura 2 - Fluxograma da metodologia proposta para estimar qualidade local.....	89

Figura 3 - Gráficos de redução do número de variáveis e melhoria da função fitness do modelo GARF em função do número de gerações.....	92
Figura 4 - Valores médio de importância das variáveis explicativas dos modelos para estimar a altura dominante aos 2 anos com variáveis aos 12 meses.	93
Figura 5 - Gráfico de correlação da altura dominante observada x estimada, resíduos e distribuição dos resíduos.	94
Figura 6 - Gráfico de correlação da altura dominante aos 7 anos observada x estimada, resíduos e distribuição dos resíduos.	95
Figura 7 - Histograma de frequência dos cenários testados para a análise de sensibilidade....	96

LISTA DE TABELAS

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

Tabela 1 - Tabela descritiva das variáveis que compõem o banco de dados florestal.	33
---	----

ARTIGO 2

Tabela 1 - Resultados de predição de duas classes utilizadas para a avaliação de detecção de ervas daninhas em plantios florestais juvenis.....	67
Tabela 2 - Descrição dos índices orbitais de Sentinel-2.....	77

ARTIGO 3

Tabela 1 - Tabela descritiva das variáveis descritivas para plantios aos dois anos.....	86
Tabela 2 -Tabela descritiva das variáveis climáticas.....	87
Tabela 3 - Métricas de avaliação dos modelos de altura dominante.	93
Tabela 4 - Descrição dos índices orbitais de Sentinel-2.....	105

SUMÁRIO

PRIMEIRA PARTE	14
1 INTRODUÇÃO	14
2 REVISÃO DE LITERATURA.....	17
2.1 Inventário Florestal	17
2.2 Séries Temporais.....	18
2.3 <i>Random Forest</i>	19
2.4 <i>Extreme Gradient Boosting Tree</i>	20
3. CONSIDERAÇÕES FINAIS.....	22
REFERÊNCIAS	23
SEGUNDA PARTE - ARTIGOS	27
ARTIGO 1 - PROJEÇÃO DO COMPORTAMENTO ESPECTRAL PARA A DETECÇÃO PRECOCE DE ANOMALIAS EM FLORESTAS CLONAIAS DE EUCALIPTO	27
1 INTRODUÇÃO	30
2 MATERIAL E MÉTODOS	31
2.1 Caracterização da área de estudo	31
2.2 Amostragem e caracterização das anomalias florestais	32
2.3 Estruturação dos dados espectrais	33
2.4 Sistema de detecção de anomalias florestais	35
2.5 Modelos Autoregressivo aplicados	38
2.5.1 <i>Random Forest (RF)</i>	38
2.5.2 <i>Extreme Gradient Boosting Tree – XGBoost</i>	39
2.5.3 Análise da performance dos modelos.....	39
3 RESULTADOS	40
4 DISCUSSÃO	44
5 CONCLUSÃO.....	46
REFERÊNCIAS	47
ARTIGO 2 – UMA ABORDAGEM BASEADA EM MACHINE LEARNING E DADOS ORBITAIS PARA MAPEAMENTO DE MATOCOMPETIÇÃO EM PLANTIOS CLONAIAS JOVENS DE EUCALIPTO	53
1 INTRODUÇÃO	56
2 MATERIAL E MÉTODOS	57
2.1 Caracterização da área de estudo	57
2.2 Caracterização da base de dados das plantações de <i>Eucalyptus</i>	58
2.3 Caracterização dos dados orbitais	59
2.4 Modelagem para detecção da ocorrência de matocompetição	60
2.5 Medidas de Desempenho	63
3 RESULTADOS	63

3.1	Comportamento dos dados	63
3.2	Avaliação do modelo GARF e de precisão do modelo desenvolvido.....	65
4	DISCUSSÃO	69
5	CONCLUSÃO.....	70
	REFERÊNCIAS	72
	ANEXO A	77
ARTIGO 3 – UMA ABORDAGEM PARA ESTIMATIVA PRECOCE DO POTENCIAL PRODUTIVO DE FLORESTAS DE EUCALIPTO EM FUNÇÃO DA QUALIDADE DE PLANTIO, EDAFOCLIMÁTICO E SENSORIAMENTO REMOTO		
	80	
1	INTRODUÇÃO	83
2	MATERIAL E MÉTODOS	85
2.1	Caracterização da Área de Estudo.....	85
2.2	Caracterização da base de dados de inventário florestal contínuo e cadastrais 86	
2.3	Caracterização dos dados climáticos	87
2.4	Pré-processamento de imagens Sentinel-2 e cálculo de variáveis	87
2.5	Estimativa do Sítio aos 12 meses	88
2.5.1	Modelo RF baseado em GA para estimar Hdom aos 12 meses.	89
2.5.2	Estimativa do Índice de Sítio aos 7 anos com base no Hdom aos 12 meses....	90
2.5.3	Avaliação e validação das abordagens de modelagem	91
3	RESULTADOS	92
3.1	Seleção de variáveis	92
3.2	Estimativa de Hdom	93
3.3	Estimativa do índice de sítio	94
3.4	Simulação do IS em relação ao clima e à qualidade do plantio.....	95
4	DISCUSSÃO	97
5	CONCLUSÃO.....	99
	REFERÊNCIAS	101
	ANEXO A	105

PRIMEIRA PARTE

1 INTRODUÇÃO

A produção de madeira de florestas plantadas no Brasil é voltada principalmente para a produção de celulose e carvão vegetal. A cadeia produtiva florestal contribui significativamente para o PIB per capita do país, com 6,3% (IBÁ, 2024), desempenhando um papel crucial no agronegócio e na economia brasileira. Atualmente, o Brasil possui 9,94 milhões de hectares de florestas plantadas (IBÁ, 2024). Para manter o abastecimento de madeira às indústrias de celulose e carvão, é necessário realizar monitoramentos constantes dessas florestas. Estimar a produtividade, detectar anomalias e outras perturbações é fundamental para fornecer informações essenciais ao planejamento florestal. Essas informações são vitais para que os gestores florestais possam tomar as melhores decisões.

Cada vez mais, há uma necessidade crescente de informações geradas das florestas para apoiar políticas econômicas, ambientais e sociais. As tecnologias de sensoriamento remoto oferecem uma ampla gama de dados em um curto período, com cobertura espacial que abrange toda a área dos plantios. Isso permite a aquisição de inúmeros produtos que servem de suporte para o monitoramento orbital dessas florestas, além de possibilitar o desenvolvimento e a parametrização subsequente de modelos para atender a uma variedade ainda maior de necessidades de informação (WHITE et al., 2016). Dados de sensoriamento remoto mais abundantes cada vez são necessários para descrever informações florestais detalhadas (ZHU et al., 2020).

A aquisição de imagens de satélite possibilita uma série de atividades, como a detecção e identificação de anomalias nos plantios, análise de déficit hídrico, ocorrência de incêndios, entre outras (CHUVIECO et al., 2010; DINDAROGLU et al., 2021; MCCANN et al., 2017; MIAO et al., 2024; PENOT; MERLIN, 2023; RACETIN; KRTALIĆ, 2021; REMBOLD et al., 2013; SHUAI; BASSO, 2022). Além de seu uso no monitoramento, as imagens de satélite também permitem estimar atributos florestais, abordagens que tradicionalmente são realizadas pelo inventário florestal (LI et al., 2020; LIU et al., 2019; MIRANDA et al., 2022; NANDY; SRINET; PADALIA, 2021; SANTOS et al., 2023; ZHOU; FENG, 2023).

Embora os inventários florestais tradicionais busquem precisão e consistência nas interpretações e medições subjetivas, eles são caros para serem realizados, a inclusão de tecnologias de sensoriamento remoto no inventário florestal oferece várias vantagens, como a possibilidade de uma área maior de amostragem, aquisição de informações a um custo reduzido,

obtenção periódica e contínua desses dados, e monitoramento constante da floresta (THOMPSON et al., 2007; WHITE et al., 2016). Algumas informações que não podem ser obtidas diretamente através de imagens de satélite podem ser derivadas por meio de modelagem (BROSOFKSKE et al., 2014; WHITE et al., 2016). Diante disso, o objetivo desta tese foi utilizar imagens de satélite em diferentes abordagens e etapas do ciclo de vida das florestas de eucalipto para auxiliar no monitoramento, extração de informações e facilitar a tomada de decisão dos gestores florestais, de forma embasada e assertiva.

A tese está organizada em quatro partes principais. A primeira parte apresenta uma breve revisão bibliográfica de alguns tópicos que são abordados ao longo do trabalho. O segundo capítulo detalha a detecção de anomalias em plantios florestais de eucalipto, utilizando os algoritmos de *machine learning* *Random Forest* (RF) e *eXtreme Gradient Boosting* (XGBoost) em séries temporais de imagens Sentinel-2. A metodologia emprega uma abordagem autorregressiva para gerar uma expectativa do valor do pixel em momentos futuros, comparando-o com o valor observado. Se o valor observado estiver abaixo do percentil 10% dos modelos ou abaixo de 0.15 no índice de vegetação utilizado (NDVI), o pixel é identificado como uma provável anomalia. Para potencializar a capacidade de predição dos modelos, é necessária a confirmação da anomalia em três imagens consecutivas. Devido à alta frequência de captura das imagens do Sentinel-2, o retorno de uma anomalia confirmada pelo algoritmo ocorre em até 15 dias. Variáveis de cadastro dos plantios, como zonas climáticas, material genético, entre outras, são incorporadas aos modelos autorregressivos, permitindo a criação de um sistema automatizado de identificação de anomalias em plantios florestais. Esta abordagem possibilita o monitoramento constante da floresta e a identificação de anomalias em tempo real, facilitando a tomada de decisões pelos gestores florestais e mitigando possíveis aumentos nas perturbações.

No capítulo 3, é apresentada uma abordagem para detecção automatizada de matocompetição em florestas jovens com menos de 9 meses de idade. Utilizando informações de campo sobre áreas com e sem ocorrência de matocompetição, foram extraídos índices de vegetação, água e terreno a partir de imagens de satélite que pudessem explicar a ocorrência da matocompetição. Devido à grande disponibilidade de variáveis de entrada, foi utilizado um método de seleção de variáveis com o algoritmo genético e o *Random Forest* para classificação da matocompetição. Esta abordagem otimiza o conjunto de variáveis que explicam a competição entre mato e os plantios florestais de eucalipto.

O último capítulo, capítulo 4, busca estimar a altura dominante de florestas de eucalipto aos 2 anos de idade, utilizando informações satelitais e climáticas coletadas ao completar um

ano de idade dos plantios. Esta estimativa antecipada da qualidade local permite gerar um inventário florestal precoce e obter informações que seriam coletadas apenas aos 2 anos. Esta abordagem gera uma série de percepções e análises de sensibilidade sobre o comportamento da floresta em cenários extremos.

Todos os capítulos apresentados permitem gerar uma série de percepções sobre as florestas e trazem uma nova visão sobre o manejo florestal de eucaliptos. As ferramentas desenvolvidas e apresentadas nesta tese servirão como apoio à tomada de decisões pelos gestores, tornando essas decisões cada vez mais assertivas e antecipando possíveis problemas que poderiam ocorrer nas florestas de eucalipto.

2 REVISÃO DE LITERATURA

2.1 Inventário Florestal

Para aproveitar eficientemente os recursos florestais e maximizar seu potencial, é essencial dispor de informações abrangentes e confiáveis, obtidas através de inventários florestais (SOUSA et al., 2023). A realização de inventários florestais, na contextualização do Brasil, abrange uma complexa e abrangente aquisição de informações para a gestão florestal. Esses inventários visam caracterizar populações florestais em termos qualitativos, quantitativos e dinâmicos, utilizando técnicas de mapeamento, medição e amostragem florestal para garantir dados precisos e de custo compatível (SCOLFORO; MELLO, 2006). Deste modo, essas informações devem-se precisas, e espacialmente detalhada. De forma abrangente, a maioria dos inventários realizados no Brasil, são realizados com equipes em campo, com lançamento de parcelas, e mensuração das circunferências das árvores e medição da altura delas (OLIVEIRA et al., 2021). E essas informações exigem informações anuais dessas mensuradas, o que eleva consideravelmente os custos.

O sucesso dos inventários florestais depende do equilíbrio entre o nível de precisão desejado, os recursos financeiros disponíveis e o tempo necessário para a coleta de dados. Esse equilíbrio é alcançado por meio de métodos e processos de amostragem, onde uma parte representativa da população é medida para gerar informações confiáveis sobre os recursos florestais disponíveis (PELLICO NETTO; BRENA, 1997).

As empresas florestais, continuamente procuram alternativas que consigam confiabilidade de aquisição dessas informações, com custo menor, e com o mesmo grau de acurácia. Assim, iniciativas como LiDAR (do inglês *Light Detection And Ranging*) são realizadas. Embora a tecnologia LiDAR seja uma alternativa eficiente e amplamente aplicada em inventários florestais em países como Estados Unidos, Finlândia e Suécia, ainda existem obstáculos para sua utilização em outros países como o Brasil (OLIVEIRA et al., 2021). Apesar da excelente confiabilidade e precisão, é extremamente caro. Deste modo, outras iniciativas, como o uso de imagens de satélites, uso de drones são aplicadas na área.

A aquisição de informações de dados de imagens de satélites permite de maneira econômica e precisa de estimar parâmetros da floresta. Vários estudos têm demonstrado que informações espectrais, como as dos satélites Landsat e Sentinel-2, podem ser utilizadas para estimar diversas variáveis florestais (DOS REIS et al., 2018; MIRANDA et al., 2022; REIS et al., 2019; SANTOS et al., 2023; WU et al., 2016).

2.2 Séries Temporais

As séries temporais são um conjunto de dados de uma determinada variável, descrita e ordenada ao longo de um período de tempo. Diferente de outras análises estatísticas, as informações que compõem os bancos de dados de séries temporais são organizadas em períodos distintos e sucessivos, sendo dependentes entre si. Esse caráter dependente exige técnicas e conceitos estatísticos especiais para seu estudo. Os problemas de séries temporais da vida real são extremamente complexos para serem modeladas usando abordagens simples, especialmente quando alta precisão é necessária (ÖMER FARUK, 2010). A previsão de séries temporais, portanto, é um método essencial para prever tendências futuras com base em dados históricos que possuem características temporais (LUO et al., 2021a). Esses problemas são relevantes em diversas áreas, como previsão de demanda, estimativa de preços futuros de matérias-primas, e cobertura de custos de fabricação, entre outras. Cada uma dessas aplicações requer uma análise cuidadosa e uma modelagem precisa para garantir resultados confiáveis e úteis.

Existem várias abordagens para a previsão de séries temporais, cada qual com suas particularidades. Métodos estatísticos tradicionais são comumente utilizados, como os métodos de análise temporal Box-Jenkins (BOX et al., 2015), que incluem modelos auto-regressivos (AR), média móvel (MA), média móvel autoregressiva (ARMA), média móvel auto-regressiva integrada (ARIMA), entre outros (BOX et al., 2015; KHASHEI; BIJARI, 2011; MURAT et al., 2018; ÖMER FARUK, 2010). Esses métodos são lineares, pois as restrições dos valores futuros são restritas a funções lineares de observações anteriores (KHASHEI; BIJARI, 2011). Para superar as limitações dos modelos lineares, inúmeras abordagens não lineares foram propostas. De acordo com Khashei e Bijari (2011), inúmeros métodos foram propostos para superar a limitação linear dos modelos de séries temporais. Os quais podem ser inclusos o modelo bilinear, o modelo *threshold autoregressive* (TAR), o modelo heterocedástico condicional autorregressivo (ARCH), dinâmica caótica e redes neurais artificiais, como perceptrons multicamadas (MLPs), redes de função de base radial (RBF) e *Long Short-Term Memory Networks* (LSTM) (ALI, Muhammad et al., 2021; AZAD et al., 2022; REUSS et al., 2021; WANG, Zhen et al., 2020; ZHANG, He; SRINIVASAN; YANG, 2021). Outros algoritmos de *machine learning* comumente utilizados que podem ser citados são máquinas de vetores de suporte (SVMs), sistema de inferência *fuzzy* adaptável (ANFIS), *random forest* (RF), *eXtreme Gradient Boosting* (XGBoost), entre outros (HAZNEDAR; KILINC, 2022; REUSS et al., 2021; TIWARI; KUMARASWAMIDHAS; GARG, 2022; XIAN; CHEN; CHENG, 2022a).

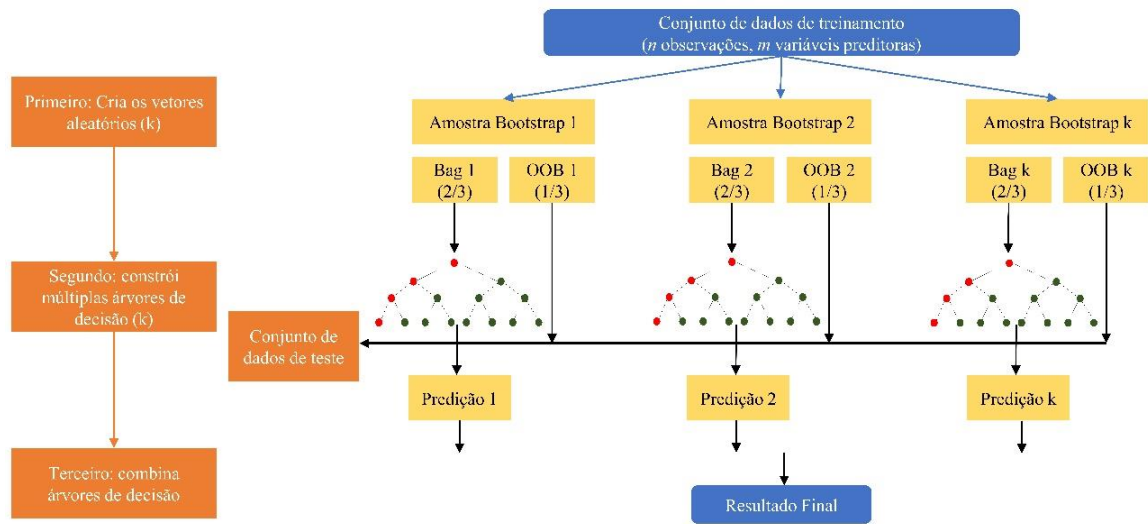
Modelos não lineares baseados em machine learning, como RF e XGBoost para séries temporais tem produzido resultados convincentes nos últimos anos (LUO et al., 2021b; LV et al., 2021; REGIER et al., 2023; SUN, Mengmeng et al., 2023). O principal desafio destes modelos tem sido ajustar a sua parametrização e algoritmos de aprendizagem.

2.3 *Random Forest*

O *Random Forest* (RF) é um algoritmo de aprendizado de máquina, amplamente utilizado em problemas de regressão e classificação usando múltiplas árvores de decisão e agregação *bootstrap* (*bagging*) (BURKART; ROBERT; HUBER, 2021). foi introduzido por Breiman (2001). *Bagging* envolve treinar cada árvore de decisão em diferentes amostras de dados. A ideia é não depender de árvores de decisão individuais, mas combinar múltiplas árvores de decisão para determinar a saída final com maior precisão (BURKART; ROBERT; HUBER, 2021).

O RF serve como um exemplo clássico de aplicação do *bagging*. Ele constrói uma "floresta" de árvores de decisão e treina cada uma delas com uma amostra aleatória dos dados originais. Para reduzir o risco de *overfitting* e aumentar a estabilidade do modelo, o RF produz uma previsão final precisa e robusta ao combinar as previsões de todas as árvores. Como o RF é uma combinação de árvores de decisão, ele possui vários hiperparâmetros que controlam a construção de cada árvore de decisão, a configuração e o tamanho da floresta, além da sua aleatoriedade (PROBST; WRIGHT; BOULESTEIX, 2019). Dentre os principais, os mais citados são o número de árvores a serem construídas (*ntree*), o número de recursos selecionados aleatoriamente (*mtry*) e o profundidade de cada árvore. (*nodesize*). Cada um desses hiperparâmetros apresenta sua própria importância e influência na previsão de saída. A precisão do modelo é diretamente impactada por esses hiperparâmetros. A precisão tende a melhorar gradualmente com o aumento do número de árvores, mas nem sempre há um ganho substancial. Embora a literatura tenha sugerido alguns valores, a escolha dos recursos é fundamental para a classificação. No entanto, vale a pena experimentar diferentes combinações. A profundidade da árvore também é importante porque uma baixa pode subajuster o modelo (PARMAR; BHANDERI; SHAH, 2014). O funcionamento do RF de forma ilustrativa é esquematizado na figura 1.

Figura 1 - Fluxograma do funcionamento do *Random Forest* para problemas de regressão e classificação.



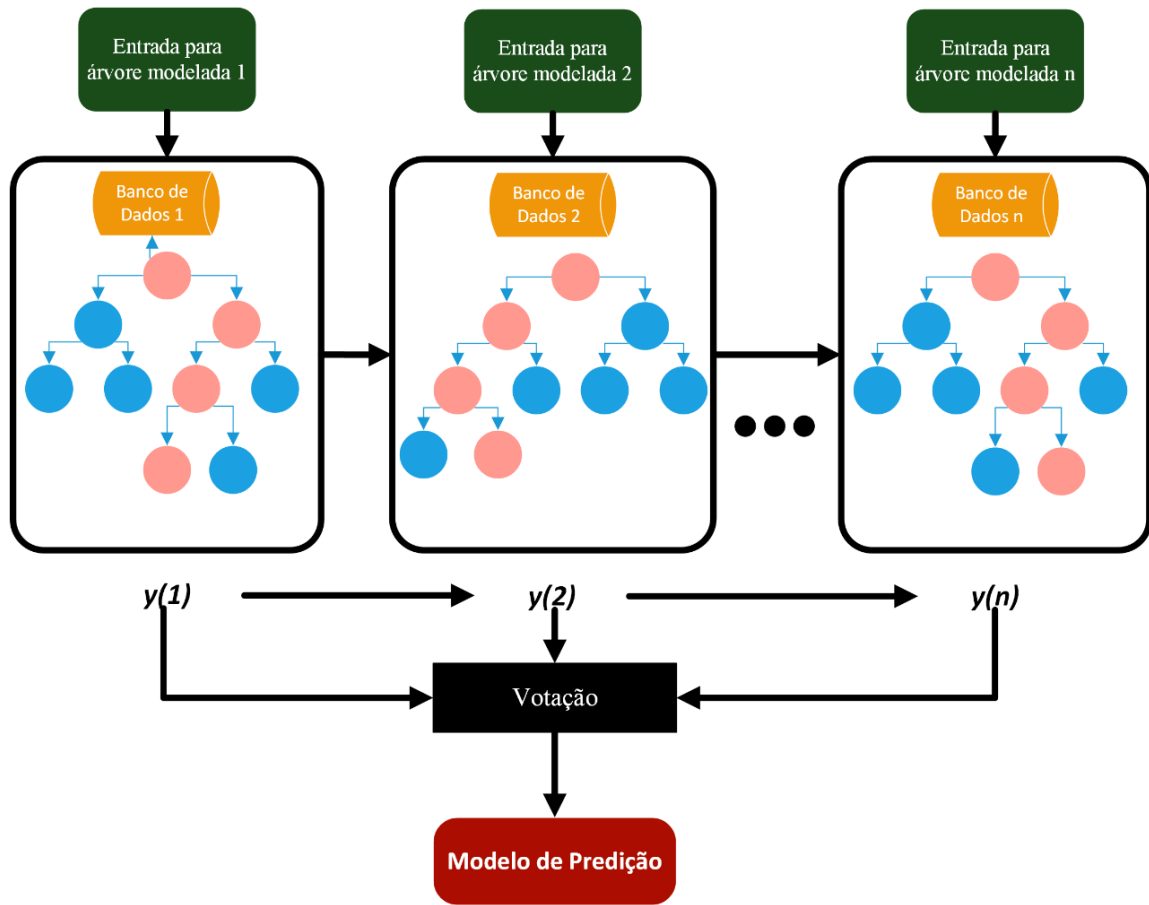
Fonte: Adaptado (CHENG et al., 2019; RODRIGUEZ-GALIANO et al., 2016).

2.4 Extreme Gradient Boosting Tree

O *eXtreme Gradient Boosting* (XGBoost) é um algoritmo de *machine learning* baseado em uma árvore de gradiente em *boosting* (CHEN, Tianqi et al., 2015). O XGBoost é uma implementação eficiente e escalonável da estrutura de aumento de gradiente de Friedman (FRIEDMAN, 2001; FRIEDMAN; HASTIE; TIBSHIRANI, 2000). O XGBoost foi projetado principalmente para velocidade e desempenho usando árvores de decisão com gradiente de reforço (DHALIWAL; NAHID; ABBAS, 2018). Para um desempenho eficiente, o XGBoost implementa a arquitetura de árvore de decisão com reforço de gradiente que pode gerar alta precisão em tarefas de classificação e regressão (ZHU et al., 2021).

O XGBoost é uma versão avançada do modelo *Gradient Boosting Regression Trees* (GBRT). O GBRT é um modelo de aprendizado de máquina que utiliza uma série de árvores de regressão básicas, combinadas de maneira sequencial para melhorar a precisão do modelo. Na ilustração da figura 2, é apresentado o funcionamento para a regressão do modelo de XGBoost.

Figura 2 - Fluxograma de regressão do algoritmo de *machine learning* *Extreme Gradient Boosting* (XGBoost).



Fonte: Do Autor (2024).

O XGBoost aprimora essa técnica, permitindo a adição adaptativa de mais árvores para aumentar a capacidade do modelo. Esta abordagem não apenas melhora a performance, mas também incorpora mecanismos de regularização que ajudam a prevenir o *overfitting*, tornando o XGBoost uma escolha poderosa e eficiente para tarefas de regressão e classificação. A função de minimização do XGBoost é a seguinte:

$$Obj(\theta) = \sum_{i=1}^n L\left(Y_i, \hat{Y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

Em que: θ : parâmetros do modelo. L : função de perda. $\hat{Y}$: valor estimado para a iteração $t - 1$. Y : valor observado. $f_t(x_i)$: valor estimado da árvore na iteração t . $\Omega(f_t)$: termo de regularização que penaliza modelos mais complexos.

3. CONSIDERAÇÕES FINAIS

Considerando o cenário atual, a vasta quantidade de informações disponíveis permite gerar uma série de produtos que podem auxiliar na gestão de florestas plantadas. No entanto, para alcançar os resultados esperados, é necessário utilizar diferentes tecnologias e metodologias. Esse contexto exige esforços crescentes na implementação de soluções que permitam uma gestão eficiente da floresta, desde o monitoramento automatizado e inteligente para a detecção de anomalias até a obtenção da produtividade florestal com o uso de imagens de satélite. A utilização de métodos de inteligência computacional, como modelos de *machine learning*, é fundamental, pois esses métodos demonstram excelente desempenho na modelagem e classificação de imagens. Eles oferecem a capacidade de monitoramento constante da floresta, permitindo a identificação precoce de problemas, a estimativa precisa da produtividade e a análise detalhada de fatores como déficit hídrico e ocorrência de incêndios. Essas tecnologias não apenas facilitam a tomada de decisões embasadas e assertivas, mas também contribuem para a sustentabilidade e eficiência da gestão florestal.

REFERÊNCIAS

- ALI, M. et al. Predicting the Direction Movement of Financial Time Series Using Artificial Neural Network and Support Vector Machine. **Complexity**, v. 2021, p. 1–13, 2 dez. 2021.
- AZAD, A. S. et al. Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study. **Sustainability**, v. 14, n. 3, p. 1843, 5 fev. 2022.
- BOX, G. E. P. et al. **TIME SERIES ANALYSIS: Forecasting and Control**. Hoboken: John Wiley & Sons, 2015.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- BROSOFKSKE, K. D. et al. A Review of Methods for Mapping and Prediction of Inventory Attributes for Operational Forest Management. **Forest Science**, v. 60, n. 4, p. 733–756, 7 ago. 2014.
- BURKART, N.; ROBERT, S.; HUBER, M. F. Are you sure? Prediction revision in automated <scp>decision-making</scp>. **Expert Systems**, v. 38, n. 1, 12 jan. 2021.
- CHEN, T. et al. Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.
- CHENG, L. et al. Applying a random forest method approach to model travel mode choice behavior. **Travel Behaviour and Society**, v. 14, p. 1–10, jan. 2019.
- CHUVIECO, E. et al. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. **Ecological Modelling**, v. 221, n. 1, p. 46–58, jan. 2010.
- DHALIWAL, S. S.; NAHID, A.-A.; ABBAS, R. Effective Intrusion Detection System Using XGBoost. **Information**, v. 9, n. 7, p. 149, 21 jun. 2018.
- DINDAROGLU, T. et al. Evaluation of geomorphometric characteristics and soil properties after a wildfire using Sentinel-2 MSI imagery for future fire-safe forest. **Fire Safety Journal**, v. 122, p. 103318, jun. 2021.
- DOS REIS, A. A. et al. Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods. **New Zealand Journal of Forestry Science**, v. 48, n. 1, p. 1, 16 dez. 2018.
- FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, 1 out. 2001.
- FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). **The Annals of Statistics**, v. 28, n. 2, 1 abr. 2000.

HAZNEDAR, B.; KILINC, H. C. A Hybrid ANFIS-GA Approach for Estimation of Hydrological Time Series. **Water Resources Management**, v. 36, n. 12, p. 4819–4842, 6 set. 2022.

IBÁ. **Anuário Estatístico da Indústria Brasileira de Árvores**. Brasília: [s.n.].

KHASHEI, M.; BIJARI, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. **Applied Soft Computing**, v. 11, n. 2, p. 2664–2675, mar. 2011.

LI, Y. et al. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. **Scientific Reports**, v. 10, n. 1, p. 9952, 19 jun. 2020.

LIU, Y. et al. Estimation of the forest stand mean height and aboveground biomass in Northeast China using SAR Sentinel-1B, multispectral Sentinel-2A, and DEM imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 151, n. November 2018, p. 277–289, 2019.

LUO, J. et al. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. **Results in Physics**, v. 27, p. 104462, ago. 2021a.

LUO, J. et al. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. **Results in Physics**, v. 27, p. 104462, ago. 2021b.

LV, C.-X. et al. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. **BMC Infectious Diseases**, v. 21, n. 1, p. 839, 19 dez. 2021.

MCCANN, C. et al. Multi-temporal mesoscale hyperspectral data of mixed agricultural and grassland regions for anomaly detection. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 131, p. 121–133, set. 2017.

MIAO, J. et al. Modeling strategies and influencing factors in retrieving canopy equivalent water thickness of mangrove forest with Sentinel-2 image. **Ecological Indicators**, v. 158, p. 111497, jan. 2024.

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, jan. 2022.

MURAT, M. et al. Forecasting daily meteorological time series using ARIMA and regression models. **International Agrophysics**, v. 32, n. 2, p. 253–264, 1 abr. 2018.

NANDY, S.; SRINET, R.; PADALIA, H. Mapping Forest Height and Aboveground Biomass by Integrating ICESat-2, Sentinel-1 and Sentinel-2 Data Using Random Forest Algorithm in Northwest Himalayan Foothills of India. **Geophysical Research Letters**, v. 48, n. 14, 28 jul. 2021.

OLIVEIRA, C. P. DE et al. Modeling and Spatialization of Biomass and Carbon Stock Using LiDAR Metrics in Tropical Dry Forest, Brazil. **Forests**, v. 12, n. 4, p. 473, 13 abr. 2021.

ÖMER FARUK, D. A hybrid neural network and ARIMA model for water quality time series prediction. **Engineering Applications of Artificial Intelligence**, v. 23, n. 4, p. 586–594, jun. 2010.

PARMAR, H.; BHANDERI, S.; SHAH, G. **Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters**. International Conference on Information Science. **Anais...Kerala**: 2014.

PELLICO NETTO, S.; BRENA, D. A. **Inventário Florestal**. Curitiba: Editorado pelos autores, 2007.

PENOT, V.; MERLIN, O. Estimating the Water Deficit Index of a Mediterranean Holm Oak Forest From Landsat Optical/Thermal Data: A Phenomenological Correction for Trees Casting Shadow Effects. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 16, p. 5867–5880, 2023.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, 28 maio 2019.

RACETIN, I.; KRTALIĆ, A. Systematic Review of Anomaly Detection in Hyperspectral Remote Sensing Applications. **Applied Sciences**, v. 11, n. 11, p. 4878, 26 maio 2021.

REGIER, P. et al. Effects of random forest modeling decisions on biogeochemical time series predictions. **Limnology and Oceanography: Methods**, v. 21, n. 1, p. 40–52, 11 jan. 2023.

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

REMBOLD, F. et al. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. **Remote Sensing**, v. 5, n. 4, p. 1704–1733, 8 abr. 2013.

REUSS, F. et al. Comparison of Long Short-Term Memory Networks and Random Forest for Sentinel-1 Time Series Based Large Scale Crop Classification. **Remote Sensing**, v. 13, n. 24, p. 5000, 9 dez. 2021.

RODRIGUEZ-GALIANO, V. F. et al. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. **Biogeosciences**, v. 13, n. 11, p. 3305–3317, 6 jun. 2016.

SANTOS, J. S. et al. Predicting eucalyptus plantation growth and yield using Landsat imagery in Minas Gerais, Brazil. **Ecological Informatics**, v. 75, p. 102120, jul. 2023.

SCOLFORO, J. R. S.; MELLO, J. M. DE. **Inventário Florestal**. 1. ed. Lavras: UFLA, 2006.

SHUAI, G.; BASSO, B. Subfield maize yield prediction improves when in-season crop water deficit is included in remote sensing imagery-based models. **Remote Sensing of Environment**, v. 272, p. 112938, abr. 2022.

SOUSA, C. S. C. DE et al. Size of sampling units appropriate for management level forest inventory in Eastern Amazon, Brazil. **Revista em Agronegócio e Meio Ambiente**, v. 16, n. 2, p. 1–13, 30 jun. 2023.

SUN, M. et al. Reconstruction of a Monthly 1 km NDVI Time Series Product in China Using Random Forest Methodology. **Remote Sensing**, v. 15, n. 13, p. 3353, 30 jun. 2023.

THOMPSON, I. D. et al. Accuracy of forest inventory mapping: Some implications for boreal forest management. **Forest Ecology and Management**, v. 252, n. 1–3, p. 208–221, nov. 2007.

TIWARI, S. K.; KUMARASWAMIDHAS, L. A.; GARG, N. Comparison of SVM and ARIMA Model in Time-Series Forecasting of Ambient Noise Levels. Em: [s.l: s.n.]. p. 777–786.

WANG, Z. et al. Monthly streamflow prediction using a hybrid stochastic-deterministic approach for parsimonious non-linear time series modeling. **Engineering Applications of Computational Fluid Mechanics**, v. 14, n. 1, p. 1351–1372, 1 jan. 2020.

WHITE, J. C. et al. Remote Sensing Technologies for Enhancing Forest Inventories: A Review. **Canadian Journal of Remote Sensing**, v. 42, n. 5, p. 619–641, 2 set. 2016.

WU, C. et al. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. **Journal of Applied Remote Sensing**, v. 10, n. 3, p. 035010, 8 ago. 2016.

XIAN, S.; CHEN, K.; CHENG, Y. Improved seagull optimization algorithm of partition and XGBoost of prediction for fuzzy time series forecasting of COVID-19 daily confirmed. **Advances in Engineering Software**, v. 173, p. 103212, nov. 2022.

ZHANG, H.; SRINIVASAN, R.; YANG, X. Simulation and Analysis of Indoor Air Quality in Florida Using Time Series Regression (TSR) and Artificial Neural Networks (ANN) Models. **Symmetry**, v. 13, n. 6, p. 952, 27 maio 2021.

ZHOU, Y.; FENG, Z. Estimation of Forest Stock Volume Using Sentinel-2 MSI, Landsat 8 OLI Imagery and Forest Inventory Data. **Forests**, v. 14, n. 7, p. 1345, 29 jun. 2023.

ZHU, X. et al. Prediction of rockhead using a hybrid N-XGBoost machine learning framework. **Journal of Rock Mechanics and Geotechnical Engineering**, v. 13, n. 6, p. 1231–1245, dez. 2021.

ZHU, Y. et al. Estimation of Forest Biomass in Beijing (China) Using Multisource Remote Sensing and Forest Inventory Data. **Forests**, v. 11, n. 2, p. 163, 31 jan. 2020.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - PROJEÇÃO DO COMPORTAMENTO ESPECTRAL PARA A DETECÇÃO PRECOCE DE ANOMALIAS EM FLORESTAS CLONAIAS DE EUCALIPTO

PROJECTION OF SPECTRAL BEHAVIOR FOR EARLY DETECTION OF ANOMALIES IN CLONAL EUCALYPTUS FORESTS

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

RESUMO

Florestas desempenham um papel crucial na economia global, mas estão sujeitas a diversos fatores que podem afetar negativamente seu crescimento e produção. Notadamente, anomalias florestais podem acarretar impactos irreversíveis quanto ao crescimento das florestas. Assim, o monitoramento precoce destas anomalias é crucial. O estudo apresenta uma abordagem baseada em *machine learning* (ML) para a sua detecção precoce em florestas clonais de eucalipto no Brasil. Os dados dos plantios comerciais foram distribuídos segundo um amplo gradiente edafoclimático. Imagem de satélite (Sentinel-2) e modelos autorregressivos de ML (XGBoost e *Random Forest* (RF) foram utilizados para prever a ocorrência do evento. O método aplicado no presente estudo parte da premissa que se existir discrepância entre o NDVI avaliado no momento presente *versus* NDVI predito por ML há o surgimento de anomalia florestal. Num primeiro momento, avaliou-se a precisão dos modelos de XGBoost e RF para definição de qual era o melhor. A partir disso, o modelo selecionado foi expandido com o uso de regressão quantílica. De maneira geral, os resultados demonstram que o método XGBoost expandido, com uso da regressão quantílica, foi capaz de detectar anomalias em florestas de eucalipto com precisão e no estágio inicial de ocorrência das mesmas. A abordagem aqui proposta permite o fornecimento de informações cruciais para apoiar a tomada de decisão florestal, no que tange controle operacional em campo direcionado a mitigação de perda de crescimento florestal, contribuindo assim para uma gestão mais eficaz e sustentável.

Palavras-chave: *Random Forest*; XGBoost; modelo autorregressivo; silvicultura.

ABSTRACT

Forests play a crucial role in the global economy, but they are subject to several factors that can negatively affect their growth and production. Notably, forest anomalies can have irreversible impacts on forest growth. Early monitoring of these anomalies is crucial, so the present study presents an approach based on machine learning (ML) for their early detection in clonal eucalyptus forests in Brazil. Data from clonal eucalyptus forests distributed across a broad soil-climatic gradient, Sentinel-2 data and autoregressive ML models XGBoost and Random Forest (RF) were used to identify forest anomalies early. The method applied in the present study is based on the premise that if there is a discrepancy between the NDVI assessed at present versus the NDVI predicted by ML, a forest anomaly will emerge. Initially, the accuracy of the XGBoost and RF models was evaluated to determine which was the best. From this, the selected model was expanded using quantile regression. In general, the results demonstrate that the ML model XGBoost expanded using quantile regression and was able to detect anomalies in eucalyptus forests accurately and at the initial stage of their occurrence. The approach proposed here allows the provision of crucial information to support forestry decision-making regarding operational control in the field aimed at mitigating loss of forest growth, thus contributing to more effective and sustainable management.

Keywords: Random Forest; XGBoost; autoregressive model; eucalyptus.

1 INTRODUÇÃO

Nos últimos anos observa-se uma intensificação de oscilação climática entre períodos de ocorrência de La Niña e El Niño. Essas variações de temperatura são cada vez mais extremas devido a esses fenômenos e mudanças climáticas (CORDERO et al., 2024; LIU et al., 2023; PASTANA et al., 2021; SANTOS et al., 2021). Identifica-se uma maior frequência e magnitude, o que acarreta grandes impactos à sociedade, e consequentemente ao agronegócio por condições climáticas extremas (vendavais, crises hídricas, temperaturas altas etc.) (DUARTE; RIBEIRO, 2023; PASTANA et al., 2021). Nesse contexto de maiores variações climáticas, as florestas de eucalipto se tornam susceptíveis não somente a redução de crescimento, como também ao surgimento das mais variadas anomalias (DUARTE; RIBEIRO, 2023).

De maneira geral, essas condições são causadas por condições bióticas e abióticas que não correspondem ao padrão esperado de crescimento de uma floresta, podendo ainda indiretamente aumentar a susceptibilidade a doenças, pragas e incêndios. Logo, detectar precocemente quaisquer anomalias é um fator chave para mitigar as perdas produtivas, socioambientais e econômicas. Na literatura há diversos estudos e aplicações para detecção de anomalias, que transitam desde ciências da saúde (RACETIN; KRTALIĆ, 2021), segurança de rede (LIM et al., 2024), sensoriamento remoto até a área florestal (GINO et al., 2023; MCCANN et al., 2017; REMBOLD et al., 2013). Entretanto, há ainda uma considerável lacuna quando avaliado as florestas clonais de eucalipto, principalmente aos métodos empregados e a forma de modelar o problema.

O preenchimento dessa lacuna requer um aprofundamento envolvendo a integração entre a análise de séries temporais e algoritmos de reconhecimento de padrões. Logo, o uso de informação de sensoriamento remoto pode desempenhar um papel fundamental na dinâmica das anomalias, mas principalmente na forma como esse evento surge temporalmente, permitindo assim uma gestão eficiente e direcionado ao longo do tempo (WU et al., 2020). Em séries temporais, modelos autorregressivos são ferramentas estatísticas que permitem, por exemplo, a modelagem da altura do dossel florestal em nível de povoamento, assim como a estimativas de crescimento e produtiva florestal (AHMED et al., 2015; CZERWINSKI; KING; MITCHELL, 2014; SANTOS et al., 2023; SHARMA et al., 2011). O uso de modelos lineares como a média móvel integrada autorregressiva (ARIMA) e de médias móveis integrativas autorregressivas sazonais (SARIMA) apresentam precisão desejável na capacidade preditiva (DABRAL; MURRY, 2017; WANG et al., 2015; XU; CHAN; ZHANG, 2019; ZHANG, 2003).

Além dos modelos autorregressivos lineares, técnicas de aprendizado de máquina, como *Random Forest* (RF) e *extreme gradient boosting* (XGBoost) são amplamente utilizados para modelar séries temporais, devido a sua capacidade de generalização e desempenho satisfatório (CHEN et al., 2017; KARASU; ALTAN, 2019; QIU et al., 2017; ZAMANI JOHARESTANI et al., 2019) e XGBoost (WANG; GUO, 2020; XIAN; CHEN; CHENG, 2022b; ZHANG; HU; YANG, 2022).

Dado à sensibilidade da floresta de eucalipto, a qualquer oscilação climática (SCOLFORO et al., 2020), por exemplo, apenas as séries temporais com alta frequência (por exemplo, 5 dias para Sentinel-2) poderão verdadeiramente descrever cada processo de mudança de um índice de vegetação, capturando assim as mudanças abruptas na paisagem (WU et al., 2020). Em escalas temporais mais longas, a detecção de anomalias em florestas de eucalipto tende a ser imprecisas devido à velocidade de surgimento de fatores de confundimento, como plantas daninhas (KOBAYASHI et al., 2024).

Le Maire et al. (2024) apresenta um sistema dinâmico denominado Sistema de Alerta de Dossel Florestal (FOCAS) para a detecção de anomalias usando Random Forest. No presente estudo, exploramos melhorias e novos modelos para detectar anomalias desde o momento de sua ocorrência. Este trabalho aprimora significativamente o sistema dinâmico FOCAS, permitindo prever a evolução dos pixels com maior precisão. A seleção do melhor modelo para a predição do baseline, a análise detalhada do efeito da adição da regressão quantílica aos modelos autorregressivos e a aplicação dos modelos finais a grandes áreas, abrangendo diversas condições climáticas do Norte ao Sul do Brasil, foram alguns dos avanços obtidos. Este estudo representa um avanço significativo na detecção precoce e na resposta eficaz à detecção de anomalias, contribuindo para o aumento da sustentabilidade e para a mitigação da perda de crescimento das florestas.

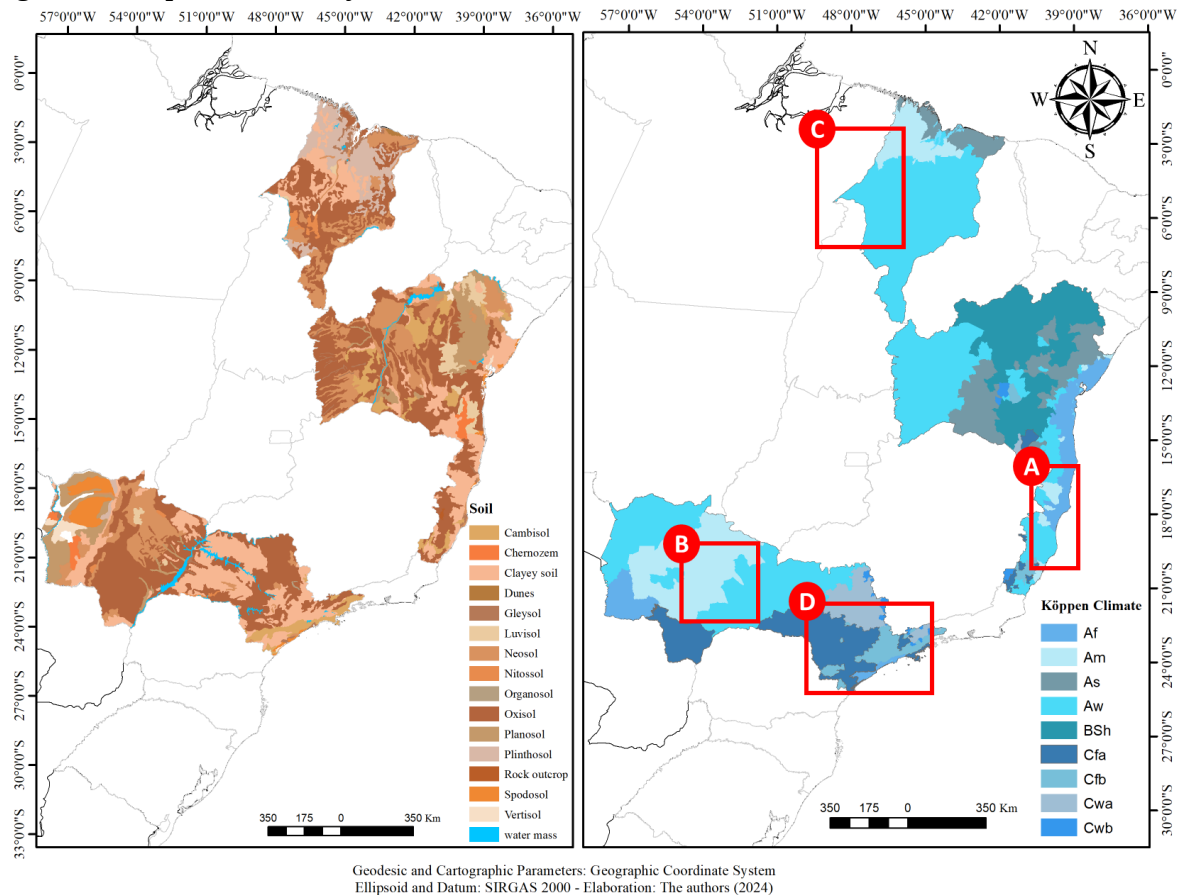
2 MATERIAL E MÉTODOS

2.1 Caracterização da área de estudo

A área do estudo é composta por florestas clonais de eucalipto, distribuídas em 4 regiões do Brasil. Essas regiões apresentam um clima diversificado, aqui rotulado como A, B, C e D (Figura 1). De acordo com a classificação de Köppen, os principais tipos climáticos incluem: Clima Tropical de Savana (Aw), Clima Tropical Úmido ou Subúmido (Am), Clima Tropical Equatorial (Af), Clima Subtropical Úmido (Cfa), Clima Subtropical Oceânico (Cfb), Clima

Subtropical Úmido no Inverno Seco (Cwa) e Clima Subtropical de Altitude (Cwb) (ALVARES et al., 2013). Já a precipitação anual nessas regiões varia entre 1000 e 1600 mm, com temperaturas médias anuais entre 24 e 28°C. Quanto ao solo, há predominância de Latossolos, Argissolos, Neossolos e Cambissolos. A altitude varia de 3 a 1150 metros acima do nível do mar.

Figura 3 - Mapa de localização da área de estudo.



Fonte: Do autor (2024).

2.2 Amostragem e caracterização das anomalias florestais

Foram levantadas informações das diferentes florestas em relação aos seus materiais genéticos, altitude do terreno, espaçamento entre plantas, idade do plantio, ciclo de rotação do plantio e variáveis ambientais, divididas em zona climáticas, risco climático e zona fisiográfica. A categoria de zona climática é baseada no índice de aridez, que avalia o balanço hídrico atmosférico pela relação entre chuva e evapotranspiração. A zona de risco climático é definida pela precipitação média anual e sua variabilidade interanual (ALVARES et al., 2023). R1 indica

alta chuva e baixa variabilidade; R2, alta chuva e alta variabilidade; e R4, pouca chuva e alta variabilidade. A zona fisiográfica categoriza a drenagem comparativa e a elevação predominante (STAPE; ALVARES, 2023). A área de estudo apresenta uma ampla variedade de espaçamentos, para trazer capacidade realista de aplicação. A altitude média varia de acordo com a região. Cada região do estudo possui uma quantidade significativa de zonas climáticas e materiais genéticos distintos, conforme demonstrado na Tabela 1.

Tabela 1 - Tabela descritiva das variáveis que compõem o banco de dados florestal.

Região	Altitude Média (m)	Espaçamento (N)	Material Genético (N)	Variáveis Ambientais (N)	Rotação (max)
A	108	130	118	19	4
B	406	208	86	17	3
C	208	107	62	18	3
D	732	122	144	13	4

Em que - N: número de observações diferentes na base de dados florestal.

A ocorrência de anomalias fora coletada periodicamente em campo entre o período de janeiro de 2018 e dezembro de 2022. Uma vez constatada a ocorrência de anomalia, ela compôs o banco de dados de anomalias, somando 4372 observações para os quatros sites. O objetivo era caracterizar o padrão ocasionado por diferentes tipos de anomalias que podem afetar florestas de eucalipto nas regiões de estudo. As anomalias que podem impactar as florestas de eucalipto incluem pragas, geada, déficit hídrico (seca), incêndio de madeira em pé, vendavais, furto de madeira em pé, enchentes, doenças, problemas nutricionais, fitotoxicidade/queima por herbicida, etc. A compreensão dessas sazonalidades e a tendência/frequência dos eventos é crucial para o reconhecimento de padrões em séries temporais, permitindo definir anomalias potenciais com maior precisão. A validação de campo para detectar qual a real anomalia observada, se deu através de laudo técnico para a confirmação da anomalia e tipo de a ocorrência.

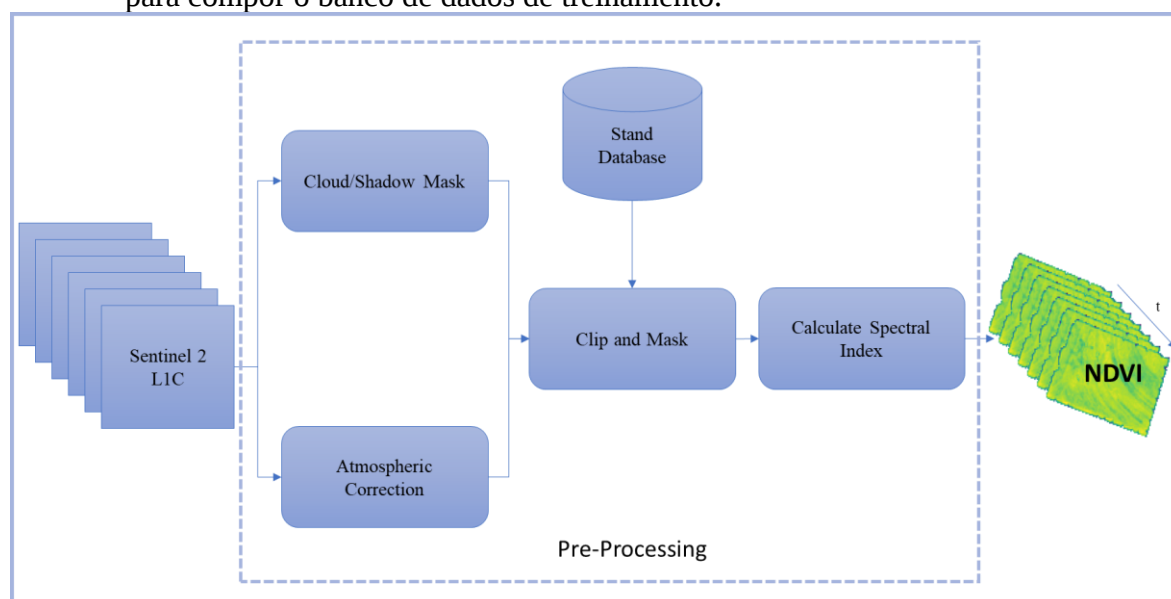
2.3 Estruturação dos dados espectrais

Primeiramente, foi realizado um recorte e uma máscara para as áreas de ocorrência dos plantios florestais. A obtenção das variáveis de sensoriamento remoto se deu através do ambiente de Sistemas de Informações Geográficas (SIG), utilizando imagens do satélite Sentinel-2 L1C da COPERNICUS/ESA. Foi adquirida uma série temporal de cada floresta ao nível de pixel, a partir de 1 ano de idade até o final do ciclo de rotação, variando no intervalo

de tempo de 2018 a 2023. Essas imagens se sobrepõem às parcelas monitoradas ao longo do tempo. As aquisições do Sentinel-2 compreendem 13 bandas espectrais com resoluções espaciais que variam de 10 m a 60 m. Neste estudo, utilizamos imagens corrigidas atmosféricamente, correspondentes a produtos de refletância de superfície de Nível 2A, também conhecidos como refletância de Fundo da Atmosfera (BOA). A consistência temporal da refletância do BOA é crucial para a análise de séries temporais, permitindo identificar mudanças sutis nas propriedades da vegetação e minimizando as variações causadas por fatores atmosféricos, como vapor d'água e aerossóis. A eliminação de dados ruidosos foi através de correções radiométricas e atmosféricas, bem como um tratamento prévio na remoção de nuvens. O Sen2Cor é comumente usado para corrigir imagens do Sentinel-2 (LOUIS et al., 2016), enquanto o LaSRC é utilizado para imagens Landsat (VERMOTE et al., 2016). Pesquisas recentes indicam que uma versão adaptada do LaSRC para o Sentinel-2 oferece melhor consistência temporal da refletância na vegetação tropical (CHRAIBI et al., 2022). Portanto, neste estudo, utilizamos o LaSRC para corrigir as imagens do Sentinel-2. A remoção de nuvens é um processo crítico no tratamento inicial das imagens, sendo utilizado a função S2cloudless (ANZE ZUPANC, 2017; SKAKUN et al., 2022) que automatiza essa operação e gera as melhores respostas (GAO et al., 2024).

A seleção do índice de vegetação selecionado se deu através da sua ligação com características florestais, tais como Índice de Área Foliar, teor de clorofila do dossel ou conteúdo de água do dossel (GUERRIC LE MAIRE et al., 2024). Para compor a estrutura do banco de dados, e formação do mosaico de imagens, foi calculado o NDVI – *Normalized Difference Vegetation Index* (ROUSE JR, 1973), que é uma das variáveis mais interessantes para monitoramento da vegetação florestal (LE MAIRE et al., 2011; MA et al., 2020; ZHU; LIU, 2015). Para complementar o banco de dados espectral, foram inseridas sete variáveis características das imagens Sentinel-2. Assim, foram incluídas o dia de aquisição dentro do ano (DoYS2), porcentagem de nuvens da imagem (cloud_cover), o ângulo zenital e azimute do sol no momento da aquisição da imagem (Sun_Zenith, Sun_Azimuth) e a incidência de visualização, o ângulo zenital de aquisição da imagem (View_Zenith, View_Azimuth), e o ângulo azimute relativo entre o sol e os ângulos de visão (Relaz). Todas essas etapas, foram desenvolvidas com software R (R CORE TEAM, 2024), e estão ilustradas na figura 2.

Figura 4 - Fluxograma ilustrativo das etapas para o pré-processamento das imagens Sentinel 2 para compor o banco de dados de treinamento.



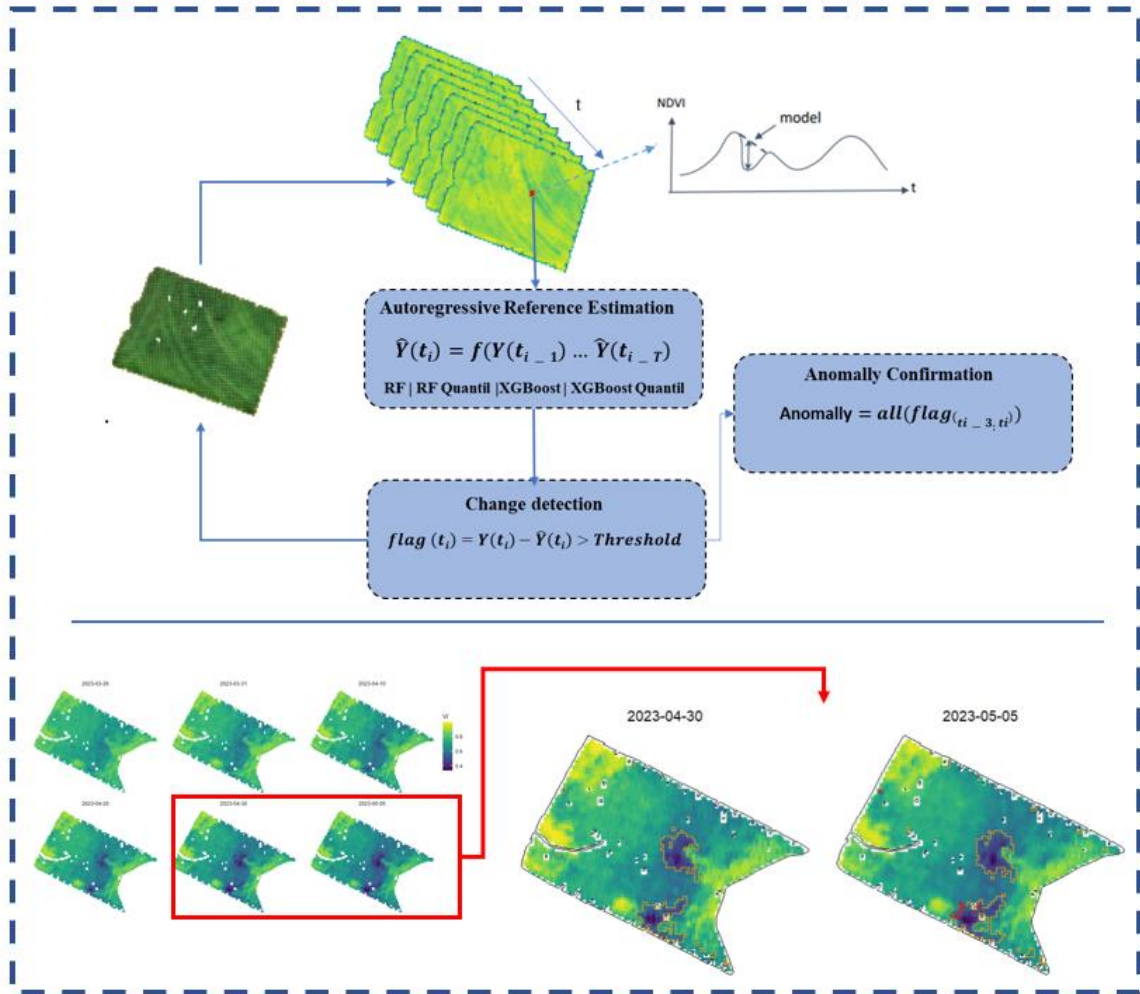
Fonte: Do autor (2024).

2.4 Sistema de detecção de anomalias florestais

Para detecção de anomalias, é necessário treinar um modelo autorregressivo, que permite previsões dos valores dos pixels no caso normal, sem anomalias. Para isso, usou-se como base, o Sistema de Alerta de Dossel Florestal (FOCAS) (LE MAIRE et al., 2024). Essa previsão forma a “baseline”, que pode ter também um intervalo estabelecido, e realizar inferências se o valor observado está dentro de um intervalo aderente (neste caso a observação é considerada “normal”), ou ao contrário se o valor é fora (observação considerada como anomalia). Assim, foram treinados modelos de *machine learning* para prever o comportamento das florestas dentro de cada intervalo de tempo.

Os modelos de série temporal *Random Forest* (RF) e o *XGBoost* foram escolhidos para o estudo, sendo posteriormente expandidos com regressão quantílica. Esses dois modelos de aprendizado de máquina foram selecionados por serem flexíveis e eficientes na computação, além de serem capazes de trabalhar com valores em série temporal. Já suas expansões com a regressão quantílica foi realizada para permitir além das estimativas pontuais, estimativas em diferentes níveis de probabilidade. A Figura 3 ilustra e descreve os procedimentos metodológicos do estudo.

Figura 5 - Fluxograma da metodologia para detecção de anomalia.



Fonte: Do autor (2024).

Treinou-se modelos autorregressivos utilizando como variáveis de entrada os valores do Índice de Vegetação por Diferença Normalizada (NDVI) pixel a pixel, ângulo zenital e azimuth da imagem Sentinel 2, material genético, altitude do terreno, espaçamento entre plantas, idade do plantio, zona climática, densidade populacional e ciclo de rotação de rotação do plantio (LE MAIRE et al., 2024). A equação (1) demonstra a forma genérica de um modelo autorregressivo. A modelo objetiva estimar o NDVI esperado (segundo padrão normal de uma floresta sem anomalia) para a data de coleta da imagem presente. Espera-se assim, que o valor observado do pixel, esteja próximo a esta estimativa.

$$\hat{Y}(t_i) = f(Y(t_{i-1}) \dots Y(t_{i-n})) \quad (1)$$

Em que: $\hat{Y}$: índice espectral estimado para o tempo t para o pixel i . Y : índice espectral observado para o tempo t para o pixel i na iteração n .

Para o treinamento do modelo autorregressivo, foi escolhido um lag (atraso) de 50 dias para o NDVI. Isso permite estabelecer uma previsão para os próximos 50 dias. Para identificar uma provável anomalia, é necessário que o valor observado do pixel esteja abaixo de um limite previamente estabelecido em relação ao NDVI esperado (Equação 2). Quando a diferença entre o NDVI observado e o esperado ultrapassa 0,15, gera-se uma sinalização de anomalia florestal. Esse valor de limiar foi determinado através de testes preliminares, que permitiram estabelecer um valor confiável para a confirmação da anomalia. A anomalia é confirmada após três sinalizações sucessivas em três imagens consecutivas dentro da linha temporal.

$$Flag(t_i) = |Y(t_i) - \hat{Y}(t_i)| > Threshold \quad (2)$$

Em que: $\hat{Y}$: índice espectral estimado para o tempo t para o pixel i no percentil 10. Y : índice espectral observado para o tempo t para o pixel i . $Flag$: provável anomalia para o tempo t para o pixel i . $Threshold$: limite determinado para apontar como anomalia.

Além de testar um valor fixo de limiar para a detecção de anomalias, foi realizada uma expansão da metodologia usando a regressão quantílica. Esta abordagem permite obter, além das estimativas pontuais, estimativas em diferentes níveis de probabilidade, fornecendo intervalos de predição. Enquanto os algoritmos de regressão RF e XGBoost padrões fornecem apenas estimativas pontuais dos valores da variável dependente, a regressão quantílica é utilizada para derivar intervalos de predição de um modelo base (HOANG; TRAN; TRAN, 2024a; MEINSHAUSEN, 2006). Este estudo recorreu à regressão quantílica para expressar uma relação funcional entre as variáveis independentes (ou seja, NDVI, variáveis climáticas e cadastrais dos plantios florestais) e a dependente, para os quantis 0,1, 0,5 e 0,9 de uma distribuição de probabilidade (Equação 3). Caso o valor de NDVI observado esteja abaixo da previsão do NDVI de 10%, gera-se uma sinalização de anomalia florestal. A anomalia é confirmada após três sinalizações sucessivas em três imagens consecutivas dentro da linha temporal. Assim, o intervalo de previsão de 10% para a variável dependente pode ser expresso por:

$$IP(x) = [Q_{0,1}(x), Q_{0,9}(x)] \quad (3)$$

Em que: IP : intervalo de previsão; $Q_{0,1}(x)$ e $Q_{0,9}(x)$ são as predições dos dois modelos XGBoost ou RF treinados com os parâmetros de regressão quantílica de 0,1 e 0,9, respectivamente.

2.5 Modelos Autoregressivo aplicados

2.5.1 *Random Forest (RF)*

O RF é um algoritmo de aprendizado de máquina introduzido por (BREIMAN, 2001b). O método é utilizado na classificação não paramétrica padrão e regressão para construir regras de predição com base em um número alto de variáveis preditoras sem análise prévia (PROBST; WRIGHT; BOULESTEIX, 2019). O RF possui vários hiperparâmetros que controlam a estrutura de cada árvore individual, a estrutura e alto número de árvores de regressão, bem como a sua aleatoriedade (PROBST; WRIGHT; BOULESTEIX, 2019). Os principais hiperparâmetros que controlam o RF são: *ntree* (número de árvores treinadas na floresta), *nodesize* (tamanho do nó do terminal de destino) e *mtry* (número de variáveis escolhidas para construir uma árvore de regressão (ou classificação)). Os dados são divididos em dois níveis de randomização, dos quais dois terços das observações são usadas para o treinamento das árvores e o outro um terço restante dos dados são usados na validação, o primeiro é chamado de agregação de *bootstrap* (“ensacamento”), e o segundo de *out-of-bag* (OOB) (WOZNICKI et al., 2019). Neste segundo nível, o RF cria muitas árvores de decisão e treina as versões reamostradas dos dados originais (de forma semelhante ao “*bagging*”), seleciona dentro desse conjunto de variáveis de forma aleatória um pequeno número de recursos que podem ser inspecionados em cada nó para divisão (GOMES et al., 2017). Cada nó é avaliado ao se utilizar métricas de erro. A estimativa do erro das amostras OOB pelas respectivas árvores e pela agregação das previsões e do erro médio quadrático médio (MSE_{OOB}) é calculado usando a equação (4) (WANG et al., 2018). O critério de parada do RF é determinado pelo *nodesize*, pois ele determina até onde cada árvore de decisão crescerá. A determinação da resposta otimizada pela criação das árvores de decisão é feita pela pontuação da estimativa de erro, a que receber o menor valor de MSE_{OOB} , ou a que mais receber votos em problemas de classificação (BELGIU; DRĂGU, 2016). Para o treinamento do modelo utilizado na metodologia, utilizou o pacote “*ranger*” do R (WRIGHT; ZIEGLER, 2017)

$$MSE_{OOB} = \frac{1}{n} \sum_{i=1}^n (O_i - P_{iOOB})^2 \quad (4)$$

Onde n : número de observações; O_i : amostra observada; P_{iOOB} : média de todas as previsões $_{OOB}$ em todas as árvores.

2.5.2 *Extreme Gradient Boosting Tree – XGBoost*

O XGBoost é um algoritmo de aprendizado de máquina introduzido por Chen et al. (2015). Dada sua flexibilidade e processamento automático de valores ausentes, é amplamente utilizado em diversos cenários de mineração de dados e competições de algoritmos (WEI; ZENG, 2021). O XGBoost é um sistema escalonável, em que combina vários classificadores mais fracos em um classificador de forma linear. Ou seja, o XGBoost é caracterizado com uma técnica *ensemble learning*, onde várias árvores de decisão mais fracas são combinadas para formar um modelo mais forte. Ela utiliza como estratégia a soma da função de perda do modelo anterior, e um termo regularizador, assim, a cada iteração do treinamento, uma nova árvore é adicionada ao modelo para minimizar a função de perda global. A função de minimização para o objetivo regularizado do XGBoost (5).

$$Obj(\theta) = \sum_{i=1}^n L\left(Y_i, \hat{Y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (5)$$

Em que: θ : parâmetros do modelo. L : função de perda. $\hat{Y}$: valor estimado para a iteração $t - 1$. Y : valor observado. $f_t(x_i)$: valor estimado da árvore na iteração t . $\Omega(f_t)$: termo de regularização que penaliza modelos mais complexos.

O objetivo é determinar os parâmetros θ que reduzam a função de perda. Isso é feito por meio de um processo iterativo em que uma árvore adicional é adicionada ao modelo para reduzir a função de perda global. O XGBoost encontra a melhor divisão em cada nó da árvore usando uma abordagem de aproximação, tornando-o mais eficiente computacionalmente do que outros métodos baseados em árvores. Para implementação do modelo, utilizou pacote “xgboost” do R (CHEN et al., 2024).

2.5.3 *Análise da performance dos modelos*

Ao todo, foram utilizadas informações da série temporal de 2.076 talhões para o treinamento dos modelos. Para a avaliação dos modelos de previsão dos valores da variável de interesse (NDVI), os indicadores utilizados incluíram o coeficiente de determinação (R^2) (Equação 5), o Bias (Equação 6) e a raiz do erro quadrático médio (RMSE) (Equação 7). O método empregado foi uma classificação não supervisionada.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (5)$$

$$Bias = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)}{n} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (7)$$

Em que: Y_i é o valor observado da variável de interesse; $\hat{Y}_i$ é o valor estimado da variável de interesse; $\bar{Y}$ é o valor médio da variável de interesse.

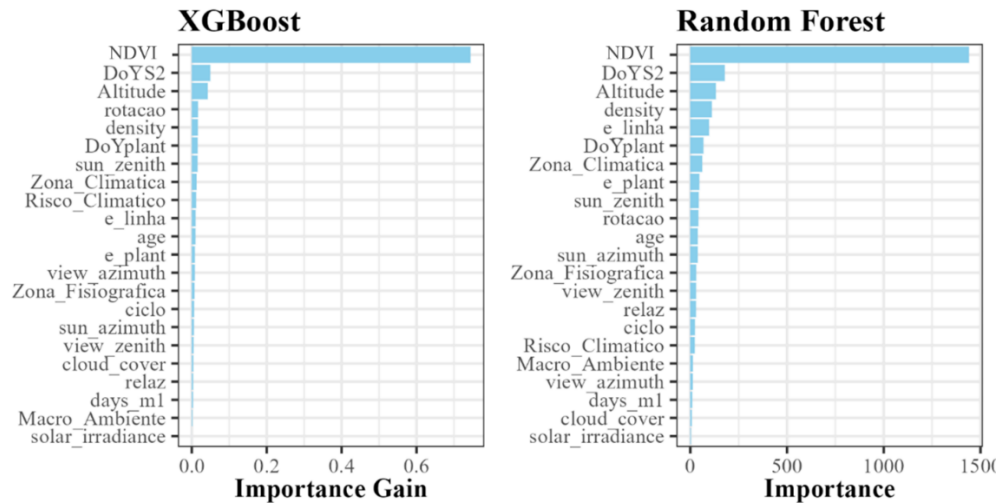
Para validar a detecção de anomalias, foi realizado um levantamento em 50 talhões ao longo de 2024. Após o monitoramento, os pontos identificados como anomalias foram visitados e avaliados usando uma matriz de confusão para estimar a taxa de acertos. A validação focou no melhor modelo para determinar sua taxa de acerto, analisando tanto a estratégia com um valor limiar fixo quanto com o percentil de 10%. Quando uma anomalia detectada correspondia a uma ocorrência real, era registrada como verdadeiro positivo; caso contrário, como falso positivo. Anomalias reais não detectadas eram classificadas como falsos negativos, e a ausência correta de anomalias, como verdadeiros negativos. Essa abordagem permitirá uma análise detalhada da eficácia das estratégias de detecção de anomalias, fornecendo uma visão clara de seu desempenho e confiabilidade.

3 RESULTADOS

Os resultados do processo de cálculo para os modelos autorregressivos para os dados de treinamento são apresentados na Tabela 2. Os modelos foram calibrados nos 4 locais, e apresentaram precisão equivalente. Para o modelo RF, foram utilizados os parâmetros *mtry* igual a 8 e *ntree* igual a 500. Este modelo apresentou um R^2 de 0.94, um MSE de 0.0004, e um RMSE de 0.021. Enquanto o modelo XGBoost, foram utilizados um número de rounds igual a 2000, também apresentou um R^2 de 0.94, um MSE de 0.0004 e um RMSE de 0.021. Os modelos de detecção de anomalias apresentaram alta precisão, conforme estatísticas avaliadas. Esses resultados indicam que ambos os modelos são eficazes para a detecção de anomalias, com desempenho semelhante nas métricas avaliadas. Já a figura 4, apresenta a importância de cada variável para o RF e XGBoost. Nota-se que as variáveis mais relevantes em ambos os modelos foram NDVI com lag de 50 dias, altitude e DoYS2. A altitude é variável que possibilite a

discriminação entre local/clima. Outras variáveis também contribuem, porém com grau de importância consideravelmente menor entre os modelos.

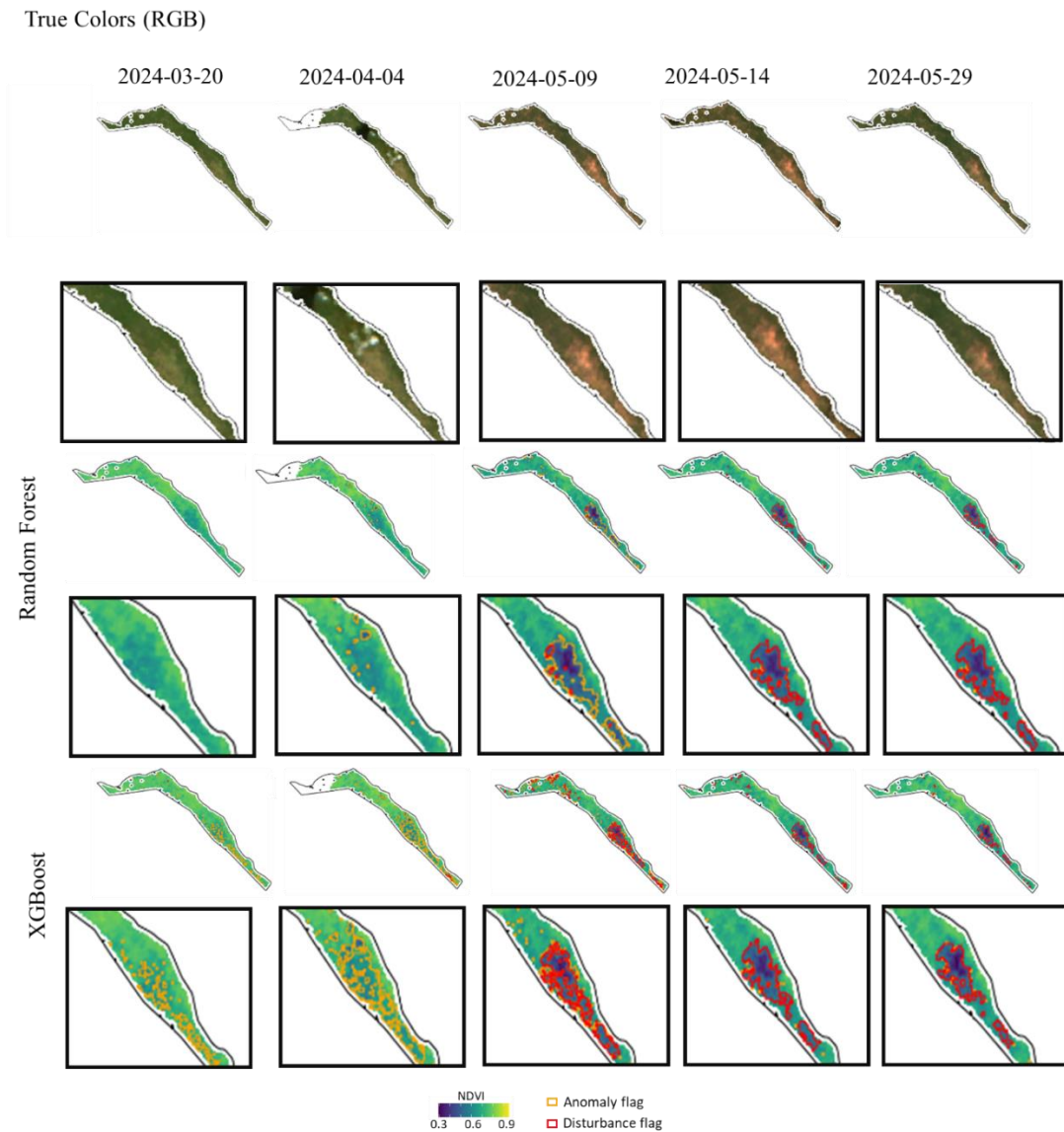
Figura 6 - Valores médios de importância das variáveis explicativas dos modelos RF e XGBoost para prever o valor de NDVI.



Fonte: Do autor (2024).

Uma vez que a precisão de ambos os modelos foi próxima, a definição do melhor modelo a ser expandido através da regressão quantílica foi realizado com base na capacidade de qual modelo identifica de mais precocemente uma anomalia florestal. Utilizou-se nesta etapa a base de dados coletada em campo das anomalias florestais. A figura 5 elenca alguns exemplos, onde percebe-se que XGBoost identifica a presença de anomalia já na primeira ocorrência, enquanto RF tende a necessitar de uma maior área de ocorrência de anomalia para fazer a mesma detecção. Resumidamente, foi diagnosticado que XGBoost consegue identificar uma anomalia florestal já na primeira imagem com ocorrência da mesma, enquanto RF necessita de 3 imagens, em média. Sendo assim, XGBoost foi definido como o modelo mais flexível para identificação de anomalias precoces.

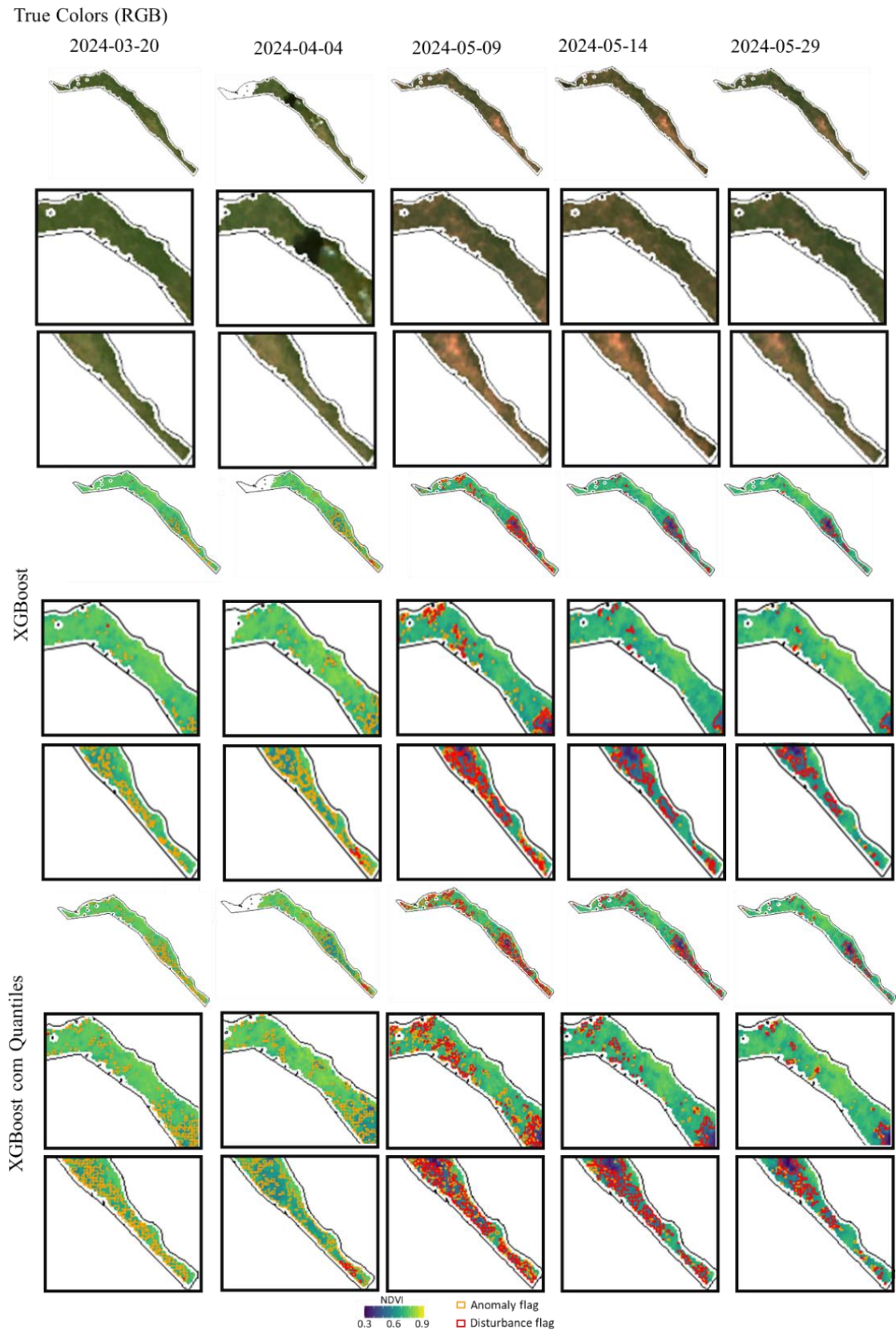
Figura 7 – Dinâmica das mudanças e detecção de anomalias para um talhão com RF e XGBoost.



Fonte: Do autor (2024).

O XGBoost padrão é capaz de fornecer estimativas pontuais dos valores de NDVI. Embora a seção anterior demonstre que este modelo pode atingir resultados de predição excelentes, ele não pode fornecer informações sobre a dispersão das predições em torno do valor de NDVI. A figura 6 apresenta um comparativo entre o XGBoost e XGBoost com quantiles (XGBoostQ). Na parte superior do plantio ocorre a desfolha ocasionada por formigas, o XGBoostQ é mais sensível a perturbações ocorrida no intervalo de tempo, identificando melhor a área de ocorrência da anomalia.

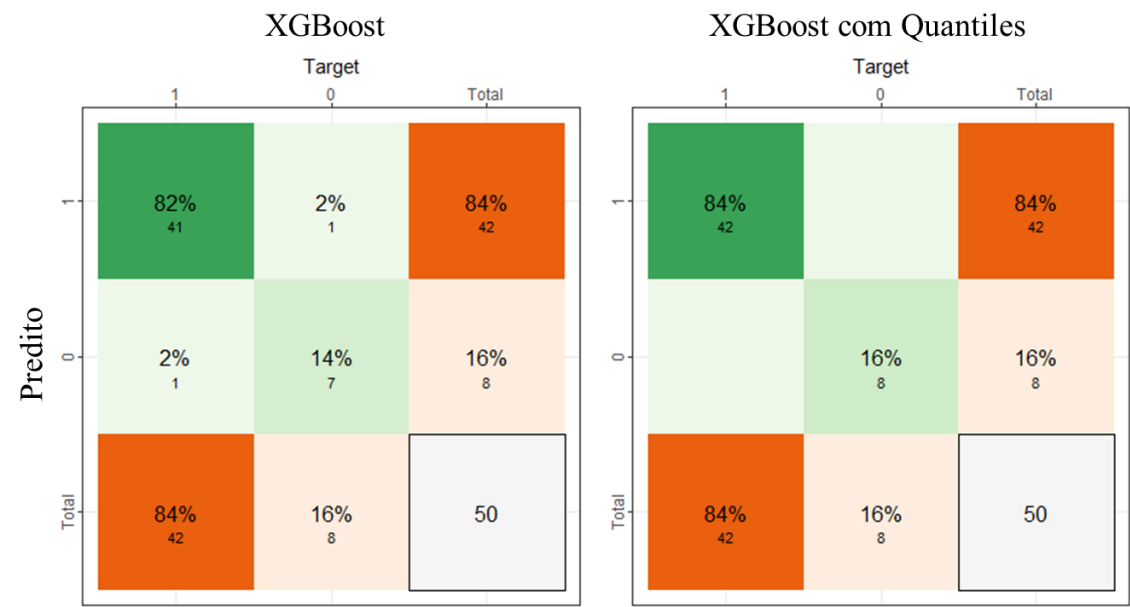
Figura 8- Mapa com mudanças e detecção de anomalias para um talhão com XGBoost e XGBoost com quantiles.



Fonte: Do autor (2024).

A Figura 7 apresenta a matriz de confusão para os 50 talhões com validação realizada em campo, onde a classificação do modelo XGBoost e XGBoost foram contrastados. Como pode ser observado, ambos os modelos apresentaram uma taxa de sucesso acima de 95%, acertando. A incorporação da regressão em quantis permite a incorporação do componente probabilístico para a detecção de anomalias. A introdução do componente de incerteza,

Figura 9 - Matriz de confusão com os dados de validação para XGBoost e XGBoost com quantiles.



Fonte: Do autor (2024).

4 DISCUSSÃO

Florestas clonais de eucalipto estão sujeitas a uma série de fatores que podem impactar seu crescimento e produtividade. A identificação precoce de qualquer irregularidade ou anomalia nas florestais torna-se então crucial para a adoção de medidas de mitigação. Devido ao vasto território das florestas, o monitoramento por equipes de campo torna-se inviável. Neste estudo, modelos de ML foram aplicados para prever o valor do NDVI na imagem presente, permitindo diagnosticar se há alguma anomalia na área de interesse e determinar a área de ocorrência das mesmas. Os resultados mostraram que os modelos foram eficientes na determinação de anomalias, sendo o XGBoost mais flexível e antecipativo quando comparado ao RF. O modelo XGBoost é mais sensível a inclusão de variáveis preditoras, enquanto o

algoritmo RF é menos influenciável pela inclusão destas variáveis (DÍAZ-URIARTE; ALVAREZ DE ANDRÉS, 2006; GEORGANOS et al., 2018; LI et al., 2020; RAMDANI; FURQON, 2022).

Detectar anomalias em florestas é crucial e requer compreensão de como elas se manifestam, sua periodicidade e a escala espacial em que são observáveis. Por exemplo, a incidência de incêndio, ocasiona quedas abruptas e sincronizadas nos valores de índices de vegetação, enquanto a exploração madeireira tende a resultar num declínio gradual de forma isolada. A seca e os danos provocados pelos insetos também levam a um declínio gradual da vegetação, porém, de forma sincronizada em regiões relativamente extensas. Embora o conhecimento sobre agentes causadores, como pragas, possa ajudar na identificação de árvores danificadas, é igualmente importante reconhecer que outras causas, como o estresse, podem gerar respostas fisiológicas semelhantes. Portanto, entender a aparência das árvores saudáveis é fundamental para interpretar a presença de anomalias em dados de sensoriamento remoto (HALL et al., 2016).

Além disso, os modelos desenvolvidos para detecção de eventos anormais requerem cobertura espacial abrangente e localmente específica. Condições climáticas adversas, como nuvens e neblina, podem limitar a disponibilidade de dados e prejudicar a precisão das análises (WHITE et al., 2014). Portanto, é essencial ter um entendimento detalhado dos valores dos pixels e monitorar mudanças abruptas nos índices de vegetação ao longo do tempo para identificar possíveis anomalias. A utilização de dados do SENTINEL-2, com sua frequência de captura de imagens de 5 dias e resolução espacial de 10 metros permite a aplicação de técnicas de composição de imagens. O SENTINEL-2 oferece uma frequência de revisita temporal adequada, e proporciona observações livres de efeitos atmosféricos. Isso reduz significativamente o número de lacunas nos dados e minimiza as desvantagens associadas nas composições de imagens e nos resultados de detecção de alterações. (HERMOSILLA et al., 2016). O uso de ML para detectar anomalias florestais tem sido estudado (PEERBHAY et al., 2021; POTIC et al., 2023; SOLÓRZANO; GAO, 2022; SOMCHING et al., 2020). No entanto, este estudo se destaca pela abordagem não supervisionada adotada para classificar eventos anormais. A utilização dos modelos de ML permite determinar com alta precisão o valor previsto do pixel na data da coleta das imagens. Isso significa que o método deste estudo permite detectar desvios em relação a esse modelo durante a detecção (KOPP; PEVNÝ; HOLEŇA, 2020).

Por fim, a adição da regressão quantílica as previsões do modelo de XGBoost melhorou a precisão preditiva para detectar anomalias. A previsão espacial do índice de vegetação para a

grade amostral, obtida pelo modelo XGBoost, permitiu calcular diferentes quantis da distribuição condicional, capturando não apenas a tendência central, mas também a dispersão. Essa amplitude da distribuição possibilita inferir uma queda dentro de um intervalo aceitável e determinar se o valor do índice está dentro dos percentis baixo e alto esperados. Um valor abaixo do percentil 10% indicam uma irregularidade técnica. O uso de regressão quantílica permite encontrar previsões mais assertivas, como encontrado no estudo de Córdoba et al. (2021). O uso de regressão quantílica baseada em XGBoost garante confiabilidade no intervalo de previsão, pois permite obter quantis condicionais quando estimativas pontuais por si só são insuficientes (BAO et al., 2023; HOANG; TRAN; TRAN, 2024b; YU et al., 2023). Com o desenvolvimento de um método de detecção de anomalias para florestas utilizando séries temporais de imagens do Sentinel-2, os próximos desafios incluem a classificação detalhada dessas anomalias em diferentes tipos. Este passo é essencial para compreender os tipos de anomalias detectadas. Isso inclui descobrir a identificação de incêndios florestais, doenças florestais, vendavais ou efeitos de estresse ambiental. Estudos futuros devem se concentrar na utilização de técnicas avançadas de ML para classificar essas anomalias com maior precisão. Eles também devem investigar como integrar dados adicionais, como variáveis meteorológicas e de solo, para melhorar ainda mais a capacidade preditiva dos modelos.

5 CONCLUSÃO

O método proposto neste estudo permite a determinação da presença de uma provável anomalia florestal de maneira precoce. O modelo de XGBoost foi selecionado, utilizando a regressão quantílica, para a detecção de anomalias em florestas a partir de séries temporais de imagens do Sentinel-2. Este modelo não apenas demonstrou alta precisão na identificação de anomalias, mas também ajudou a incorporar a incerteza no processo. A consideração da incerteza permite uma melhor avaliação da variabilidade dos dados, oferecendo uma detecção mais confiável e confiável das anomalias florestais. Esse avanço é essencial para o bom monitoramento e manejo das florestas porque permite identificar e responder a mudanças sutis no ecossistema que poderiam ter sido ignoradas pelas abordagens convencionais.

REFERÊNCIAS

- AHMED, O. S. et al. Characterizing stand-level forest canopy cover and height using Landsat time series, samples of airborne LiDAR, and the Random Forest algorithm. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 101, p. 89–101, mar. 2015.
- ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.
- ALVARES, C. A. et al. Decision-Tree Application to Predict and Spatialize the Wood Productivity Probabilities of Eucalyptus Plantations. **Forests**, v. 14, n. 7, p. 1334, 29 jun. 2023.
- ANZE ZUPANC. **Improving cloud detection with machine learning**. Disponível em: <<https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13>>. Acesso em: 3 jul. 2024.
- BAO, T. et al. **Dynamic XGBoost-based Quantile Predictor for Real-time Electricity Price Forecasting**. 2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG). **Anais...IEEE**, 3 dez. 2023.
- BELGIU, M.; DRĂGU, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- CHEN, G. et al. Multiple Random Forests Modelling for Urban Water Consumption Forecasting. **Water Resources Management**, v. 31, n. 15, p. 4715–4729, 11 dez. 2017.
- CHEN, T. et al. Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.
- CHEN, T. et al. **xgboost: Extreme Gradient Boosting**, 2024.
- CHRAIBI, E. et al. Stability in time and consistency between atmospheric corrections: Assessing the reliability of Sentinel-2 products for biodiversity monitoring in tropical forests. **International Journal of Applied Earth Observation and Geoinformation**, v. 112, p. 102884, ago. 2022.
- CORDERO, R. R. et al. Extreme fire weather in Chile driven by climate change and El Niño–Southern Oscillation (ENSO). **Scientific Reports**, v. 14, n. 1, p. 1974, 23 jan. 2024.
- CÓRDOBA, M. et al. A spatially based quantile regression forest model for mapping rural land values. **Journal of Environmental Management**, v. 289, p. 112509, jul. 2021.
- CZERWINSKI, C. J.; KING, D. J.; MITCHELL, S. W. Mapping forest growth and decline in a temperate mixed forest using temporal trend analysis of Landsat imagery, 1987–2010. **Remote Sensing of Environment**, v. 141, p. 188–200, fev. 2014.

DABRAL, P. P.; MURRY, M. Z. Modelling and Forecasting of Rainfall Time Series Using SARIMA. **Environmental Processes**, v. 4, n. 2, p. 399–419, 10 jun. 2017.

DÍAZ-URIARTE, R.; ALVAREZ DE ANDRÉS, S. Gene selection and classification of microarray data using random forest. **BMC Bioinformatics**, v. 7, n. 1, p. 3, 6 dez. 2006.

DUARTE, M. L.; RIBEIRO, A. Influência do El Niño e La Niña na produtividade de plantios de Eucalipto em distintas regiões no Brasil. **Ciência Florestal**, v. 33, n. 1, p. e61334, 28 mar. 2023.

FRANKLIN, S. E. **Remote Sensing for Sustainable Forest Management**. [s.l.] CRC Press, 2001.

GEORGANOS, S. et al. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. **IEEE Geoscience and Remote Sensing Letters**, v. 15, n. 4, p. 607–611, abr. 2018.

GEORGE S. PURITCH. **Nonvisual remote sensing of trees affected by stress--a review**. Technical Report 30 ed. Victoria, British Columbia, Canada: Canadian Forestry Service, 1981. v. Technical Report 30

GINO, V. L. S. et al. Integrating Unsupervised Machine Intelligence and Anomaly Detection for Spatio-Temporal Dynamic Mapping Using Remote Sensing Image Series. **Sustainability**, v. 15, n. 6, p. 4725, 7 mar. 2023.

GOMES, H. M. et al. Adaptive random forests for evolving data stream classification. **Machine Learning**, v. 106, n. 9–10, p. 1469–1495, 13 out. 2017.

HALL, R. J. et al. Remote sensing of forest pest damage: a review and lessons learned from a Canadian perspective. **The Canadian Entomologist**, v. 148, n. S1, p. S296–S356, 12 ago. 2016.

HERMOSILLA, T. et al. Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring. **International Journal of Digital Earth**, v. 9, n. 11, p. 1035–1054, 13 nov. 2016.

HOANG, N.-D.; TRAN, V.-D.; TRAN, X.-L. Predicting Compressive Strength of High-Performance Concrete Using Hybridization of Nature-Inspired Metaheuristic and Gradient Boosting Machine. **Mathematics**, v. 12, n. 8, p. 1267, 22 abr. 2024a.

HOANG, N.-D.; TRAN, V.-D.; TRAN, X.-L. Predicting Compressive Strength of High-Performance Concrete Using Hybridization of Nature-Inspired Metaheuristic and Gradient Boosting Machine. **Mathematics**, v. 12, n. 8, p. 1267, 22 abr. 2024b.

KARASU, S.; ALTAN, A. **Recognition Model for Solar Radiation Time Series based on Random Forest with Feature Selection Approach**. 2019 11th International Conference on Electrical and Electronics Engineering (ELECO). **Anais...IEEE**, nov. 2019.

KOBAYASHI, S. et al. Analysis of the relationship between L-band SAR backscatter and understory weed density in eucalyptus plantation forests. **GIScience & Remote Sensing**, v. 61, n. 1, 31 dez. 2024.

KOPP, M.; PEVNÝ, T.; HOLEŇA, M. Anomaly explanation with random forests. **Expert Systems with Applications**, v. 149, p. 113187, jul. 2020.

LE MAIRE, G. et al. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. **Remote Sensing of Environment**, v. 115, n. 10, p. 2613–2625, out. 2011.

LE MAIRE, G. et al. Disturbance monitoring in Eucalyptus plantation forest using Sentinel-2 time series. **Remote Sensing of Environment**, v. 2, p. 1–10, 2024.

LI, Y. et al. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. **Scientific Reports**, v. 10, n. 1, p. 9952, 19 jun. 2020.

LIM, W. et al. Future of generative adversarial networks (GAN) for anomaly detection in network security: A review. **Computers & Security**, v. 139, p. 103733, abr. 2024.

LIU, Y. et al. Nonlinear El Niño impacts on the global economy under climate change. **Nature Communications**, v. 14, n. 1, p. 5887, 21 set. 2023.

LOUIS, J. et al. **SENTINEL-2 SEN2COR: L2A Processor for Users**. [s.l.: s.n.].

MA, J. et al. Analyzing Ecological Vulnerability and Vegetation Phenology Response Using NDVI Time Series Data and the BFAST Algorithm. **Remote Sensing**, v. 12, n. 20, p. 3371, 15 out. 2020.

MCCANN, C. et al. Multi-temporal mesoscale hyperspectral data of mixed agricultural and grassland regions for anomaly detection. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 131, p. 121–133, set. 2017.

MEINSHAUSEN, N. Quantile Regression Forests. **J. Mach. Learn. Res.**, v. 7, p. 983–999, 12 jan. 2006.

PASTANA, D. N. B. et al. Strong El Niño reduces fruit production of Brazil-nut trees in the eastern Amazon. **Acta Amazonica**, v. 51, n. 3, p. 270–279, set. 2021.

PEERBHAY, K. et al. Improving the unsupervised mapping of riparian bugweed in commercial forest plantations using hyperspectral data and LiDAR. **Geocarto International**, v. 36, n. 4, p. 465–480, 25 fev. 2021.

PETER A. MURTHA. Remote sensing and vegetation damage: a theory for detection and assessment. **Photogrammetric Engineering and Remote Sensing**, v. 44, n. 9, p. 1147–1158, 1978.

POTIĆ, I. et al. Improving Forest Detection Using Machine Learning and Remote Sensing: A Case Study in Southeastern Serbia. **Applied Sciences**, v. 13, n. 14, p. 8289, 18 jul. 2023.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, 28 maio 2019.

QIU, X. et al. Oblique random forest ensemble via Least Square Estimation for time series forecasting. **Information Sciences**, v. 420, p. 249–262, dez. 2017.

R CORE TEAM. **R: A language and environment for statistical computing**. ViennaR Foundation for Statistical Computing, , 2024.

RACETIN, I.; KRTALIĆ, A. Systematic Review of Anomaly Detection in Hyperspectral Remote Sensing Applications. **Applied Sciences**, v. 11, n. 11, p. 4878, 26 maio 2021.

RAMDANI, F.; FURQON, M. T. The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification. **F1000Research**, v. 11, p. 1069, 20 set. 2022.

REMBOLD, F. et al. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. **Remote Sensing**, v. 5, n. 4, p. 1704–1733, 8 abr. 2013.

ROUSE JR, J. W. ET AL. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. p. 1–8, 1973.

SANTOS, E. B. et al. Spatio-temporal variability of wet and drought events in the Paraná River basin—Brazil and its association with the El Niño—Southern oscillation phenomenon. **International Journal of Climatology**, v. 41, n. 10, p. 4879–4897, 7 ago. 2021.

SANTOS, J. S. et al. Predicting eucalyptus plantation growth and yield using Landsat imagery in Minas Gerais, Brazil. **Ecological Informatics**, v. 75, p. 102120, jul. 2023.

SHARMA, R. P. et al. Modelling dominant height growth from national forest inventory individual tree data with short time series and large age errors. **Forest Ecology and Management**, v. 262, n. 12, p. 2162–2175, dez. 2011.

SKAKUN, S. et al. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. **Remote Sensing of Environment**, v. 274, p. 112990, jun. 2022.

SOLÓRZANO, J. V.; GAO, Y. Forest Disturbance Detection with Seasonal and Trend Model Components and Machine Learning Algorithms. **Remote Sensing**, v. 14, n. 3, p. 803, 8 fev. 2022.

SOMCHING, N. et al. Using machine learning algorithm and landsat time series to identify establishment year of para rubber plantations: a case study in Thalang district, Phuket Island, Thailand. **International Journal of Remote Sensing**, v. 41, n. 23, p. 9075–9100, 1 dez. 2020.

STAPE, J. L.; ALVARES, C. A. Physiographic position drives Eucalyptus productivity in Mato Grosso do Sul, Brazil. **Série Técnica IPEF**, v. 26, n. 48, p. 289–293, 2023.

VERMOTE, E. et al. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. **Remote Sensing of Environment**, v. 185, p. 46–56, nov. 2016.

WANG, B. et al. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. **Ecological Indicators**, v. 88, p. 425–438, maio 2018.

WANG, W. et al. Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition. **Water Resources Management**, v. 29, n. 8, p. 2655–2675, 7 jun. 2015.

WANG, Y.; GUO, Y. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. **China Communications**, v. 17, n. 3, p. 205–221, mar. 2020.

WEI, H.; ZENG, Q. Research on sales Forecast based on XGBoost-LSTM algorithm Model. **Journal of Physics: Conference Series**, v. 1754, n. 1, p. 012191, 1 fev. 2021.

WHITE, J. C. et al. Pixel-Based Image Compositing for Large-Area Dense Time Series Applications and Science. **Canadian Journal of Remote Sensing**, v. 40, n. 3, p. 192–212, 4 maio 2014.

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

WRIGHT, M. N.; ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 77, n. 1, 2017.

WU, L. et al. Multi-Type Forest Change Detection Using BFAST and Monthly Landsat Time Series for Monitoring Spatiotemporal Dynamics of Forests in Subtropical Wetland. **Remote Sensing**, v. 12, n. 2, p. 341, 20 jan. 2020.

XIAN, S.; CHEN, K.; CHENG, Y. Improved seagull optimization algorithm of partition and XGBoost of prediction for fuzzy time series forecasting of COVID-19 daily confirmed. **Advances in Engineering Software**, v. 173, p. 103212, nov. 2022.

XU, S.; CHAN, H. K.; ZHANG, T. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. **Transportation Research Part E: Logistics and Transportation Review**, v. 122, p. 169–180, fev. 2019.

YU, H. et al. Ultra-Short-Term Operating Reserve Requirement Assessment of Power System Based on Improved XGboost Quantile Regression. 2023 2nd Asia Power and Electrical Technology Conference (APET). **Anais...IEEE**, 28 dez. 2023.

ZAMANI JOHARESTANI, M. et al. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. **Atmosphere**, v. 10, n. 7, p. 373, 4 jul. 2019.

ZHANG, C.; HU, D.; YANG, T. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. **Reliability Engineering & System Safety**, v. 222, p. 108445, jun. 2022.

ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, v. 50, p. 159–175, jan. 2003.

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

**ARTIGO 2 – UMA ABORDAGEM BASEADA EM MACHINE LEARNING E DADOS
ORBITAIS PARA MAPEAMENTO DE MATOCOMPETIÇÃO EM PLANTIOS
CLONAIIS JOVENS DE EUCALIPTO**

AN APPROACH BASED ON MACHINE LEARNING AND ORBITAL DATA FOR
MAPPING WEED COMPETITION IN YOUNG CLONAL PLANTS OF EUCALYPTUS

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

RESUMO

O monitoramento periódico dos plantios florestais jovens é uma ferramenta crucial para o direcionamento eficaz de atividades de combate e controle da matocompetição. Este estudo objetivou através de uma abordagem que combina modelo de *machine learning* a dados orbitais permitir a identificação e classificação da ocorrência de matocompetição em plantios clonais de eucalipto até 9 meses de idade. Dados de imagens Sentinel-2 em múltiplas datas foram utilizados, assim como banco de dados que contém a informação de ocorrência de matocompetição aferido em campo. Devido a vasta disponibilidade de métricas orbitais passíveis de extração das imagens Sentinel-2, uma meta-heurística foi utilizada para a seleção das variáveis independentes que melhor permitissem a classificação da ocorrência da matocompetição. Com o subconjunto de variáveis independentes selecionadas, um modelo *Random Forest* (RF) foi treinado. Os resultados encontrados indicaram que 4 métricas extraídas do Sentinel-2 foram suficientes para predição com desejável precisão da ocorrência da matocompetição. Além disso, o modelo RF teve acurácia superior a 92% na classificação da ocorrência da matocompetição, indicando sua viabilidade para uso nos plantios clonais jovens de eucalipto. Ao final, a abordagem proposta no estudo permite a geração de mapas de classificação das áreas com ocorrência de matocompetição. Esses mapas auxiliam na tomada de decisão por parte dos gestores florestais, determinando o melhor momento e as florestas que precisam de controle de matocompetição.

Palavras-chave: Seleção de variáveis; Algoritmo Genético; *Random Forest*.

ABSTRACT

Periodic monitoring of young forest plantations is a crucial tool for effectively directing activities to combat and control weed competition. This study aimed to use an approach that combines a machine learning model with orbital data to identify and classify the occurrence of weed competition in clonal eucalyptus plantations up to 9 months old. Data from Sentinel-2 images on multiple dates were used, as well as a database that contains information on the occurrence of weed competition measured in the field. Due to the vast availability of orbital metrics that can be extracted from Sentinel-2 images, a meta-heuristic was used to select the independent variables that best allowed the classification of the occurrence of weed competition. With the subset of independent variables selected, a Random Forest (RF) model was trained. The results indicated that four metrics extracted from Sentinel-2 were sufficient to predict the occurrence of weed competition with desirable accuracy. In addition, the RF model was more than 92% accurate in classifying the occurrence of weed competition, indicating its viability for use in young clonal eucalyptus plantations. In the end, the approach proposed in the study allows the generation of classification maps of areas where weed competition occurs. These maps help forest managers decide the best time and the forests that need weed control.

Keywords: Variable selection; Genetic Algorithm; RandomForest.

1 INTRODUÇÃO

Um dos principais desafios enfrentados pelo setor florestal é referente ao controle eficaz da matocompetição que compete com as árvores por nutrientes, água, luz solar e espaço de cultivo (BERGE; AASTVEIT; FYKSE, 2008; IQBAL et al., 2019). Mitigar esse problema de forma rápida é fundamental para garantir um crescimento adequado das árvores de modo a não reduzir o potencial produtivo esperado da floresta. A ocorrência da matocompetição, em especial, no 1º ano de vida de uma floresta de eucalipto pode comprometer seu potencial produtivo em até 15% (STAPE; BINKLEY; RYAN, 2004). Neste contexto, a identificação precisa e instantânea da ocorrência de matocompetição em qualquer nível nos plantios florestais é crucial. Sendo assim, o uso de imagens de satélites torna-se chave ao processo, uma vez que estas imagens auxiliam no monitoramento e reconhecimento de padrões em plantios florestais. Logo, a diferenciação entre solo exposto, matocompetição e plantios florestais através de dados espectrais é uma abordagem que pode auxiliar no manejo específico de silvicultura de precisão (GEBBERS; ADAMCHUK, 2010; LÓPEZ-GRANADOS et al., 2016; SLAUGHTER; GILES; DOWNEY, 2008; SOUZA et al., 2020).

Para a devida classificação e identificação dos locais com ocorrência de matocompetição, modelos flexíveis como os de *machine learning* tornam-se críticos. O *Random Forest* (RF), por exemplo, é amplamente reconhecido como eficaz para aplicação em problemas de classificação de imagens e regressão (BALHA; SINGH, 2022; BORUP et al., 2023; JUNAID et al., 2023; SUN; YUAN; LI, 2023; WAGER; ATHEY, 2018). O RF é aplicado com sucesso no mundo florestal para estimativa do estoque de volume, biomassa, altura do dossel e classificação de imagens (MA et al., 2023; NANDY; SRINET; PADALIA, 2021; TORRE-TOJAL et al., 2022a, 2022b). Um ponto de atenção quanto a aplicação do RF, no entanto, é que sua modelagem não exige a seleção prévia das variáveis independentes críticas ao processo, ou seja, há o risco da interpretabilidade do modelo se tornar complexa, além do risco da geração de estimativas inconsistentes por presença de multicolinearidade entre múltiplas variáveis (SABATIA; BURKHART, 2013). A extração de muitos índices de vegetação a partir de imagens de satélite, inclusive, dificulta a modelagem e sua interpretação, especialmente devido à alta correlação entre as variáveis.

Diversas abordagens de seleção de variáveis são encontradas na literatura, com destaque para meta-heurísticas baseadas em algoritmos evolutivos, como Algoritmo Genético (AG), Otimização por Enxame de Partículas (PSO), Otimização de Colônia de Formigas (ACO) e Algoritmo de Otimização de Lobos Cinzentos (GWO) (ALZUBI et al., 2020; HASSAN et al.,

2022; LIFANDALI; ABGHOOR; CHIBA, 2023; MIRANDA et al., 2022; PRAJAPATI; DAS; GOURISARIA, 2023; ROSTAMI; BERAHMAND; FOROUZANDEH, 2021). Nesse contexto, o algoritmo genético se destaca por sua capacidade de selecionar variáveis em conjuntos de dados de sensoriamento remoto e climáticos, permitindo estimar a altura de florestas nativas (MIRANDA et al., 2022), parametrização de modelos não lineares (MONTI et al., 2022), e otimizar o roteamento de veículos no combate a incêndios florestais (LI et al., 2024). O estágio inicial dos plantios é determinante para o sucesso e a produtividade futura das florestas, e a presença de mato pode influenciar negativamente nesse processo (STAPE; BINKLEY; RYAN, 2004). Com base nesses avanços, a combinação de sensoriamento remoto com *machine learning* é uma abordagem viável para estudar a competição de matocompetição em plantios florestais (MACIEL et al., 2022a; QUAN et al., 2023).

Há uma lacuna em abordagens que permitam uma precisa classificação da ocorrência de matocompetição em plantios jovens de eucalipto. A combinação de dados obtidos por imagens de satélite, assim como pela seleção das variáveis críticas por algoritmo genético e a consequente modelagem através de *Random Forest* soam como promissores para permitir uma evolução natural no controle do processo de controle da matocompetição. Assim, o estudo objetivou através de uma abordagem que combina modelo de *machine learning* a dados orbitais permitir a identificação e classificação da ocorrência de matocompetição em plantios clonais de eucalipto até 9 meses de idade. Em específico, esse estudo buscou aplicar Algoritmo Genético para reduzir a dimensionalidade das variáveis preditoras no modelo *Random Forest*, além de automatizar a seleção de um conjunto ideal de variáveis para a detecção da ocorrência de matocompetição.

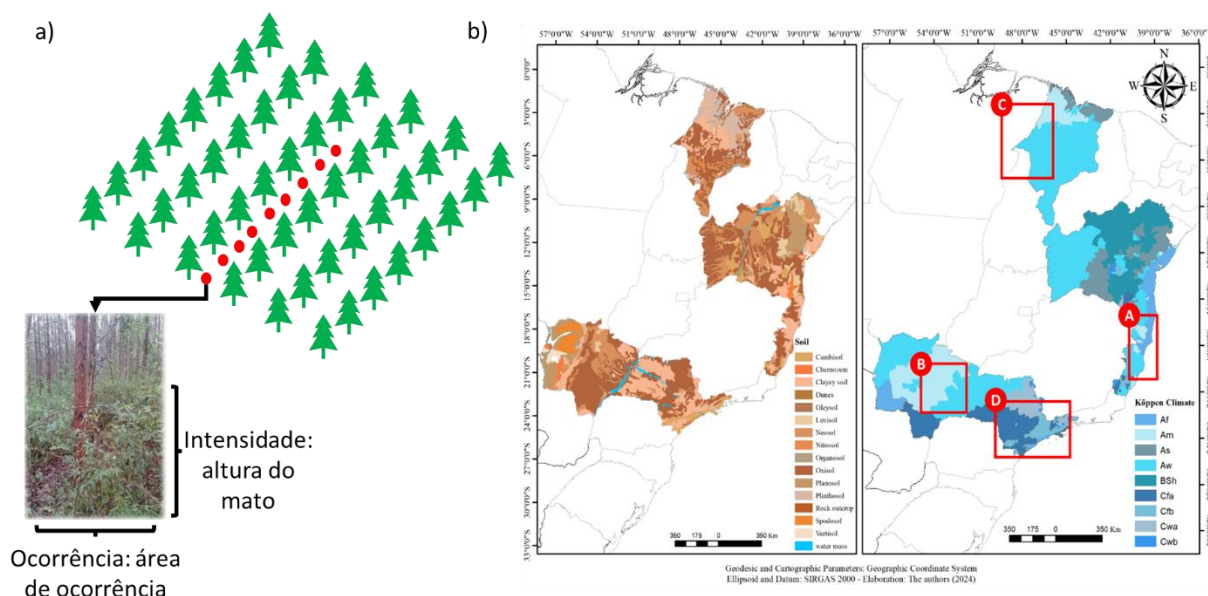
2 MATERIAL E MÉTODOS

2.1 Caracterização da área de estudo

A área de estudo abrange um extenso gradiente longitudinal e latitudinal, e engloba diversas regiões do Brasil, como ilustrado na Figura 1. O clima predominante é tropical, mas varia de acordo com a região, apresentando diferentes padrões de temperatura, clima e regime pluviométrico, como demonstrado no mapa de classificação de Köppen (ALVARES et al., 2013). A temperatura média dessas regiões varia entre 25 e 29 °C, com regime de pluviométricos (chuva) anuais, variando entre 1000 mm e 2200 mm. A área de estudo está dividida em quatro regiões, denominadas A, B, C e D (Figura 1). Foram avaliados mais de 1000

talhões para a amostra de treinamento, distribuídos ao longo dessas quatro áreas, com uma área total aproximada de 53.000 hectares. Os critérios de seleção das áreas de plantio incluíram o trimestre de plantio, a área de estudo e a classe de intensidade da matocompetição.

Figura 10 - (a) Exemplo de caminhamento dos transectos de campo em um povoamento, e avaliação da matocompetição. (b) Localização dos quatro locais de estudo.



Fonte: Do autor (2024).

2.2 Caracterização da base de dados das plantações de *Eucalyptus*

aplicação aérea e mecanizada de herbicidas ou defensivos químicos. Além disso, drones também são utilizados para a aplicação aérea desses produtos.

Entre 2019 e 2023, campanhas de campo foram realizadas para identificar a ocorrência de matocompetição em plantios de eucalipto com até 9 meses de idade. Nas diferentes florestas, foram realizados caminhamentos em transectos, onde se amostrou a ocorrência (cobertura do solo) e a intensidade (altura) da matocompetição. Com base nas coletas, calculou-se para cada transecto o grau de matocompetição (MC), conforme descrito por Stape et al. (2023) (Equação 1). Valores superiores a 50% indicam alta ocorrência (intensidade) de matocompetição. Na Figura 1, é apresentado um esquema ilustrativo de como é realizada a amostragem da matocompetição. O grau de matocompetição é determinado utilizando-se os índices "Ocorrência" e "Intensidade", cada um variando de 0 a 10. Esses índices são multiplicados para obter o Grau de Matocompetição, que varia de 0 a 100. Foram utilizados 1079 talhões espalhados nas 4 áreas de estudo, ou seja, foram avaliados 1079 transectos e calculado um índice de grau de matocompetição para cada um desses transectos. Esse índice foi criado para medir a competição de ervas daninhas nos plantios florestais de eucalipto. Para calcular o grau de mato competição é utilizado a equação a seguir:

$$MC = \frac{10}{3}(O_i - 1)\frac{10}{9}(I_i) \quad (1)$$

Em que: *MC*: grau de matocompetição; *O_i*: ocorrência de matocompetição da amostra *i*. *I_i*: intensidade de matocompetição da amostra *i*.

2.3 Caracterização dos dados orbitais

As imagens do satélite Sentinel-2 foram adotadas pois garantem uma recorrência de cinco dias, permitindo obter informações orbitais próximas às datas das coletas de dados do inventário de matocompetição. Esse satélite possui 13 bandas com resoluções espaciais de 10 m, 20 m e 60 m, onde foram utilizadas cenas do Sentinel-2 nível 1C durante o período de estudo, com a aplicação do Sen2Cor para correção atmosférica. As cenas foram pré-processadas para os blocos da área de interesse, gerando máscaras de sombra e nuvem através da combinação do FMask (QIU; ZHU; HE, 2019) e do s2cloudless (ANZE ZUPANC, 2017; SKAKUN et al., 2022). Para obter essas informações e realizar o processamento necessário, foi utilizado o Google Earth Engine, uma plataforma poderosa que permite o acesso e a análise de grandes volumes de dados geoespaciais de forma eficiente e integrada.

As coordenadas de cada transecto dos plantios foram associadas às bandas dos pixels correspondentes, sendo possível a construção de um banco de dados. Logo, a classificação de ocorrência de matocompetição utilizou um total de 54 índices orbitais (Anexo A). Os índices selecionados são essenciais para diversas aplicações, incluindo monitoramento da saúde da vegetação, detecção de estresse hídrico, identificação de áreas queimadas, avaliação da cobertura do solo, entre outras análises ambientais e agrícolas (ABDI, 2020; DINDAROGLU et al., 2021; FAN et al., 2022; JUNTILA et al., 2023; MA et al., 2023; MIAO et al., 2024b; NANDY; SRINET; PADALIA, 2021).

2.4 Modelagem para detecção da ocorrência de matocompetição

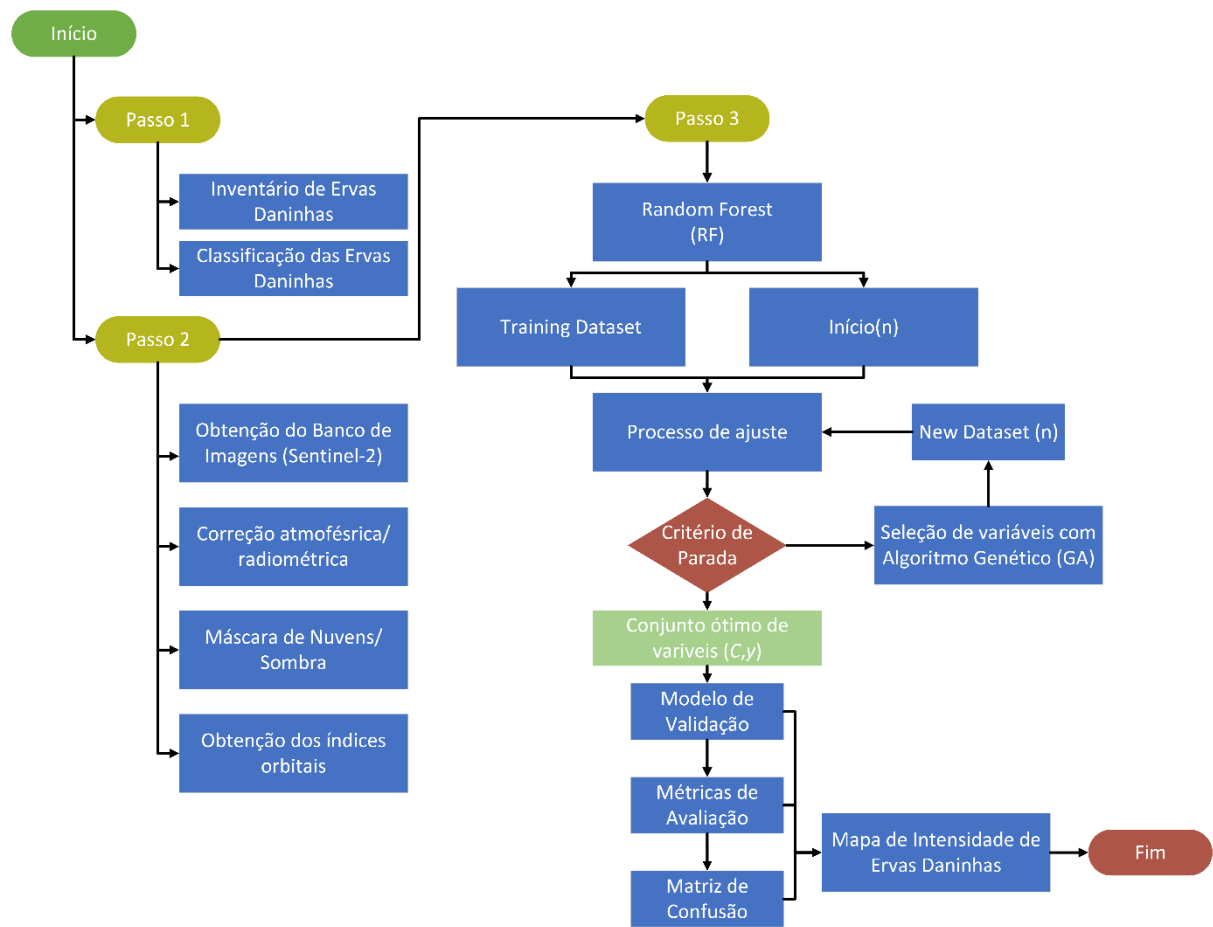
A detecção de objetos ou reconhecimento de padrões é baseado na classificação de imagens. Esses problemas são amplamente encontrados na literatura. Para classificar a matocompetição, é essencial investigar, dentro do vasto espaço de variáveis obtidas por sensoriamento remoto, quais são as mais eficazes para detectar a competição do mato com as linhas de plantios. A seleção dessas variáveis é crucial para compor um conjunto ótimo que explique a variável de interesse, utilizando um modelo robusto para essa avaliação. Entre os diversos algoritmos e procedimentos disponíveis para a classificação, adotou-se o modelo de *machine learning* conhecido como *Random Forest* (RF). O RF é amplamente reconhecido por sua eficácia em análises espectrais (MA et al., 2023; MIAO et al., 2024b; MUTANGA; ADAM; CHO, 2012; NANDY; SRINET; PADALIA, 2021; WERE et al., 2015), destacando-se como uma abordagem robusta e precisa para tarefas de classificação em vastos conjuntos de dados. Desenvolvido por Breiman (2001), o RF baseia-se no princípio de *bootstrapping* para realizar a classificação, construindo um conjunto de árvores de decisão a partir de amostras de *bootstrap*.

O RF é também um algoritmo muito utilizado para a seleção de variáveis, não apenas para a classificação ou predição, mas que pode gerar um longo processo manual de modelagem. Por outro lado, a utilização de múltiplas variáveis independentes num modelo tende a ser redundante, além de introduzir multicolinearidade. Portanto, é importante selecionar e explorar de forma ampla e otimizada as variáveis mais relevantes, não apenas para otimizar o desempenho da aplicação, mas também para facilitar uma compreensão mais profunda dos resultados com base nas variáveis escolhidas. Nesse contexto, utilizou-se o algoritmo genético (AG) para a seleção autônoma de um conjunto ideal de variáveis independentes fornecidas ao modelo de classificação RF. A partir de agora, adotou-se o algoritmo GARF como sendo a

combinação do AG (seleção de variáveis) e o RF (Classificação) para a resolução do problema de classificação, que considerou quatros classes espaciais de intensidade de matocompetição (alta ($MC > 50$), média ($25 < MC < 50$), baixa ($10 < MC < 25$) e aproximadamente nula ($MC < 10$)). Um fluxograma dos passos utilizados no decorrer da metodologia é apresentado na Figura 2.

A eficácia do AG depende significativamente de sua parametrização (HALIM et al., 2021; MOSAYEBI; SODHI, 2020; OKSUZ et al., 2023), e por isso, testes preliminares foram realizados visando ajustar adequadamente o algoritmo. Os operados do AG utilizados foram: seleção por torneio, cruzamento realizado por um ponto de corte e mutação aleatória do bit aleatório, e como de critério de parada um total de 50 gerações (iterações). A diversidade genética foi mantida em uma população de 100 indivíduos, utilizando-se uma probabilidade de 10% para mutações e 50% para a troca aleatória de genes. O melhor conjunto de dados obtido pelo AG foi utilizando como modelo de avaliação. Todos os algoritmos foram implementados em R 4.4.1 (R CORE TEAM, 2024). Nos cálculos do modelo RF foi utilizado o pacote randomForest (LIAW; WIENER, 2002). Na figura 2 ilustra o fluxograma da metodologia, enquanto os detalhes de cada etapa são apresentados nos parágrafos seguintes. Como apresentado acima, um aspecto muito significativo na implementação do AG para seleção se variáveis é o conjunto de variáveis selecionadas (n).

Figura 11 - Fluxograma da metodologia aplicada para a predição da ocorrência espacial de matocompetição em plantios comerciais.



Fonte: Do autor (2024).

A avaliação do desempenho do GARF focou em dois critérios principais: a minimização do erro médio quadrático (*out-of-bag* - OOB) na estimativa da variável de interesse e a redução do conjunto de variáveis preditoras. A função multiobjetiva de adequação (*fitness*) combina a razão entre o erro OOB das variáveis selecionadas pelo AG e o erro OOB máximo (determinado através dos testes preliminares) com a proporção entre o número de variáveis selecionadas pelo AG (n) e o total de variáveis testadas (Equação 2). Esse enfoque destaca a busca por um equilíbrio entre precisão da classificação e a simplicidade do modelo. Em que: erroOOB: erro médio quadrático (*out-of-bag* - OOB); n : número de variáveis selecionadas; NTV: número total de variáveis.

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

2.5 Medidas de Desempenho

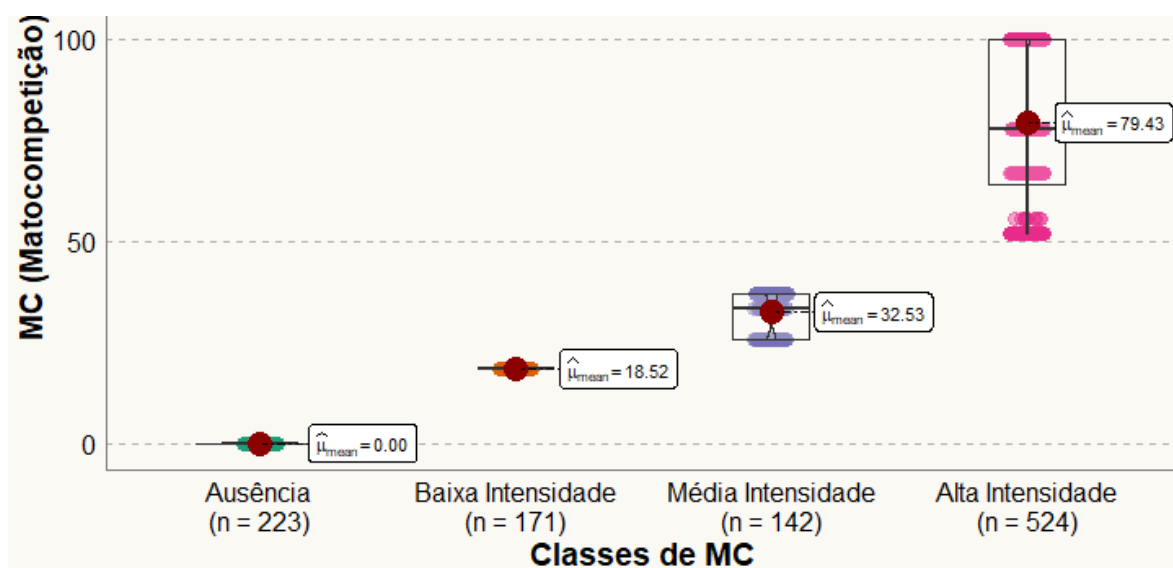
Utilizou-se uma como critério de avaliação, medidas que avaliassem o desempenho da classificação, tais como: a) sensibilidade: mede a proporção de positivos corretamente identificados pelo modelo; b) especificidade: avalia a proporção de negativos corretamente identificados. c) valor preditivo positivo: indica a proporção de positivos preditos que são corretos; d) valor preditivo negativo: reflete a proporção de negativos preditos que são corretos; e) precisão: proporção total de previsões corretas; f) acurácia balanceada: média da sensibilidade e especificidade, oferecendo uma medida de desempenho equilibrada; g) medida F1: média harmônica da precisão e sensibilidade, fornecendo uma única pontuação que considera tanto a precisão quanto a sensibilidade do modelo; e h) coeficiente de recall: equivalente à sensibilidade. Essas métricas ajudam a avaliar a eficácia e o equilíbrio do modelo de classificação. Por fim, um retorno ao campo foi realizado sendo sorteados 400 pontos amostrais ao longo de todo o gradiente de estudo, focando em plantios com idade de até 9 meses. Essa nova amostragem de campo teve como objetivo testar a habilidade do modelo desenvolvido em identificar corretamente a ocorrência de matocompetição. Para validação dessas amostras, utilizou-se uma matriz de confusão como critério auxiliar de avaliação.

3 RESULTADOS

3.1 Comportamento dos dados

Na Figura 3, apresenta-se um gráfico de violino, utilizado para realizar uma análise exploratória dos dados de MC para as amostras de campo. Este gráfico é eficaz para visualizar a distribuição dos valores de MC em diferentes classes de intensidade. O gráfico de violino combina características dos gráficos de caixa e dos histogramas, permitindo uma visualização clara tanto da densidade quanto da dispersão dos dados em cada classe de intensidade, proporcionando uma compreensão mais profunda da distribuição dos dados utilizados neste estudo.

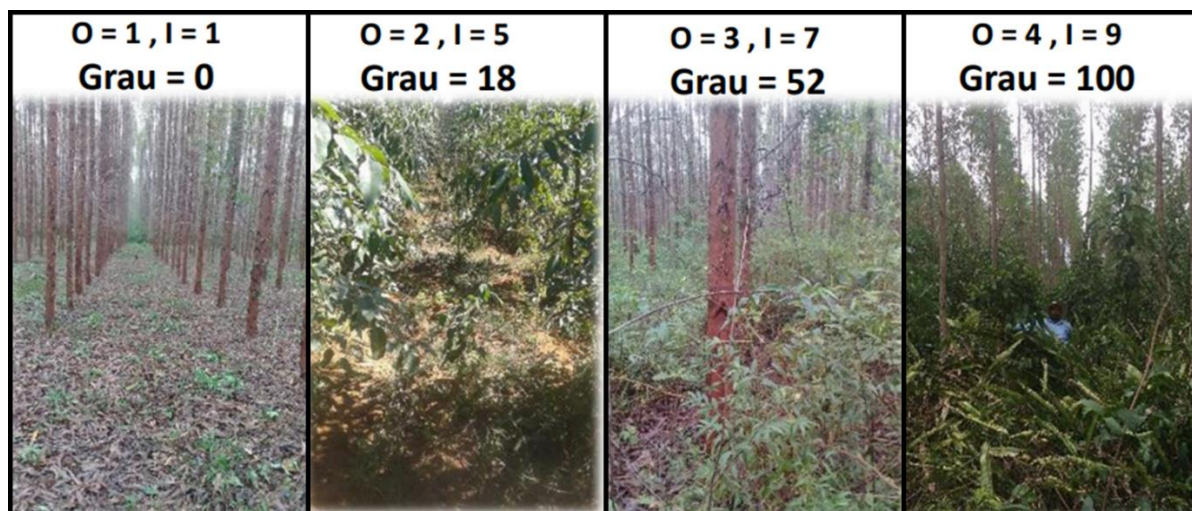
Figura 12 - Gráfico de violino para análise exploratória dos dados.



Fonte: Do autor (2024).

A amostra de dados utilizada para o treinamento do modelo de classificação da intensidade do grau de matocompetição revelou que 48% das amostras foram classificadas como de alta intensidade. A Figura 4 auxilia na compreensão dos níveis de matocompetição encontrados em campo e sua associação com os talhões mensurados. Ela demonstra, para diferentes níveis de intensidade e ocorrência, qual é o grau de MC. Observa-se que florestas classificadas com alta intensidade de matocompetição apresentam um alto grau de infestação, com mato com altura acima de 1 metro. Para matocompetição de intensidade média, há um nível significativo de ocorrência, com mato de até 1 metro de altura. Florestas classificadas com baixa intensidade mostram uma baixa ocorrência de mato, com vegetação de altura abaixo de 40 cm, praticamente rente ao solo. Talhões classificados como ausência de matocompetição não apresentam competição, tanto em termos de intensidade quanto de ocorrência.

Figura 13 - Panorama do comportamento do grau de matocompetição que ocorre nos plantios florestais de eucalipto.

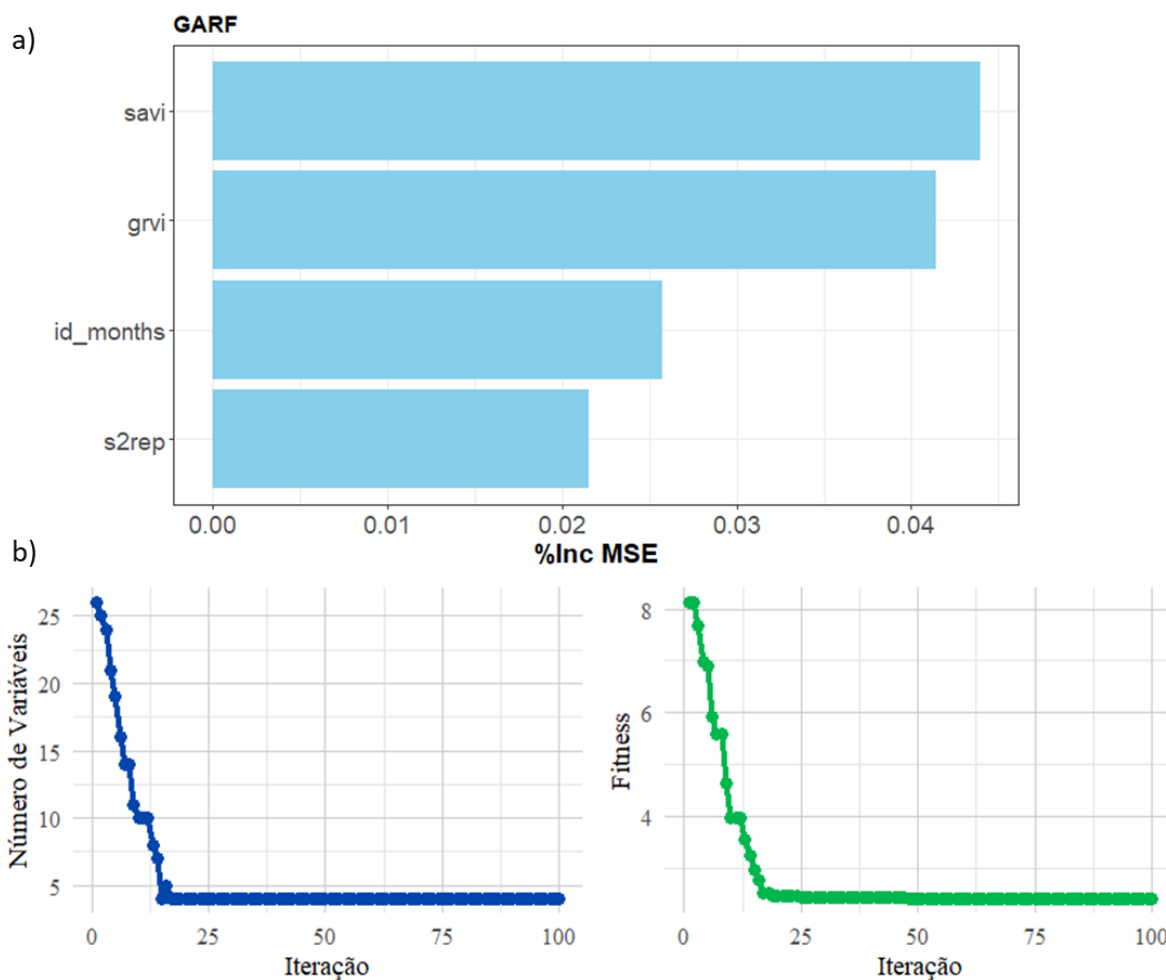


Fonte: Do autor (2024).

3.2 Avaliação do modelo GARF e de precisão do modelo desenvolvido

Um total de 54 índices orbitais e uma variável referente a idade da floresta foram utilizados. A aplicação de AG mostrou que através de um subconjunto de 4 variáveis é suficiente para que o modelo RF seja ajustado ao menor erro possível (Figura 5A). As quatro variáveis utilizadas no GARF, ou seja, selecionadas por AG e utilizadas na modelagem RF foram: idade dos plantios em meses, índice de vegetação verde-vermelho (GRVI), índice de vegetação de solo ajustado (SAVI) e a posição da borda vermelha (S2REP). O processo iterativo de aprimoramento do AG contribuiu para a redução significativa de variáveis, o que ajuda a diminuir a complexidade do modelo e mitiga o risco de *overfitting* (Figura 5B). A seleção otimizada de subconjunto de variáveis diminuiu a redundância das variáveis de entrada e aumentou o poder explicativo das variáveis selecionadas, melhorando a interpretação do modelo.

Figura 14 - Valores médios de importância das variáveis explicativas do modelo para classificar a ocorrência de matocompetição nas florestas jovens de eucalipto (a). Gráficos de redução do número de variáveis e valor da função de aptidão do GARF em função do número de gerações (b).



Fonte: Do autor (2024).

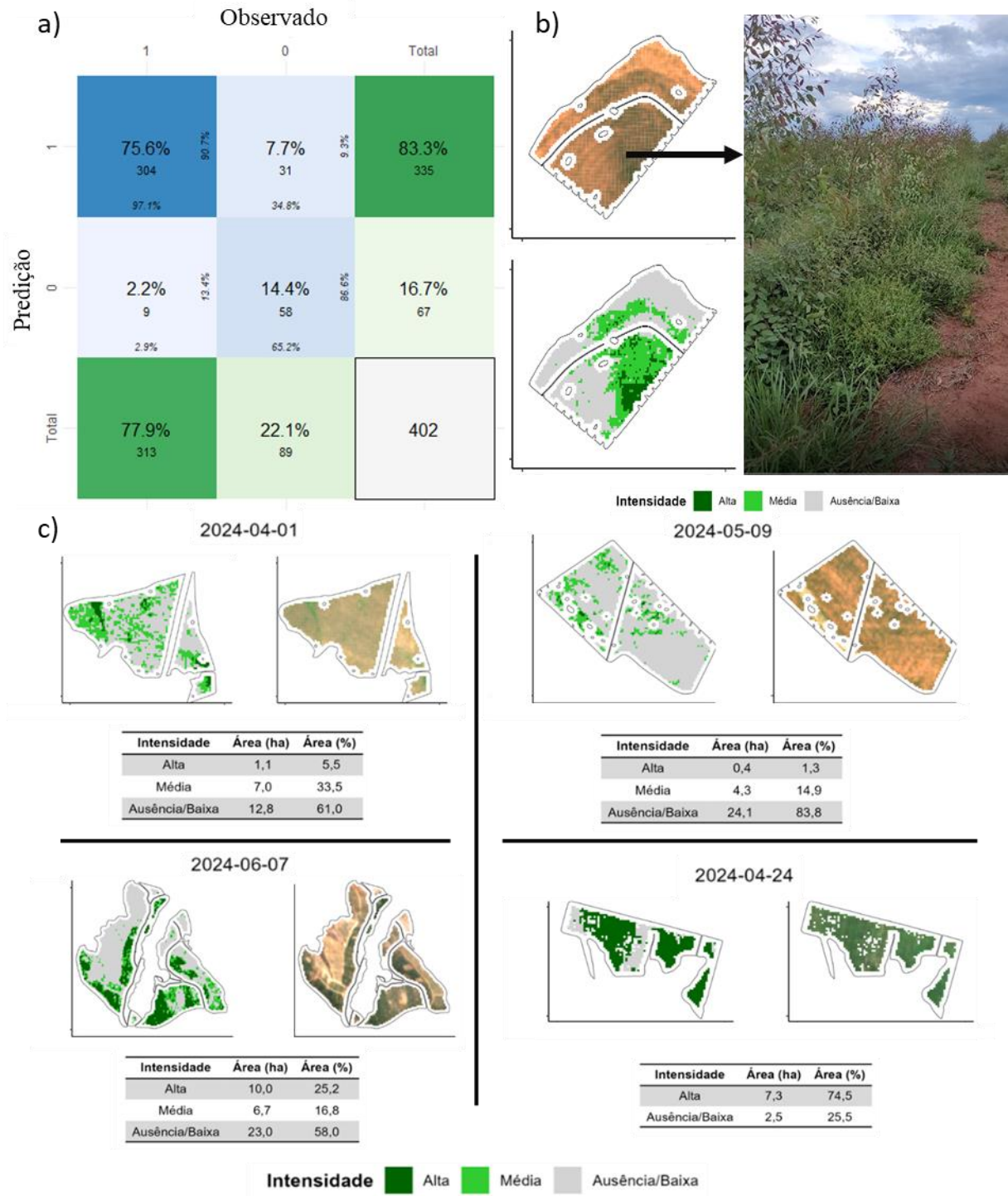
A Tabela 1 indica o quão satisfatório foi o modelo de classificação de ocorrência de matocompetição. O modelo apresentou alta sensibilidade (recall) para todas as classes de MC. Os F1 scores foram de 0,97 para MC referente a baixa/ausência de matocompetição, 0,83 para MC referente a alta intensidade de matocompetição, e 0,71 para MC referente a intensidade média de matocompetição. A precisão e a acurácia foram superiores a 70% em todas as classes de MC.

Tabela 2 - Resultados de predição de duas classes utilizadas para a avaliação de detecção de ervas daninhas em plantios florestais juvenis.

Estatísticas	Intensidade de matocompetição (MC)			
	nula	baixa	média	alta
Sensibilidade	0,98	0,97	0,59	1
Especificidade	0,99	1	0,98	0,9
Falso positivo	0,99	1	0,88	0,71
Falso negativo	0,96	0,95	0,91	1
Precisão	0,99	1	0,88	0,71
Recall	0,98	0,97	0,59	1
F1	0,98	0,98	0,71	0,83
Acurácia	0,98	0,98	0,79	0,95

A Figura 6(a) apresenta a matriz de confusão para a base de dados de validação, onde a classificação do modelo ajustado foi contrastada com a validação de campo. Como pode ser observado, o modelo apresenta uma taxa de sucesso de 90%, acertando principalmente os pontos de ocorrência de matocompetição. Já na Figura 6 (b) e Figura (c), é apresentada a classificação de um talhão validado em campo. A observação de matocompetição competindo com as linhas de plantio pode ser visualizada tanto na imagem de satélite quanto na foto tirada em campo. Ela apresenta ainda um mapa de classificação para alguns dos talhões avaliados no processo de validação. Observa-se que a ocorrência de matocompetição possui uma assinatura característica que a diferencia do solo, permitindo uma clara distinção entre ambos. O modelo detectou pixels com a presença de mato com base nessa assinatura espectral característica.

Figura 15 – a) Matriz de confusão com os dados de validação; b) Exemplo de saída da classificação do mato, comparado com a imagem de satélite e a foto de campo; c) Mapas com a classificação de MC comparadas com imagens em cores verdadeiras, separando pelo nível de intensidade.



Fonte: Do autor (2024).

4 DISCUSSÃO

Este estudo apresentou uma abordagem que combina modelo de *machine learning* a dados orbitais permitir a identificação e classificação da ocorrência de matocompetição em plantios clonais de eucalipto até 9 meses de idade. A detecção de ocorrência de matocompetição em plantios florestais de eucalipto permite a criação de mapas que servem de apoio direcionado à tomada de decisão por meio dos gestores florestais. A presença de matocompetição nos estágios iniciais dos plantios florestais de eucalipto é um componente chave na competição por recursos ambientais, o que impacta diretamente na perda de produtividade potencial (STAPE; BINKLEY; RYAN, 2004).

Embora a produtividade máxima de florestas em povoamentos plantados de rápido crescimento dependa significativamente da nutrição florestal, a presença de matocompetição pode anular os efeitos benéficos da fertilização, devido à competição por recursos limitados, como água, nutrientes e luz (DENG et al., 2020; GRAAT et al., 2018; MACIEL et al., 2022b; MOMOLLI et al., 2023). A matocompetição retarda o crescimento inicial do eucalipto (COLMANETTI et al., 2017; GRAAT et al., 2018) e seu controle tardio pode prejudicar a produtividade dos plantios na rotação (AVISAR et al., 2023).

A rápida detecção de ocorrência de matocompetição se torna uma ferramenta crucial. Nesse sentido, o modelo de classificação construído neste estudo se apresenta como uma ferramenta promissora. Os resultados obtidos pela classificação da matocompetição com imagens de Sentinel-2 oferecem uma versatilidade para criação de mapas das áreas de ocorrência. A recorrência de imagens disponíveis e escala espacial, garantem a confiabilidade necessária para construção de mapas de localização, número e cobertura da matocompetição.

O algoritmo genético demonstrou-se altamente eficiente na seleção automatizada de um conjunto ótimo que explicada a variável resposta. Trabalhos com seleção de variáveis, demonstram a capacidade do GA em encontrar soluções ótimas ou sub-ótimas (ALIČKOVIĆ; SUBASI, 2017; DAS; DAS; GHOSH, 2017; HONG et al., 2018; LACERDA et al., 2022; MIRANDA et al., 2022). Os resultados do modelo RF otimizado, classificaram com precisão da ocorrência de matocompetição em plantios jovens. Trabalhos na literatura confirmam a capacidade do modelo RF em classificar corretamente a ocorrência de matocompetição (DE CASTRO et al., 2018; ISLAM et al., 2021; QUAN et al., 2023).

A metodologia proposta selecionou de forma automatizada e eficiente um subconjunto ótimo de variáveis para a classificação de matocompetição com linhas de plantios de *Eucalyptus juvenis*, reduzindo a redundância e eliminando recursos ruidosos que frequentemente levam a

classificações incorretas. Kiala Odindi e Mutanga (2022) encontraram resultados promissores para mapeamento de erva daninha *Parthenium* usando dados de imagem Sentinel-2 multidas usando *Tree-based Pipeline Optimization Tool* (TPOT). Esta ferramenta reduziu a dimensionalidade do conjunto de dados, diminuindo os custos computacionais e resultando em ligeiro aumento nas precisões de classificação. Casady, Hanley e Seelan (2005) também tiveram sucesso ao classificar spurge folhoso (*Euphorbia esula* L.) usando o classificador de máxima verossimilhança.

O modelo com a maior capacidade de classificação utilizou apenas quatro variáveis. Uma variável importante na classificação de matocompetição, a idade do plantio é um fator importante para verificar se há a competitividade com os plantios florestais. A competitividade tende a diminuir a medida que ocorre a floresta envelhece pelo fechamento do dossel. O GRVI foi considerado a variável preditora mais robusta no estudo e serve para distinguir entre pixels cobertos por folhas verdes e outras condições, como solo descoberto e água. A correlação positiva do GRVI indica uma maior quantidade de folhagem verde, que representa a matocompetição. Outra variável selecionada, o SAVI é uma variável comumente utilizada no monitoramento do solo, onde a assinatura espectral do solo difere de outros tipos de cobertura de terra (BINTE MOSTAFIZ; NOGUCHI; AHAMED, 2021). A assinatura espectral do SAVI permite captar as mudanças na vitalidade da vegetação com atenção às condições de fundo do solo (SASHIKKUMAR et al., 2017). Já o S2REP é uma variável sentinel-2 sensível ao nitrogênio contido nas plantas (conteúdo de clorofila) e ao status de crescimento (ALI et al., 2022), sendo esta utilizada para prever altura e o crescimento de biomassa em povoamentos jovens, tendo forte correlação com ocorrência de matocompetição (RAHIMZADEH-BAJGIRAN et al., 2020).

Ainda existem desafios como separar a ocorrência da matocompetição por tipo, por linha e entrelinha, assim como para plantios acima de 9 meses idade. Estudos futuros podem abordar essas questões, bem como explorar o uso de imagens de satélite com resolução espacial maior ou o uso de imagens de VANT (Veículos Aéreos Não Tripulados) para a classificação das plantas invasoras. Essas abordagens podem fornecer dados mais precisos e detalhados, auxiliando na gestão e controle da matocompetição.

5 CONCLUSÃO

Neste artigo, propomos um algoritmo que combina a seleção de variáveis por meio de algoritmo genético e a classificação baseada em *Random Forest* para criar mapas de

classificação de matocompetição em plantios jovens de eucalipto. Essa metodologia permite identificar de forma automatizada o conjunto de variáveis que explicam a competição do mato com os plantios florestais, eliminando erros manuais e subjetivos no processo de seleção. O processo automático torna a metodologia computacionalmente eficiente, confiável e precisa, ao reduzir a multicolinearidade e a dimensionalidade dos dados, fatores que poderiam impactar negativamente na classificação do modelo. O índice de matocompetição avalia a competição entre o mato e os plantios florestais sob várias perspectivas. A presença de matocompetição impacta diretamente a taxa de crescimento das árvores, devido à competição por nutrientes, água e luz. Portanto, é crucial realizar o monitoramento periódico das áreas de plantio para identificar a competição e permitir ações de combate ou mitigação. A abordagem proposta gera mapas de classificação como produto final. Esses mapas permitem que os gestores florestais criem estratégias de manejo de matocompetição de maneira adequada e no *timing* oportuno. Esse rastreamento de informações possibilita um controle direcionado de recursos, tornando o manejo mais eficiente.

REFERÊNCIAS

- ABDI, A. M. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. **GIScience & Remote Sensing**, v. 57, n. 1, p. 1–20, 2 jan. 2020.
- ALI, A. et al. Evaluating Sentinel-2 red edge through hyperspectral profiles for monitoring LAI & chlorophyll content of Kinnow Mandarin orchards. **Remote Sensing Applications: Society and Environment**, v. 26, p. 100719, abr. 2022.
- ALIČKOVIĆ, E.; SUBASI, A. Breast cancer diagnosis using GA feature selection and Rotation Forest. **Neural Computing and Applications**, v. 28, n. 4, p. 753–763, 2017.
- ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.
- ALZUBI, Q. M. et al. Intrusion detection system based on a modified binary grey wolf optimisation. **Neural Computing and Applications**, v. 32, n. 10, p. 6125–6137, 27 maio 2020.
- ANZE ZUPANC. **Improving cloud detection with machine learning**. Disponível em: <<https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13>>. Acesso em: 3 jul. 2024.
- AVISAR, D. et al. Safety of genetically modified glyphosate-tolerant eucalyptus designed for integrated weed management. **Advances in Weed Science**, v. 41, 16 nov. 2023.
- BALHA, A.; SINGH, C. K. Comparison of Maximum Likelihood, Neural Networks, and Random Forests Algorithms in Classifying Urban Landscape. Em: [s.l: s.n.]. p. 29–38.
- BERGE, T. W.; AASTVEIT, A. H.; FYKSE, H. Evaluation of an algorithm for automatic detection of broad-leaved weeds in spring cereals. **Precision Agriculture**, v. 9, n. 6, p. 391–405, 27 dez. 2008.
- BINTE MOSTAFIZ, R.; NOGUCHI, R.; AHAMED, T. Agricultural Land Suitability Assessment Using Satellite Remote Sensing-Derived Soil-Vegetation Indices. **Land**, v. 10, n. 2, p. 223, 23 fev. 2021.
- BORUP, D. et al. Targeting predictors in random forest regression. **International Journal of Forecasting**, v. 39, n. 2, p. 841–868, abr. 2023.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- CASADY, G. M.; HANLEY, R. S.; SEELAN, S. K. Detection of Leafy Spurge (*Euphorbia esula*) Using Multidate High-Resolution Satellite Imagery. **Weed Technology**, v. 19, n. 2, p. 462–467, 20 jun. 2005.

COLMANETTI, M. A. A. et al. Interference of *Urochloa decumbens* and *Panicum maximum* in the initial growth of six clones of *Eucalyptus urograndis*. **Australian Journal of Crop Science**, v. 11, n. 10, p. 1261–1267, 20 out. 2017.

DAS, A. K.; DAS, S.; GHOSH, A. Ensemble feature selection using bi-objective genetic algorithm. **Knowledge-Based Systems**, v. 123, p. 116–127, maio 2017.

DE CASTRO, A. et al. An Automatic Random Forest-OBIA Algorithm for Early Weed Mapping between and within Crop Rows Using UAV Imagery. **Remote Sensing**, v. 10, n. 2, p. 285, 12 fev. 2018.

DENG, Y. et al. Effects of Different Weeding Methods on the Biomass of Vegetation and Soil Evaporation in Eucalyptus Plantations. **Sustainability**, v. 12, n. 9, p. 3669, 1 maio 2020.

DINDAROGLU, T. et al. Evaluation of geomorphometric characteristics and soil properties after a wildfire using Sentinel-2 MSI imagery for future fire-safe forest. **Fire Safety Journal**, v. 122, p. 103318, jun. 2021.

FAN, X. et al. Sentinel-2 Images Based Modeling of Grassland Above-Ground Biomass Using Random Forest Algorithm: A Case Study on the Tibetan Plateau. **Remote Sensing**, v. 14, n. 21, p. 5321, 24 out. 2022.

GEBBERS, R.; ADAMCHUK, V. I. Precision Agriculture and Food Security. **Science**, v. 327, n. 5967, p. 828–831, 12 fev. 2010.

GRAAT, Y. et al. Initial Development of Eucalyptus According to Different Desiccation Periods of Signalgrass. **Planta Daninha**, v. 36, n. 0, 29 mar. 2018.

HALIM, Z. et al. An effective genetic algorithm-based feature selection method for intrusion detection systems. **Computers & Security**, v. 110, p. 102448, nov. 2021.

HASSAN, I. H. et al. An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection. **Intelligent Systems with Applications**, v. 16, p. 200114, nov. 2022.

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

IQBAL, N. et al. Investigation of alternate herbicides for effective weed management in glyphosate-tolerant cotton. **Archives of Agronomy and Soil Science**, v. 65, n. 13, p. 1885–1899, 10 nov. 2019.

ISLAM, N. et al. Machine Learning Based Approach for Weed Detection in Chilli Field Using RGB Images. Em: [s.l.: s.n.]. p. 1097–1105.

JUNAID, M. et al. Mapping LULC Dynamics and Its Potential Implication on Forest Cover in Malam Jabba Region with Landsat Time Series Imagery and Random Forest Classification. **Sustainability**, v. 15, n. 3, p. 1858, 18 jan. 2023.

JUNTTILA, S. et al. Estimating local-scale forest GPP in Northern Europe using Sentinel-2: Model comparisons with LUE, APAR, the plant phenology index, and a light response function. **Science of Remote Sensing**, v. 7, p. 100075, jun. 2023.

KIALA, Z.; ODINDI, J.; MUTANGA, O. Determining the Capability of the Tree-Based Pipeline Optimization Tool (TPOT) in Mapping Parthenium Weed Using Multi-Date Sentinel-2 Image Data. **Remote Sensing**, v. 14, n. 7, p. 1687, 31 mar. 2022.

LACERDA, T. H. S. et al. Feature selection by genetic algorithm in nonlinear taper model. **Canadian Journal of Forest Research**, 20 jan. 2022.

LI, X. et al. A Parallel Genetic Algorithm With Variable Neighborhood Search for the Vehicle Routing Problem in Forest Fire-Fighting. **IEEE Transactions on Intelligent Transportation Systems**, p. 1–17, 2024.

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

LIFANDALI, O.; ABGHOOR, N.; CHIBA, Z. Feature Selection Using a Combination of Ant Colony Optimization and Random Forest Algorithms Applied To Isolation Forest Based Intrusion Detection System. **Procedia Computer Science**, v. 220, p. 796–805, 2023.

LÓPEZ-GRANADOS, F. et al. Early season weed mapping in sunflower using UAV technology: variability of herbicide treatment maps against weed thresholds. **Precision Agriculture**, v. 17, n. 2, p. 183–199, 20 abr. 2016.

MA, T. et al. A Novel Vegetation Index Approach Using Sentinel-2 Data and Random Forest Algorithm for Estimating Forest Stock Volume in the Helan Mountains, Ningxia, China. **Remote Sensing**, v. 15, n. 7, p. 1853, 30 mar. 2023.

MACIEL, J. C. et al. Growth, Nutrient Accumulation, and Nutritional Efficiency of a Clonal Eucalyptus Hybrid in Competition with Grasses. **Forests**, v. 13, n. 8, p. 1157, 22 jul. 2022a.

MACIEL, J. C. et al. Growth, Nutrient Accumulation, and Nutritional Efficiency of a Clonal Eucalyptus Hybrid in Competition with Grasses. **Forests**, v. 13, n. 8, p. 1157, 22 jul. 2022b.

MIAO, J. et al. Modeling strategies and influencing factors in retrieving canopy equivalent water thickness of mangrove forest with Sentinel-2 image. **Ecological Indicators**, v. 158, p. 111497, jan. 2024.

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, jan. 2022.

MOMOLLI, D. R. et al. Weed Coexistence in Eucalyptus Hybrid Stands Decreases Biomass and Nutritional Efficiency Mid-Rotation. **Forests**, v. 14, n. 9, p. 1816, 5 set. 2023.

MONTI, C. A. U. et al. Hybrid Method for Fitting Nonlinear Height–Diameter Functions. **Forests**, v. 13, n. 11, p. 1783, 28 out. 2022.

MOSAYEBI, M.; SODHI, M. **Tuning genetic algorithm parameters using design of experiments**. Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion. **Anais...**New York, NY, USA: ACM, 8 jul. 2020.

MUTANGA, O.; ADAM, E.; CHO, M. A. High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm. **International Journal of Applied Earth Observation and Geoinformation**, v. 18, n. 1, p. 399–406, 2012.

NANDY, S.; SRINET, R.; PADALIA, H. Mapping Forest Height and Aboveground Biomass by Integrating ICESat-2, Sentinel-1 and Sentinel-2 Data Using Random Forest Algorithm in Northwest Himalayan Foothills of India. **Geophysical Research Letters**, v. 48, n. 14, 28 jul. 2021.

OKSUZ, M. K. et al. A genetic algorithm integrated with the initial solution procedure and parameter tuning for capacitated P-median problem. **Neural Computing and Applications**, v. 35, n. 8, p. 6313–6330, 19 mar. 2023.

PRAJAPATI, S.; DAS, H.; GOURISARIA, M. K. **Feature Selection using Ant Colony Optimization for Microarray Data Classification**. 2023 6th International Conference on Information Systems and Computer Networks (ISCON). **Anais...IEEE**, 3 mar. 2023.

QIU, S.; ZHU, Z.; HE, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. **Remote Sensing of Environment**, v. 231, p. 111205, set. 2019.

QUAN, L. et al. Multimodal remote sensing application for weed competition time series analysis in maize farmland ecosystems. **Journal of Environmental Management**, v. 344, p. 118376, out. 2023.

R CORE TEAM. **R: A language and environment for statistical computing**. ViennaR Foundation for Statistical Computing, , 2024.

RAHIMZADEH-BAJGIRAN, P. et al. Forest Potential Productivity Mapping by Linking Remote-Sensing-Derived Metrics to Site Variables. **Remote Sensing**, v. 12, n. 12, p. 2056, 26 jun. 2020.

ROSTAMI, M.; BERAHMAND, K.; FOROUZANDEH, S. A novel community detection based genetic algorithm for feature selection. **Journal of Big Data**, v. 8, n. 1, p. 2, 4 dez. 2021.

SABATIA, C. O.; BURKHART, H. E. Height and Diameter Relationships and Distributions in Loblolly Pine Stands of Enhanced Genetic Material. **Forest Science**, v. 59, n. 3, p. 278–289, 14 jun. 2013.

SASHIKKUMAR, M. C. et al. Remote sensing for recognition and monitoring of vegetation affected by soil properties. **Journal of the Geological Society of India**, v. 90, n. 5, p. 609–615, 28 nov. 2017.

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

SKAKUN, S. et al. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. **Remote Sensing of Environment**, v. 274, p. 112990, jun. 2022.

SLAUGHTER, D. C.; GILES, D. K.; DOWNEY, D. Autonomous robotic weed control systems: A review. **Computers and Electronics in Agriculture**, v. 61, n. 1, p. 63–78, abr. 2008.

SOUZA, M. F. DE et al. Spectral differentiation of sugarcane from weeds. **Biosystems Engineering**, v. 190, p. 41–46, fev. 2020.

STAPE, J. L.; BINKLEY, D.; RYAN, M. G. Eucalyptus production and the supply, use and efficiency of use of water, light and nitrogen across a geographic gradient in Brazil. **Forest Ecology and Management**, v. 193, n. 1–2, p. 17–31, maio 2004.

SUN, P.; YUAN, X.; LI, D. Classification of Individual Tree Species Using UAV LiDAR Based on Transformer. **Forests**, v. 14, n. 3, p. 484, 28 fev. 2023.

TORRE-TOJAL, L. et al. Above-ground biomass estimation from LiDAR data using random forest algorithms. **Journal of Computational Science**, v. 58, p. 101517, fev. 2022a.

TORRE-TOJAL, L. et al. Above-ground biomass estimation from LiDAR data using random forest algorithms. **Journal of Computational Science**, v. 58, p. 101517, fev. 2022b.

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

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

Anexo A

Tabela 3 - Descrição dos índices orbitais de Sentinel-2.

Sigla	Descrição	Fórmula
ARI1	Anthocyanin Reflectance Index	$ARI1 = \frac{1}{B3} - \frac{1}{B5}$
ARI2	Anthocyanin Reflectance Index 2	$ARI2 = \left(\frac{B8}{B2}\right) - \left(\frac{B8}{B3}\right)$
ARVI	Atmospherically Resistant Vegetation Index	$ARVI = \frac{B8 - (2 * B4 - B2)}{B8 + (2 * B4 - B2)}$
BAI	Burn Area Index	$BAI = \frac{1}{((0.1 - B4)^2 + (0.06 - B8)^2)}$
BAI2	Burn Area Index 2	$BAI2 = \left(1 - \sqrt{\frac{B6 * B7 * B8A}{B4}}\right) * \left(\frac{(B12 - B8A)}{\sqrt{B12 + B8A}} + 1\right)$
CHLRE	Chlorophyll Red Edge Index	$CHLRE = \frac{B5}{B8}$
CRI1	Carotenoid Reflectance Index 1	$CRI1 = \left(\frac{1}{B2}\right) - \left(\frac{1}{B3}\right)$
CRI2	Carotenoid Reflectance Index 2	$CRI2 = \left(\frac{1}{B2}\right) - \left(\frac{1}{B5}\right)$
CRSWIR	Continuum Removal Short Wave Infrared	$CRSWIR = \frac{B11}{B8A + (1613.7 - 864.8) * \frac{(B12 - B8A)}{(2202.4 - 864.8)}}$
DSWI	Drought Stress Water Index	$DSWI = \frac{(B7 + B3)}{(B11 + B4)}$
EVI	Enhanced Vegetation Index	$EVI = \frac{2.5 * (B8 - B4)}{(B8 + 6 * B4 - (7.5 * B2 + 1))}$
EVI2	Enhanced Vegetation Index 2	$EVI2 = \frac{2.5 * (B8 - B4)}{(B8 + 2.4 * B4 + 1)}$
GNDVI	Green Normalized Difference Vegetation Index	$GNDVI = \frac{(B8 - B3)}{(B8 + B3)}$
GRVI	Green Red Vegetation Index	$GRVI = \frac{(B3 - B4)}{(B3 + B4)}$
GSAVI	Green Soil Adjusted Vegetation Index	$GSAVI = 1.5 * \left(\frac{B8 - B3}{B8 + B3 + 0.5}\right)$
IRECI	Inverted Red-Edge Chlorophyll Index	$IRECI = (B7 - B4) * \left(\frac{B6}{B5}\right)$
LAISAVI	Leaf Area Index Soil Adjusted Vegetation Index	$LAISAVI = \frac{-\log\left(\frac{0.371 + 1.5 * (B8 - B4)}{(B8 + B4 + 0.5)}\right)}{2.4}$
MCARI	Modified Chlorophyll Absorption in Reflectance Index	$MCARI = \frac{1 - 0.2 * (B5 - B3)}{(B5 - B4)}$
MIRBI	Mid-Infrared Burn Index	$MIRBI = (10 * B12) - (9.8 * B11) + 2$
mNDVI705	Modified NDVI 705	$mNDVI705 = \frac{(B6 - B5)}{(B6 + B5 - 2 * B2)}$
MSI	Moisture Stress Index	$MSI = \frac{B11}{B8A}$

Sigla	Descrição	Fórmula
MSR	Modified Simple Ratio	$MSR = \frac{(B7 - B2)}{(B4 - B2)}$
MTCI	Meris Terrestrial Chlorophyll Index	$MTCI = \frac{(B6 - B5)}{(B5 + B4)}$
NBR	Normalized Burn Ratio	$NBR = \frac{(B12 - B8A)}{(B12 + B8A)}$
NBR2	Normalized Burn Ratio 2	$NBR2 = \frac{(B11 - B12)}{(B11 + B12)}$
NBRRAW	Normalized Burn Ratio Raw	$NBRRAW = \frac{(B8 - B12)}{(B8 + B12)}$
NDI45	Normalized Difference Index 4/5	$NDI45 = \frac{(B5 - B4)}{(B5 + B4)}$
NDII	Normalized Difference Infrared Index	$NDII = \frac{(B8A - B11)}{(B8A + B11)}$
NDVIG	Normalized Difference Vegetation Index Green	$NDVIG = \frac{B3 * (B8 - B4)}{(B8 + B4)}$
NDVI2	Normalized Difference Vegetation Index 2	$NDVI2 = \frac{B8A - B4}{B8A + B4}$
NDVI705	Normalized Difference Vegetation Index 705	$NDVI705 = \frac{(B6 - B5)}{(B6 + B5)}$
NDMI	Normalized Difference Moisture Index	$NDMI = \frac{(B8 - B11)}{(B8 + B11)}$
NDSWIR	Normalized Difference SWIR Index	$NDSWIR = \frac{(B11 - B8A)}{(B11 + B8A)}$
NDWI	Normalized Difference Water Index	$NDWI = \frac{(B3 - B8)}{(B3 + B8)}$
PSSRA	Plant Senescence Reflectance Index	$PSSRA = \frac{B7}{B4}$
PSRI	Plant Senescence Reflectance Index	$PSRI = \frac{(B4 - B2)}{B5}$
PSRINIR	Plant Senescence Reflectance Index NIR	$PSRINIR = \frac{(B4 - B2)}{B8}$
OSAVI	Optimized Soil Adjusted Vegetation Index	$OSAVI = \frac{(1 + 0.16) * (B7 - B4)}{(B7 + B4 + 0.16)}$
RVI	Ratio Vegetation Index	$RVI = \frac{B4}{B8}$
RDVI	Renormalized Difference Vegetation Index	$RDVI = \frac{(B7 - B4)}{(B7 + B4)^{0.5}}$
RENDVI	Red-Edge Normalized Difference Vegetation Index	$RENDVI = \frac{(B8 - B6)}{(B8 + B6)}$
RENDWI	Red-Edge Normalized Difference Water Index	$RENDWI = \frac{B7 - B5}{B7 + B5}$
RGRI	Red Green Ratio Index	$RGRI = \frac{(B4 + B5)}{(B2 + B3)}$
S2REP	Sentinel-2 Red Edge Position	$S2REP = \frac{705 + 35 * (0.5 * (B8 + B5) - B6)}{(B7 - B6)}$
SAVI	Soil Adjusted Vegetation Index	$SAVI = \frac{1.5 * (B8 - B5)}{(B8 + B5 + 0.5)}$
SELI	Sentinel-2 Enhanced Vegetation Index	$SELI = \frac{(B8A - B5)}{(B8A + B5)}$

Sigla	Descrição	Fórmula
SIPI	Structure Insensitive Pigment Index	$\text{SIPI} = \frac{(B8 - B2)}{(B8 - B4)}$
SR	Simple Ratio	$\text{SR} = \frac{B8}{B4}$
TCARI	Transformed Chlorophyll Absorption in Reflectance Index	$\text{TCARI} = 3 * (B5 - B4) - 0.2 * (B5 - B3) * \left(\frac{B5}{B4}\right)$
TVI	Triangular Vegetation Index	$\text{TVI} = \frac{(120 * (B6 - B3)) - (200 * (B4 - B3))}{2}$

**ARTIGO 3 – UMA ABORDAGEM PARA ESTIMATIVA PRECOCE DO
POTENCIAL PRODUTIVO DE FLORESTAS DE EUCALIPTO EM FUNÇÃO DA
QUALIDADE DE PLANTIO, EDAFOCLIMÁTICO E SENSORIAMENTO REMOTO**

AN APPROACH FOR EARLY ESTIMATION OF THE PRODUCTION POTENTIAL OF
EUCALYPTUS FORESTS BASED ON PLANTING QUALITY, SOIL AND CLIMATE
ANALYSIS AND REMOTE SENSING

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

RESUMO

O manejo florestal é essencial para a proteção e uso racional dos recursos florestais existentes. Vários fatores, incluindo o clima, são cruciais para determinar com precisão o potencial de crescimento das florestas. A metodologia proposta sugere que é possível antecipar o inventário florestal realizado em dois anos usando informações do Sentinel-2 e dados climáticos. Com base nessa suposição, foi desenvolvido um método para prever a avaliação da qualidade local de florestas clonais de eucalipto no Brasil usando dados do primeiro ano de vida dessas florestas. Para gerenciar a grande quantidade de dados disponíveis, foi utilizado um método automatizado de seleção de variáveis *wrapper* por meio de um algoritmo genético. O modelo foi ajustado usando uma *Random Forest*. Um conjunto de variáveis climáticas, de sensoriamento remoto e cadastrais foram selecionadas pelo modelo, permitindo uma estimativa precisa da altura dominante da floresta aos dois anos de idade, com um RMSE de 1,5 m. A antecipação do inventário florestal permite inferir os impactos das variações climáticas, possibilitando a criação de cenários extremos de produtividade com base na quantidade e distribuição de precipitação, bem como nos grupos de materiais clonais nas áreas em estudo. Isso fornece uma ferramenta valiosa para a tomada de decisões pelos gestores florestais, auxiliando na implementação de ações corretivas e na otimização do manejo florestal.

Palavras-chave: *Random Forest*; Precipitação; Índice de sítio; Qualidade local; Seleção de variáveis.

ABSTRACT

Forest management is essential for the protection and rational use of existing forest resources. Several factors, including climate, are crucial to accurately determine the growth potential of forests. The proposed methodology suggests that it is possible to anticipate the forest inventory carried out at 2 years using information from Sentinel-2 and climate data. Based on this assumption, a method was developed to anticipate the assessment of the local quality of clonal eucalyptus forests in Brazil, using data from the first year of life of these forests. To deal with the large amount of available data, an automated wrapper variable selection method was used through the genetic algorithm. The model was adjusted using Random Forest. A set of climate, remote sensing and cadastral variables was selected by the model, allowing for an accurate estimate of the dominant height of the forest at 2 years of age, with an RMSE of 1.5 m. The anticipation of forest inventory makes it possible to infer the impacts of climate variations, allowing the creation of extreme productivity scenarios based on the amount and distribution of precipitation, as well as on the groups of clonal materials in the areas under study. This offers a valuable tool for decision-making by forest managers, assisting in the implementation of corrective actions and in the optimization of forest management.

Keywords: Random Forest; Precipitation; Site index; Local quality; Variable selection

1 INTRODUÇÃO

Florestas de eucalipto no Brasil geralmente são povoamentos homogêneos resultado do intenso esforço no melhoramento genético e de tratamentos silviculturais realizadas pela indústria do país. Atualmente, o Brasil possui 9,94 milhões de hectares de área plantada. O eucalipto abrange 76% da área plantada, com 7,6 milhões de hectares (IBÁ, 2024) e apresenta uma produtividade estimada em $32,7 \text{ m}^3 \cdot \text{há}^{-1} \cdot \text{ano}$ (com casca), com duração média do ciclo de 6,7 anos (IBÁ, 2024). Avaliar o potencial de crescimento dessas florestas é essencial para um planejamento eficaz e sustentável. A capacidade produtiva é comumente avaliada por meio do índice de sítio (BURKHART; TOMÉ, 2012; SCOLFORO et al., 2020), uma medida crucial para prever a produtividade florestal e otimizar o manejo das florestas de eucalipto. Ela é amplamente independente da densidade do povoamento e integra os efeitos históricos do clima e do solo na produtividade florestal (ILES, 2003; SCOLFORO et al., 2020).

Diversos esforços têm sido realizados para explorar como a variação climática afeta a produtividade florestal (SABATIA; BURKHART, 2014; SCOLFORO et al., 2020). As oscilações climáticas recentes têm mostrado que o potencial do sítio pode variar significativamente, influenciando diretamente o crescimento e a produtividade das florestas de eucalipto. Além das oscilações climáticas, a qualidade do estabelecimento da floresta também pode impactar drasticamente o índice de sítio. Fatores como a qualidade do solo, o manejo inicial e as condições de plantio podem melhorar ou reduzir a produtividade florestal. A inclusão de variáveis climáticas, como a disponibilidade de água no solo e a precipitação anual, melhora a capacidade explicativa e a previsão do crescimento florestal (SCOLFORO et al., 2016, 2017, 2020).

Diante das incertezas relacionadas à qualidade florestal e ao clima vigente, o uso do sensoriamento remoto torna-se crítico. Com os avanços tecnológicos, a obtenção de imagens de satélite e dados climáticos permitem a fácil aquisição e processamento destes dados. Apesar de ter uma precisão inferior às encontradas com amostras de campo, as imagens ópticas mantêm um papel crucial no monitoramento da biomassa florestal (LE MAIRE et al., 2011). As tecnologias modernas permitem antecipar o entendimento da biomassa florestal de maneira precoce, facilitando a coleta de dados essenciais para a estimativa do índice de sítio. Combinar informações satelitais, climáticas e de qualidade florestal pode agregar valor ao processo de identificação do sítio de uma rotação específica, antecipando o planejamento em 12 meses. Isso torna a informação crítica para um manejo mais eficiente e sustentável. Utilizar dados de satélite

pode reduzir significativamente os custos associados à coleta de dados de campo, permitindo a alocação de recursos para outras áreas críticas do manejo florestal.

A aquisição de informações de satélite permite antecipar dados importantes antes da coleta de campo. A sensibilidade detectada pela reflectância do dossel florestal pode ser correlacionada com a produtividade observada meses depois. A reflectância da floresta indica sua salubridade, e muitos índices extraídos das bandas dos sensores orbitais fornecem informações sobre disponibilidade de água, déficit hídrico, taxa fotossintética e teor de clorofila, entre outros. Essas características são fundamentais para nortear a capacidade produtiva. Variáveis obtidas de imagens de satélites têm demonstrado confiabilidade nas estimativas (HU et al., 2020; MA et al., 2023; MIRANDA et al., 2022; MURA et al., 2018; SANTOS et al., 2023; ZHOU; FENG, 2023). Esses modelos, baseados em imagens de satélite e informações climáticas, podem explorar áreas amplas e fornecer estimativas precisas de sítio, facilitando o planejamento e manejo florestal. Por fim, os dados de satélite históricos permitem análises de longo prazo das tendências florestais, ajudando a identificar padrões e prever futuras mudanças na produtividade florestal

Devido à alta disponibilidade de dados, combinar todas essas informações é um processo complexo e computacionalmente intenso. Para buscar um melhor entendimento e interpretação dos modelos gerados, o uso de técnicas de *machine learning* com seleção de variáveis tem se mostrado promissor (LACERDA et al., 2022; LI et al., 2019; ZHANG et al., 2022). O Algoritmo Genético (GA) é uma dessas técnicas, permitindo analisar grandes volumes de dados de forma eficiente e gerando *insights* valiosos para o manejo florestal. Quando combinado com a robustez do Random Forest (RF), a capacidade explicativa do modelo é aprimorada, eliminando ruídos que poderiam prejudicar a precisão das previsões (LOPES et al., 2022; MIRANDA et al., 2022). Essas abordagens não só aumentam a eficiência do processamento de dados, mas também melhoram a qualidade das estimativas fornecidas, resultando em um manejo florestal mais informado e eficaz.

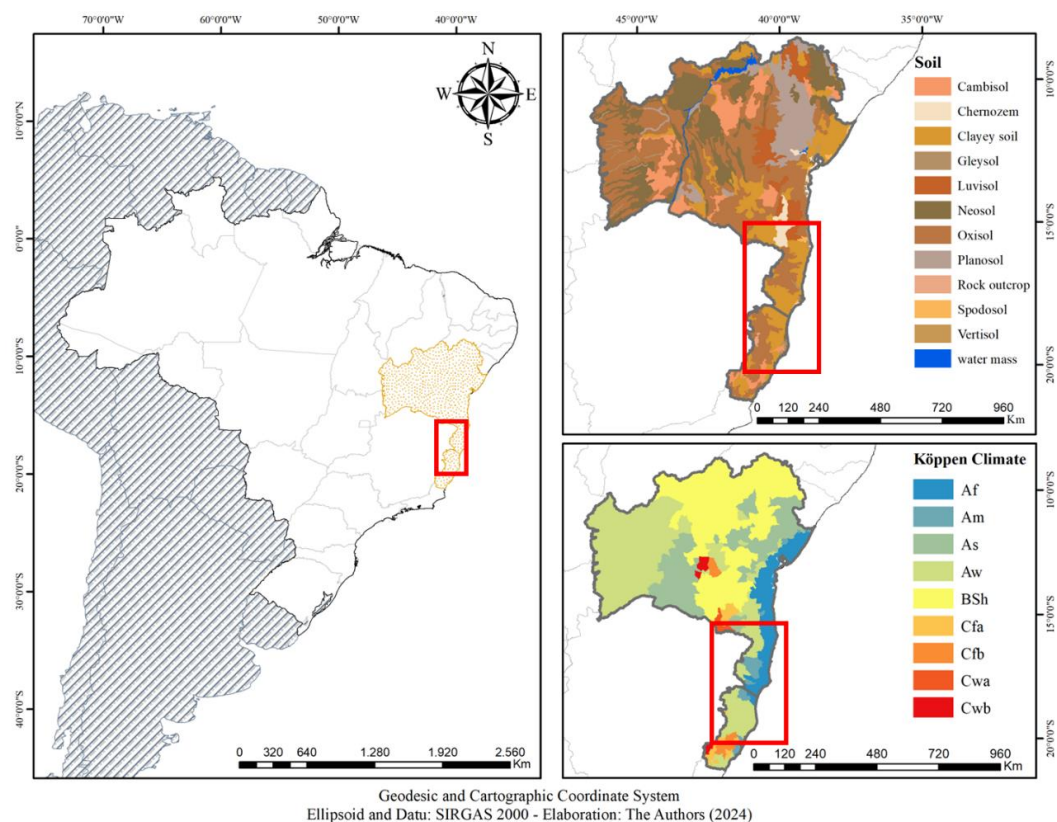
Diante do contexto apresentado, o estudo apresenta como objetivo uma abordagem que permita estimar o índice de sítio das florestas de eucalipto aos 12 meses de idade, utilizando dados edafoclimáticos, informações de satélite e qualidade de plantio, a partir da aplicação de técnicas de *machine learning*. Espera-se que essa abordagem forneça uma ferramenta valiosa para o manejo florestal, permitindo explorar a relação entre o clima e a produtividade das florestas de eucalipto de maneira mais eficiente e econômica.

2 MATERIAL E MÉTODOS

2.1 Caracterização da Área de Estudo

A área de estudo está localizada em dois estados do Brasil, ao sul do estado da Bahia, e ao norte do estado do Espírito Santo (Figura 1). O clima predominante da região é o tropical, com predominância do clima tropical de savana (Aw), clima tropical úmido ou subúmido (Am) e tropical equatorial (Af) (ALVARES et al., 2013). Na área de interesse, são conduzidos plantios de eucalipto para produção de celulose. São plantados na região clones que são cuidadosamente feitos para garantir a melhor adaptação à interação entre genótipo e ambiente (SCOLFARO et al., 2017). A densidade populacional desses plantios pode variar de 1.200 a 1.667 árvores por hectare. Após a implantação dos plantios, diversas práticas silviculturais são adotadas. A precipitação média varia entre 700 e 1500 mm. O solo da região é formado predominantemente por solos latossolos e argilosos.

Figura 16 - Mapa de localização, clima e de solo da área de estudo.



Fonte: Do Autor (2024).

2.2 Caracterização da base de dados de inventário florestal contínuo e cadastrais

Para o estudo presente, as florestas foram mensuradas durante no período de 1º de janeiro de 2019 a junho de 2024. As medições inventariadas pelo inventário florestal contínuo (IFC), apresentam parcelas circulares de 400 m², distribuídas sistematicamente em toda a área. Foram obtidas informações de diâmetro à altura do peito, altura total e de altura dominante baseada no conceito de altura média do povoamento (ASSMANN, 1970).

Foram coletadas informações para florestas aos 2 anos de idade a partir deste conjunto de dados. Para complementar o banco de dados florestal, foram obtidas informações cadastrais que poderiam auxiliar na modelagem. Esses dados incluíam informações sobre materiais genéticos, altitude, espaçamento entre plantas e entrelinhas, idade do plantio, ciclo de rotação (implantação ou condução) e variáveis ambientais, divididas em zona climáticas. Além desses dados, medições de inventário florestal qualitativo foram realizadas aos 4 meses, analisando a altura das árvores, número de fustes e fitossanidade, que foram inseridas no banco de dados. Na tabela 1, é apresentado a tabela descritiva para as diferentes zonas climáticas utilizadas para plantios aos 2 anos.

Tabela 4 - Tabela descritiva das variáveis descritivas para plantios aos dois anos.

Zona Climática	Hd				IS			
	Média	Mínimo	Máximo	Sd	Média	Mínimo	Máximo	Sd
A	16.6	9.32	21.18	1.65	29	26.4	32.6	0.981
B	15.76	8	22.13	2.29	29.6	26.7	32.1	1.08
C	16.32	10.73	20.92	1.73	28.9	25.9	32.1	1.27
D	15.13	11.45	18.46	1.52	28.6	25.6	31.4	1.32
E	14.19	10.53	18.09	1.74	27.4	24.8	29.9	1.04
Zona Climática	Altura (4 meses)				Fustes (4 Meses)			
	Média	Mínimo	Máximo	Sd	Média	Mínimo	Máximo	Sd
A	1.67	0.65	4.05	0.45	1192.14	783.33	1650	124.05
B	1.52	0.55	3.17	0.41	1175.97	683.33	1750	129.81
C	1.55	0.44	3.31	0.46	1168.45	883.33	1820	159.87
D	1.46	0.31	3.11	0.46	1117.13	783.72	1433.33	113.83
E	1.58	0.38	2.95	0.46	1082.19	800	1233.33	84.61

2.3 Caracterização dos dados climáticos

Os dados climáticos foram obtidos do NASA POWER (*Prediction Of Worldwide Energy Resources*). O NASA POWER é uma plataforma de dados climáticos globais com base nos produtos de assimilação *Modern Era Retrospective-Analysis for Research and Applications* (MERRA-2). Os valores de precipitação acumulada, temperatura média, déficit hídrico, excedente hídrico e armazenamento foram obtidos para o período do plantio até o período em que completam 12 meses de idade. As variáveis climáticas foram então atribuídas a cada parcela. A Tabela 2 contém as estatísticas resumidas para as variáveis climáticas na área de estudo.

Tabela 5 -Tabela descritiva das variáveis climáticas.

Zona Climática	Déficit Hídrico (mm)				Precipitação Acumulada (mm)			
	Média	Mínimo	Máximo	Sd	Média	Mínimo	Máximo	Sd
A	-491.96	-902.26	-56.36	125.16	1322.90	796.69	1891.75	251.21
B	-383.09	-720.67	-97.90	145.13	1403.85	965.38	1863.74	188.69
C	-495.32	-882.23	-108.37	178.41	1251.19	804.38	1791.63	213.86
D	-507.01	-877.91	-166.17	188.28	1189.50	712.03	1717.06	217.13
E	-649.81	-1018.86	-370.39	133.39	1030.28	662.50	1310.31	129.08

Zona Climática	Temperatura Média (°C)				Armazenamento Hídrico (mm)			
	Média	Mínimo	Máximo	Sd	Média	Mínimo	Máximo	Sd
A	24.40	22.69	25.18	0.37	619.05	192.05	1679.41	280.49
B	24.39	23.72	25.09	0.33	773.58	221.74	1448.10	314.30
C	24.33	23.55	25.05	0.31	651.86	158.71	1312.90	304.68
D	24.20	23.68	24.95	0.28	577.18	213.47	1218.44	263.17
E	23.97	22.90	24.80	0.34	321.87	88.88	846.04	142.03

2.4 Pré-processamento de imagens Sentinel-2 e cálculo de variáveis

Para a obtenção dos dados de satélite que correspondem aos dados amostrais de campo, foram utilizadas as imagens de satélite Sentinel-2 Nível-2A (Refletância de superfície atmosférica corrigida ortoretificada). Elas foram obtidas através da plataforma do Google Earth Engine (GEE) para o período entre 1º de janeiro de 2019 e junho de 2024. O Sentinel-2 é uma missão de geração de imagens multiespectrais de alta resolução e amplo alcance que dá suporte aos estudos de monitoramento terrestre do Copernicus. Essa missão inclui o monitoramento da vegetação, do solo e da cobertura hídrica, bem como a observação de cursos d'água interiores e

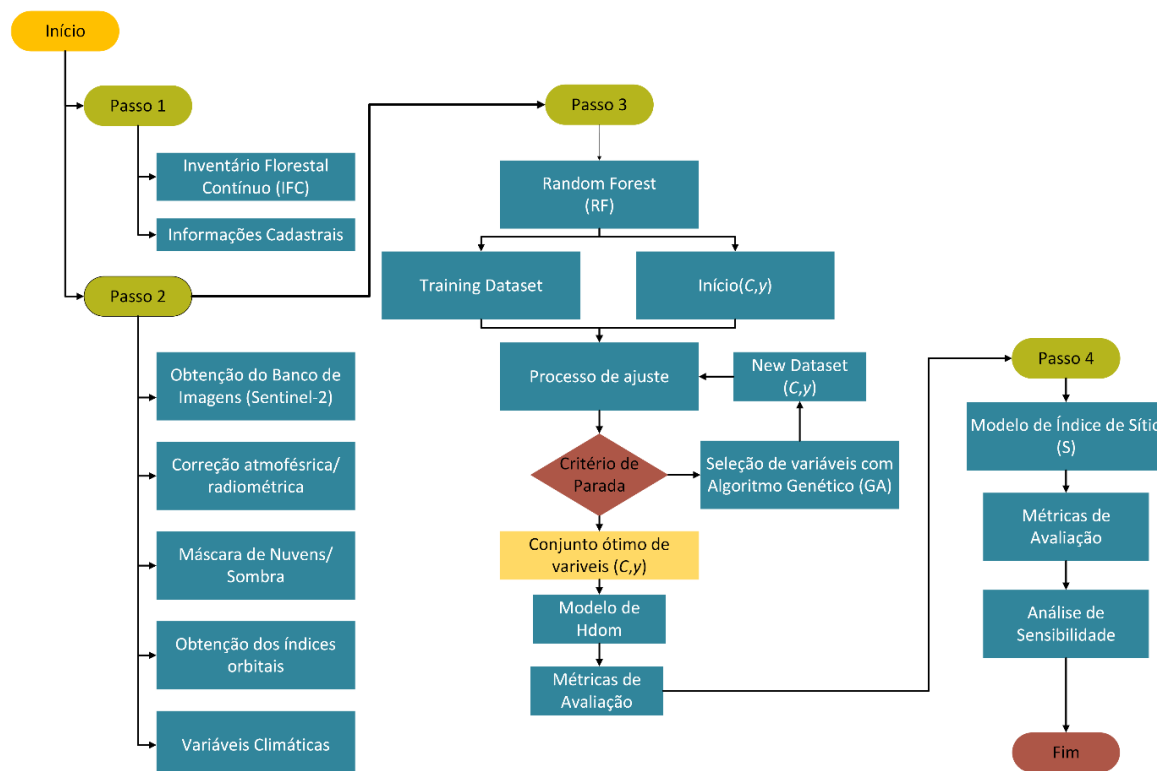
áreas costeiras. Os dados do Sentinel-2 L2 são baixados do SciHub e são processados utilizando o processador Sen2Cor, que usa a entrada Top-Of-Atmosphere (TOA) de nível 1C, que realiza sua correção atmosférica, de terreno e cirrus e gera a reflectância do fundo da atmosfera (BOA) multiplicada por 10.000 (LA CECILIA et al., 2023; MAIN-KNORN et al., 2017).

No estudo foi priorizada a obtenção de imagens que apresentassem baixa incidência de nuvens nos pontos de coletas. Além destes tratamentos da imagem corrigida, utilizou-se o algoritmo Scene Classification Layer (SCL) para mascarar nuvens, sombras e cirros finos defeituosos. Em seguida, extraiu os valores médios das variáveis características usando a plataforma de computação em nuvem GEE para os meses em que ocorreram medições de campo nas parcelas circulares de 400 m². As variáveis incluíram as bandas multiespectrais B2, B3, B4, B5, B6, B7, B8, B8A, B11 e B12. No total, foram obtidos 54 índices derivados das bandas espectrais do Sentinel-2. Os índices calculados estão listados no Anexo A.

2.5 Estimativa do Sítio aos 12 meses

Devido à complexidade das variáveis disponíveis no conjunto de dados construídos. É necessário realizar uma seleção de variáveis que seleciona um conjunto ótimo de variáveis ao modelo. E para modelar a altura dominante aos 2 anos com informações cadastrais, climáticas e de sensoriamento remoto optou-se pelo uso do algoritmo de *machine learning Random Forest* (RF). Foi utilizada a seguinte estratégia: variáveis climáticas + variáveis de sensoriamento + variáveis cadastrais + variáveis de inventário florestal aos 4 meses. Um fluxograma dos passos utilizados no decorrer da metodologia é apresentado na Figura 2.

Figura 17 - Fluxograma da metodologia proposta para estimar qualidade local.



Fonte: Do Autor (2024).

2.5.1 Modelo RF baseado em GA para estimar Hdom aos 12 meses.

O RF foi introduzido Breiman (2001), se consolidou na literatura devido a flexibilidade e capacidade de lidar com multicolinearidade, e apresenta excelentes resultados na modelagem florestal (JUNAID et al., 2023; MA et al., 2023; MIRANDA et al., 2022; REDDY; GEBRESLASIE; ISMAIL, 2017; SILVA et al., 2017; WANG et al., 2016). Incorporar múltiplas variáveis independentes em um modelo pode resultar em redundância e introduzir multicolinearidade. Assim, é crucial selecionar as variáveis mais relevantes para otimizar o desempenho do modelo e facilitar uma compreensão mais clara dos resultados. Neste estudo, utilizou-se um algoritmo genético (GA) para selecionar um conjunto ideal de variáveis independentes para o modelo de classificação RF. A combinação do GA para a seleção de variáveis e o ajuste do modelo RF com base no subconjunto selecionado é referida como GARF.

A eficácia do AG é fortemente influenciada pela sua parametrização. Para ajustar adequadamente o algoritmo, foram realizados testes preliminares. Os operadores do AG incluíram seleção por torneio, cruzamento com um ponto de corte, e mutação aleatória de bits. O critério de parada foi definido como um total de 50 gerações. A diversidade genética foi

mantida com uma população de 100 indivíduos, utilizando uma probabilidade de 10% para mutações e 50% para a troca aleatória de genes.

A avaliação do desempenho do GARF focou em dois critérios principais: a minimização do erro médio quadrático (*out-of-bag* - OOB) na estimativa da variável de interesse e a redução do conjunto de variáveis preditoras (Equação 1). A função de fitness combinou a razão entre o erro OOB das variáveis selecionadas pelo AG e o erro OOB máximo (determinado pelos testes preliminares) com a proporção entre o número de variáveis selecionadas pelo AG e o total de variáveis testadas (Equação 1). Este enfoque enfatiza o equilíbrio entre a precisão da classificação e a simplicidade do modelo. O pacote randomForest (LIAW; WIENER, 2002) em R Development Core Team (2024) foi usado.

$$fitness = \left(\frac{erroOOB}{erroOOB_{max}} + \frac{n}{NVT} \right) \quad (1)$$

Em que: *erroOOB*: erro médio quadrático (*out-of-bag* - OOB); *n*: número de variáveis selecionadas; *NVT*: número de variáveis total.

2.5.2 Estimativa do Índice de Sítio aos 7 anos com base no Hdom aos 12 meses

Scolforo (2018) sugeriu a modelagem da qualidade do local com o objetivo de estimar a produtividade incluindo variáveis climáticas. Portanto, o cálculo da qualidade do local segue a abordagem proposta por Scolforo (SCOLFORO et al., 2019a). A equação de nível de povoamento para estimar o crescimento e a produtividade no Brasil foi desenvolvida com base no déficit hídrico (SWD) e na altura dominante (H). A equação utilizada para calcular o índice de sítio está apresentada na Equação 2.

$$\hat{H}_2 = [35.4970 + 0.0085 SWD_2][1 - \exp(-0.3711t_2)]^{\frac{\ln\left(\frac{H_1}{[35.4970 + 0.0085 SWD_1]}\right)}{\ln[1 - \exp(-0.3711t_1)]}} \quad (2)$$

Em que: H_1 : altura dominante estimada para 2 anos; SWD_1 : é o déficit hídrico médio anual do solo (mm ano⁻¹) desde o momento em que as parcelas foram instaladas ($t = 0$) até t_1 ; SWD_2 : é o déficit hídrico médio anual do solo (mm ano⁻¹) desde o momento em que as parcelas foram instaladas ($t = 0$) até t_2 .

2.5.3 Avaliação e validação das abordagens de modelagem

Para avaliar o desempenho das equações de altura dominante foram usados o viés médio (Bias) (Equação 3), erro absoluto médio (MAE) (Equação 4), erro quadrático médio (RMSE) (Equação 5) e o coeficiente de determinação (R^2) (Equação 6). Além disso, foram usadas análises gráficas para avaliar o comportamento do modelo.

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (6)$$

Em que: Y_i é o valor observado da variável de interesse (anomalia detectada em campo); $\hat{Y}_i$ é o valor estimado da variável de interesse; $\bar{Y}$ é o valor médio da variável de interesse.

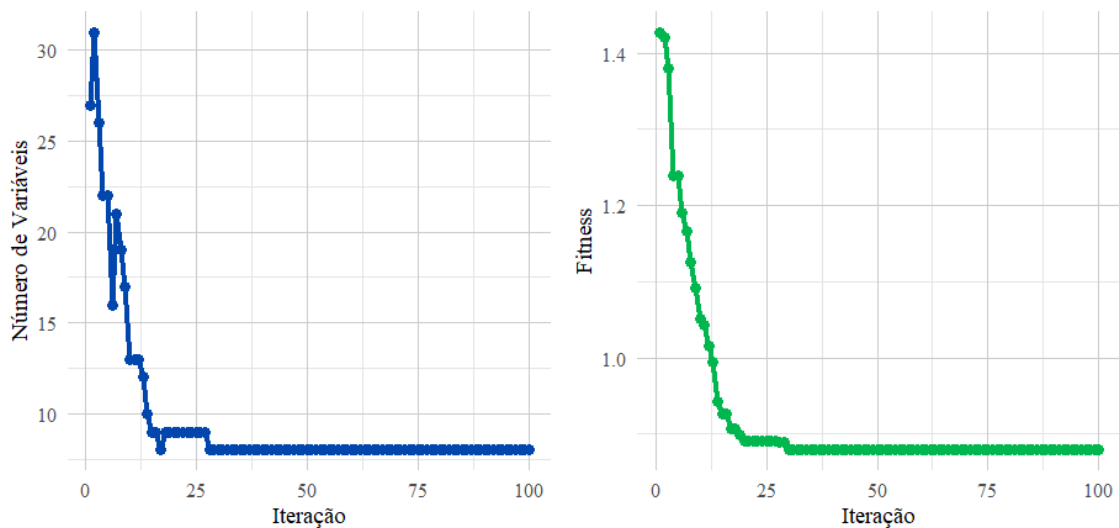
Para a validação do modelo de hdom e de índice de sítio aos 7 anos, será conduzida utilizando um conjunto de dados dividido em duas partes: 30% dos dados serão destinados ao ajuste do modelo, enquanto os 70% restantes serão utilizados para a validação. Esse método garantirá que o modelo seja treinado adequadamente e testado de maneira robusta. Além da validação básica, uma análise de sensibilidade será realizada para avaliar como o modelo responde a variações em variáveis climáticas. Esta análise incluirá simulações de precipitação em cenários extremos, incluindo precipitação menor e maior. Além das simulações com dados climáticos, será aplicado análise de sensibilidade a 10 clones de eucalipto mais produtivos, e outros 10 clone menos produtivos, representando diferentes taxas de crescimento: rápido e lento. Essa abordagem permite a criação de possíveis cenários climáticos. Com isso, é possível analisar possíveis flutuações climáticas que podem ocorrer e impactar na produtividade local.

3 RESULTADOS

3.1 Seleção de variáveis

Nesse estudo, o tamanho do subconjunto de recursos e a precisão da regressão foram utilizados como critérios de avaliação de desempenho. Primeiramente, nos experimentos, é apresentado o desempenho da seleção de variáveis baseada em *wrapper*, realizada pelo algoritmo AG, bem como o comportamento da solução de fitness ao longo das gerações (Figura 3).

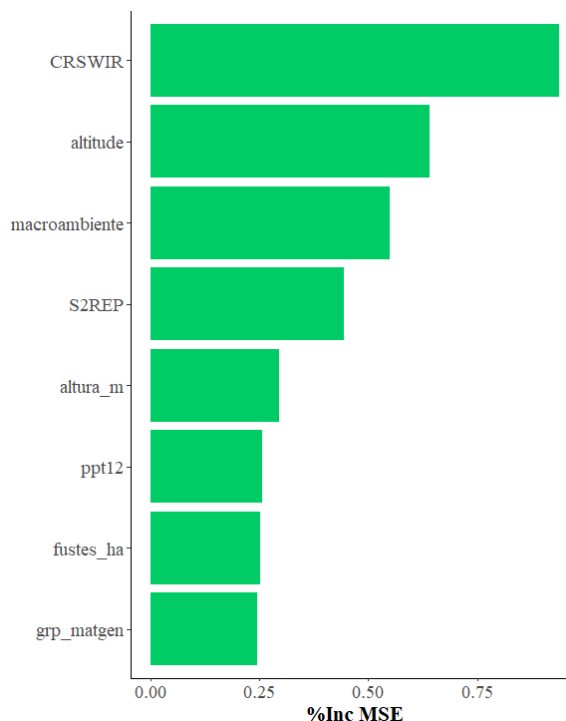
Figura 18 - Gráficos de redução do número de variáveis e melhoria da função *fitness* do modelo GARF em função do número de gerações.



Fonte: Do Autor (2024).

A Figura 3 registra o número de variáveis selecionadas pela abordagem de seleção de variáveis baseada na evolução do *wrapper* para o conjunto de dados. Observando o gráfico, nota-se que essa abordagem reduz consideravelmente a dimensionalidade, escolhendo uma pequena parte das características originais. Do total de variáveis, foram selecionadas oito variáveis para o modelo GARF. Entre elas, duas são do primeiro inventário realizado aos 120 dias (altura e número de fustes), duas são variáveis de sensoriamento remoto (CRWIR e S2REP), duas variáveis categóricas de cadastro (grupo de material genético e macroambiente), uma variável climática (precipitação acumulada aos 12 meses de idade) e uma variável de terreno (altitude). Cada uma dessas variáveis selecionadas apresenta um grau de contribuição para modelar a altura dominante, conforme mostrado na Figura 4.

Figura 19 - Valores médio de importância das variáveis explicativas dos modelos para estimar a altura dominante aos 2 anos com variáveis aos 12 meses.



Fonte: Do Autor (2024).

3.2 Estimativa de Hdom

Os dados descritos nas seções 2.3, 2.4 e 2.5 foram utilizados para estimar a altura dominante (Hdom). Nesta seção, focou-se nos resultados dessa estimativa. Foi gerado um modelo para estimar Hdom, apresentando excelente desempenho. O GARF, que utilizou variáveis aos 4 meses de idade, apresentou um RMSE de 9,20%. As métricas de precisão do modelo foram excelentes. O viés foi zero, indicando que os erros se anulam e se aproximam da normalidade. A Tabela 3 apresenta outras métricas utilizadas para avaliar a precisão dos modelos para o conjunto de dados de validação.

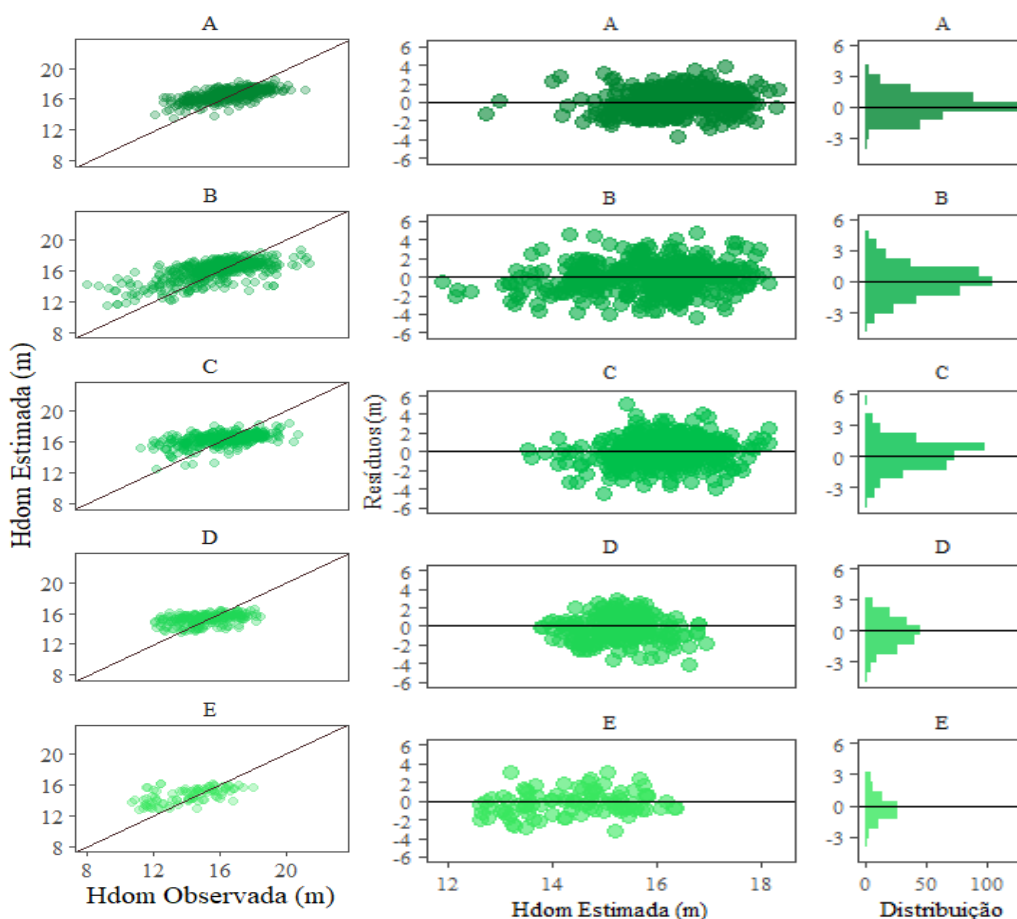
Tabela 6 - Métricas de avaliação dos modelos de altura dominante.

Modelo	Bies (m)	MAE (m)	RMSE (m)
GARF	0	1,17	1,47

Em que: MAE: erro absoluto médio; RMSE: erro quadrático médio.

A Figura 5 exibe três gráficos: correlação entre valores estimados e observados, resíduos, e dispersão dos resíduos para as cinco zonas climáticas na estimativa de Hdom. O modelo mostra correlação e dispersão semelhantes para cada uma das zonas climáticas. A distribuição dos resíduos sugere normalidade, indicando que os modelos são aderentes. Observa-se que o modelo tende a errar mais nos extremos devido à dispersão dos resíduos. Há um efeito de saturação: a altura dominante para florestas de 2 anos, utilizando informações aos 12 meses, é superestimada abaixo de 10 m e subestimada acima de 20 m.

Figura 20 - Gráfico de correlação da altura dominante observada x estimada, resíduos e distribuição dos resíduos.



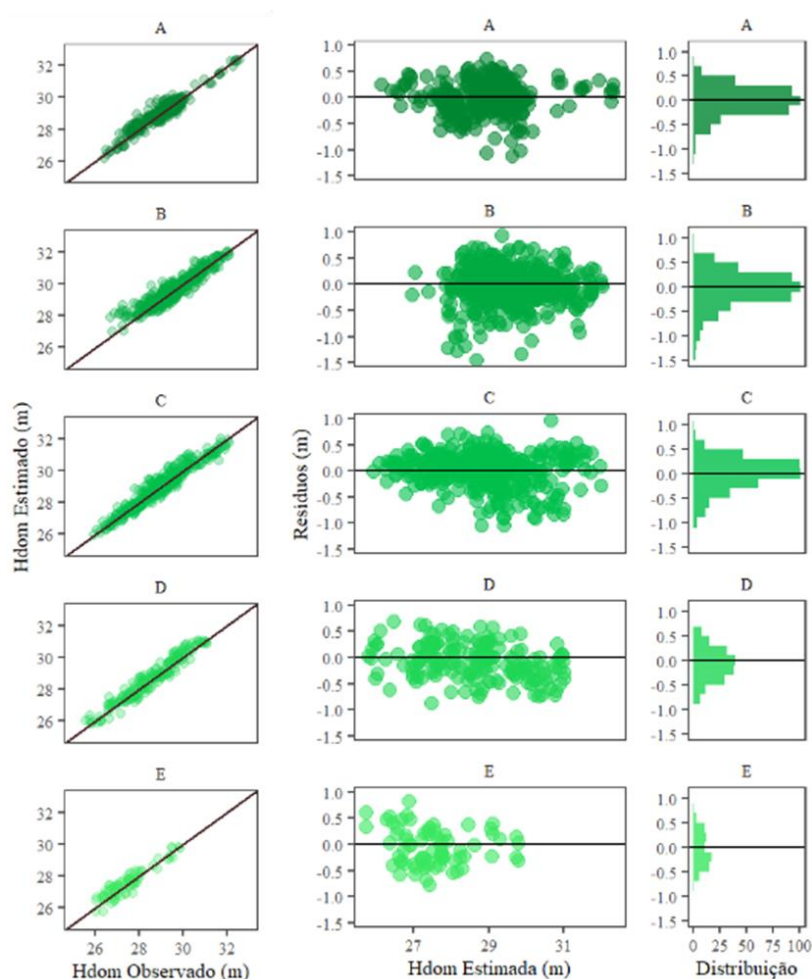
Fonte: Do Autor (2024).

3.3 Estimativa do índice de sítio

Com base na premissa de que o índice de sítio corresponde à altura dominante em uma idade específica, utilizou-se a equação de Scolforo et al. (2019) para projetar a altura dominante

aos 7 anos na base de dados de validação. Em seguida, foi realizada uma comparação gráfica envolvendo a correlação entre valores estimados e observados, resíduos, e a dispersão dos resíduos para as cinco zonas climáticas na estimativa de Hdom. O modelo apresentou correlação e dispersão semelhantes entre as zonas climáticas. A análise da distribuição dos resíduos indicou uma tendência à normalidade, sem tendência de superestimativa ou dos valores dos sítios encontrados aos 7 anos. A dispersão dos resíduos mostrou o GARF tende a apresentar maiores erros nos sítios menos produtivos da zona climática B.

Figura 21 - Gráfico de correlação da altura dominante aos 7 anos observada x estimada, resíduos e distribuição dos resíduos.



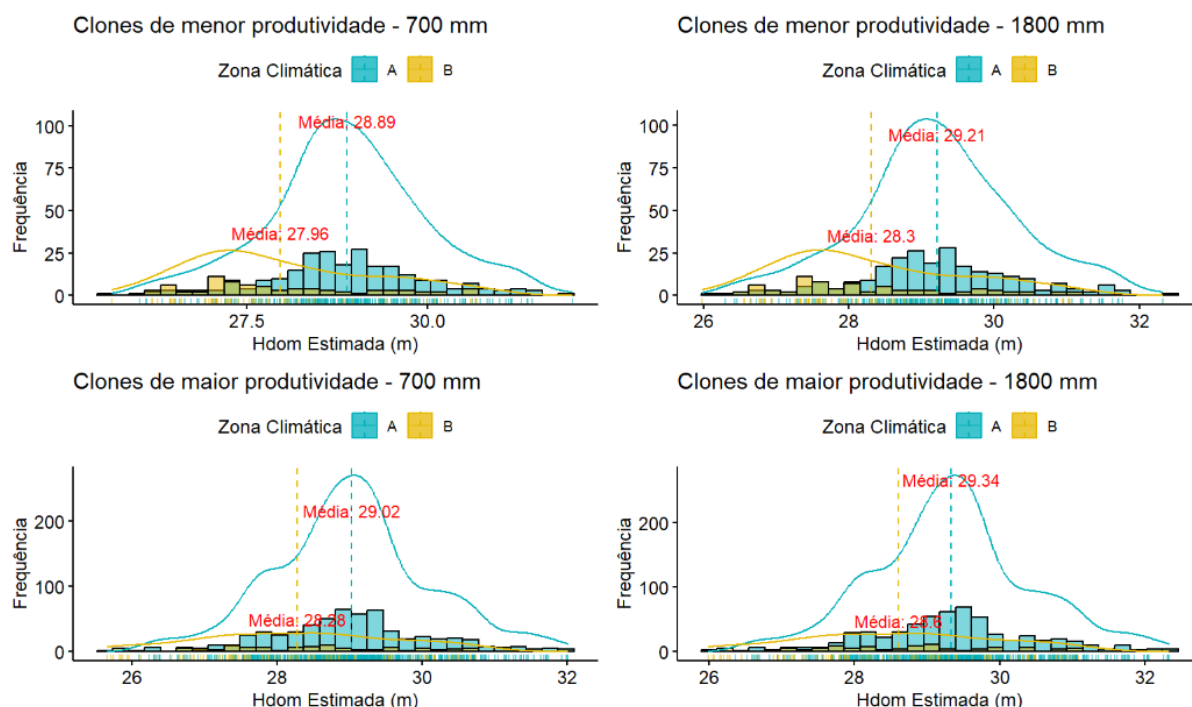
Fonte: Do Autor (2024).

3.4 Simulação do IS em relação ao clima e à qualidade do plantio

Para explorar a sensibilidade climática, foram construídos quatro cenários que exploram os extremos climáticos. Primeiramente, considerou-se uma precipitação acumulada de 700

milímetros no primeiro ano de vida da floresta e comparou-se a produtividade dos 10 clones menos produtivos com a dos 10 clones mais produtivos sob o mesmo regime hídrico. Em seguida, dois outros cenários foram simulados, com uma precipitação acumulada de 1800 milímetros, novamente comparando os 10 clones menos produtivos com os 10 clones mais produtivos. Para critérios de comparação, as zonas climáticas A, B e C foram agrupadas e denominadas como Zona Climática A, enquanto as zonas D e E foram agrupadas como Zona Climática B. Isso permitiu uma amostragem suficiente de clones desde os menos produtivos até os mais produtivos. Foi utilizando somente o modelo GARF para esta análise. A Figura 7 apresenta o histograma de frequência e as médias climáticas das amostras projetadas para a idade de 7 anos, utilizando o modelo de Scolforo et al. (2019). Observa-se que os clones menos produtivos na Zona Climática A e B apresentam uma melhora acima de 0,32 metros na média de Hdom em climas com maior precipitação. Os clones mais produtivos, por sua vez, apresentam uma melhoria de 0,32 metros na média de Hdom em condições semelhantes. Esses resultados indicam que os modelos de Hdom, utilizando variáveis climáticas e sensoriamento, são sensíveis às mudanças climáticas.

Figura 22 - Histograma de frequência dos cenários testados para a análise de sensibilidade.



Fonte: Do Autor (2024).

4 DISCUSSÃO

Com os avanços tecnológicos, o monitoramento orbital das florestas se torna cada vez mais viável. Devido à homogeneidade do plantio de florestas de eucalipto, é possível e eficiente realizar estimativas de produtividade e crescimento dessas florestas utilizando variáveis climáticas e de satélite. Antecipar o inventário florestal contínuo, tradicionalmente realizado aos 2 anos de vida das florestas de eucalipto, para o primeiro ano de idade utilizando essas variáveis, possibilita uma série de análises preditivas que, de outra forma, só poderiam ser feitas aos 2 anos.

Este estudo apresentou uma metodologia inovadora que permite realizar o inventário florestal utilizando informações climáticas e de sensoriamento remoto para gerar *insights* sobre a produtividade e o potencial de crescimento das florestas. Um modelo de *machine learning* foi empregado para modelar a altura dominante das florestas aos 2 anos de idade, utilizando dados obtidos quando as florestas tinham apenas 1 ano. Isso permitiu diagnosticar precocemente o potencial de crescimento dos plantios florestais. Os resultados mostraram que os modelos foram eficientes para estimar a altura dominante (Hdom). A seleção de variáveis foi realizada utilizando um GA, que comprovam a capacidade de encontrar um subconjunto ótimo de variáveis que explicam melhor a variável preditora (HONG et al., 2018; HU et al., 2023; MIRANDA et al., 2022; SRIVASTAVA et al., 2023). O algoritmo RF apresentou uma excelente performance na estimativa do Hdom, demonstrando uma grande capacidade de generalização e ajudando a evitar o *overfitting* (GAO; LIN; HU, 2023; SHI; ZHANG, 2023). Além disso, o RF permitiu combinar variáveis de diferentes bancos de dados em um modelo explicativo, ajudando a entender o potencial de crescimento das florestas através dessas variáveis.

Entre as variáveis de sensoriamento remoto selecionadas, constatou-se que o CRswir é um espectro importante para mapear a altura dominante. Mouret et al. (2024) identificaram a capacidade do CRswir em monitorar a morte regressiva de carvalhos. As bandas visível e infravermelha de onda curta (SWIR - bandas 11 e 12) e NIR (B8A) que compõem o índice, demonstraram grande capacidade de monitoramento florestal. O SWIR é menos susceptível a interferências atmosféricas e é usado para monitorar a saúde das plantas, estresse hídrico, conteúdo relativo de água, potencial hídrico da folha, condutância estomática e elasticidade da parede celular (CARELLA et al., 2024; HUO; PERSSON; LINDBERG, 2021; LE; HARPER; DELL, 2023; WATT et al., 2021). Os dados SWIR são cruciais, pois fornecem informações valiosas sobre a presença de água líquida no dossel da floresta, diretamente relacionada ao

estresse hídrico (LE; HARPER; DELL, 2023). A relação entre as bandas SWIR mostrou-se eficiente para monitorar a salubridade do dossel da floresta e correlacioná-la com a estima de altura dominante (Hdom). Embora menos importante, a seleção do S2REP (Sentinel-2 Red-Edge Position Index) também foi destacada pelo método proposto. O S2REP é utilizado para monitorar clorofila e é sensível ao nitrogênio da cultura, bem como ao estado de crescimento (ALI et al., 2022; GAO et al., 2024; HALLIK et al., 2019). No estudo de Rahimzadeh-Bajgiran et al. (2020), entre todas as variáveis Sentinel-2 testadas, as bandas B3, B8A e S2REP foram identificadas como as melhores para prever a altura (HT) em povoamentos jovens, enquanto o S2REP foi a melhor variável para prever a taxa de crescimento de biomassa em todas as condições florestais. Isso corrobora com o presente estudo, evidenciando a importância do S2REP para estimar a produtividade potencial florestal.

A adição da variável precipitação anual fornece uma melhoria modesta na modelagem da altura dominante. Porém, torna-o mais preciso para fornecer projeções específicas de sítio, ou para imitar mudanças climáticas de curto prazo. Embora a inclusão de variáveis climáticas, como a precipitação anual, nos modelos de altura dominante, torna-o mais sensível e perceptível à mudanças climáticas (SCOLFORO et al., 2016). Outras variáveis climáticas, como o déficit hídrico anual do solo e a disponibilidade de água no solo, a sua inclusão em modelos de crescimento de altura dominante, aumentam a capacidade explicativa ao avaliar a qualidade do local para povoamentos clonais de eucalipto no Brasil (SCOLFORO et al., 2017, 2020). A produtividade das plantações clonais de eucalipto, bem como de outras espécies de árvores, é influenciada pela disponibilidade de água no solo (SCOLFORO et al., 2020; SPITTLEHOUSE, 2003). Isso implica que a precipitação anual desempenha um papel fundamental no crescimento da floresta (SCOLFORO et al., 2020). A inclusão de variáveis como a precipitação anual é crucial para modelar o crescimento florestal de forma precisa e eficaz.

A inclusão da variável da altitude do terreno é um indicativo de sua importância para a precisão do modelo. Ela apresenta uma alta pontuação de importância no modelo RF, evidenciando diferenças na distribuição vertical da altura dominante nas florestas de eucalipto na região de interesse. A reflectância dos índices espectrais varia de acordo com as características do terreno. Para uma maior precisão nos modelos de altura dominante do dossel, é crucial combinar variáveis topográficas e de sensoriamento remoto. A altitude também é usada para estimar a biomassa acima do solo (FAN et al., 2022) e foi utilizada para classificar espécies de árvores nas regiões orientais das montanhas Qilian, na China (MA et al., 2021).

A inclusão de variáveis de cadastro como grupos clonais e zonas climáticas, embora menos importante, também foram selecionadas. Cada grupo apresenta uma taxa de crescimento

diferente de acordo com a zona climática, essa distinção é encontrada com a inclusão destas variáveis. Características fisiológicas semelhantes compartilhadas pelos clones em um grupo podem permitir o desenvolvimento de clones com maior produção (GONÇALVES, 1997; SCOLFORO et al., 2019b). Embora com menor importância entre as variáveis selecionadas, a seleção de variáveis do inventário aos 4 meses, altura e número de fustes. Trazem uma estimativa inicial e potencial destes plantios, que se caracteriza por estimativas mais precisas ao combinar com outras variáveis de outros conjuntos de dados.

A combinação de variáveis cadastrais, topográficas, climáticas e de sensoriamento remoto aos 1 anos de idade pode estimar com precisão a altura dominante das árvores de eucalipto aos 2 anos de idade. É importante ressaltar a metodologia proposta permite antecipar em um ano a estimativa de produtividade e de crescimento florestal. Permite assim, antecipar eventos de ocorrência da floresta, e planejar com maior assertividade o planejamento florestal. E realocar com maior precisão os recursos disponíveis. Como está sendo antecipado informações de inventário florestal por meio desta nova abordagem, possibilita encontrar os sítios menos produtivos, e como está numa etapa inicial do plantio, 1 ano de idade, recomendar ações corretivas que viabilizam e potencializam o crescimento produtividade daqueles locais com menor taxa de crescimento. Novos estudos devem ser desenvolvidos nesta linha de pesquisa, focando no desenvolvimento de modelos que estimem o estoque florestal atual e permitam prever com precisão o estoque futuro. Uma possibilidade para melhorar a precisão desses modelos é a inclusão de variáveis de LiDAR orbital, como o GEDI.

5 CONCLUSÃO

Informações precisas sobre as florestas são essenciais para estimar a qualidade e o potencial das florestas locais. As características do Sentinel-2, como o tempo de revisita frequente, a disponibilidade gratuita, as múltiplas bandas para monitoramento espacial e a resolução espacial, o tornam aplicável para aplicações florestais. No entanto, para um monitoramento preciso, é crucial integrar as informações orbitais disponíveis com os dados de inventário de campo. Além disso, a inclusão de variáveis climáticas permite que o modelo de altura dominante seja sensível a mudanças climáticas, gerando cenários mais realistas.

A utilização de informações satelitais e climáticas, associadas a dados cadastrais, possibilita determinar com precisão e acurácia a altura dominante da área de estudo. Para isso, utilizou-se uma metodologia de seleção de variáveis por meio de um algoritmo genético para diminuir a dimensionalidade das variáveis explicativas, alcançando êxito na proposta. No

modelo de altura dominante, foram selecionadas as seguintes variáveis para o modelo de *Random Forest*: altura aos 4 meses, número de fustes aos 4 meses, variáveis de sensoriamento remoto (CRWIR e S2REP) - médias do primeiro ano, grupo de material genético, zona climática, precipitação anual no primeiro ano e altitude do terreno. A combinação dessas variáveis aos 1 anos de idade dos plantios florestais permite estimar com precisão a altura dominante aos 2 anos. Isso possibilita antecipar o inventário florestal, que normalmente ocorreria apenas aos 2 anos de idade.

Essa antecipação do inventário florestal gera diversas vantagens, como a identificação precoce das florestas com menor produtividade, permitindo que os gestores florestais implementem ações corretivas e de mitigação. Além disso, possibilita a projeção da produtividade local e a criação de cenários que diagnosticam como os efeitos climáticos podem interferir na produtividade local. A metodologia proposta alcança sucesso na antecipação do inventário florestal de 2 anos para 1 ano de idade, tornando-se uma ferramenta importante para a tomada de decisões pelos gestores florestais.

REFERÊNCIAS

- ALI, A. et al. Evaluating Sentinel-2 red edge through hyperspectral profiles for monitoring LAI & chlorophyll content of Kinnow Mandarin orchards. **Remote Sensing Applications: Society and Environment**, v. 26, p. 100719, abr. 2022.
- ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.
- ASSMANN, E. **The Principles of Forest Yield Study: Studies in the Organic Production, Structure, Increment and Yield of Forest Stands**. [s.l.] Elsevier, 1970.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- BURKHART, H. E.; TOMÉ, M. **Modeling Forest Trees and Stands**. Dordrecht: Springer Netherlands, 2012.
- CARELLA, A. et al. Continuous Plant-Based and Remote Sensing for Determination of Fruit Tree Water Status. **Horticulturae**, v. 10, n. 5, p. 516, 16 maio 2024.
- FAN, X. et al. Sentinel-2 Images Based Modeling of Grassland Above-Ground Biomass Using Random Forest Algorithm: A Case Study on the Tibetan Plateau. **Remote Sensing**, v. 14, n. 21, p. 5321, 24 out. 2022.
- GAO, C.; LIN, H.; HU, H. Forest-Fire-Risk Prediction Based on Random Forest and Backpropagation Neural Network of Heihe Area in Heilongjiang Province, China. **Forests**, v. 14, n. 2, p. 170, 17 jan. 2023.
- GAO, S. et al. Assessment of remote-sensed vegetation indices for estimating forest chlorophyll concentration. **Ecological Indicators**, v. 162, p. 112001, maio 2024.
- GONÇALVES, F. M. A. **Adaptabilidade e estabilidade de cultivares de milho avaliadas em “safrinha” no período de 1993 a 1995**. Dissertação—Lavras: Universidade Federal de Lavras, 1997.
- HALLIK, L. et al. Reflectance Properties of Hemiboreal Mixed Forest Canopies with Focus on Red Edge and Near Infrared Spectral Regions. **Remote Sensing**, v. 11, n. 14, p. 1717, 20 jul. 2019.
- HONG, H. et al. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. **Science of the Total Environment**, v. 630, p. 1044–1056, 2018.
- HU, Y. et al. Estimating Forest Stock Volume in Hunan Province, China, by Integrating In Situ Plot Data, Sentinel-2 Images, and Linear and Machine Learning Regression Models. **Remote Sensing**, v. 12, n. 1, p. 186, 4 jan. 2020.
- HU, Y. et al. Improving the Potential of Coniferous Forest Aboveground Biomass Estimation by Integrating C- and L-Band SAR Data with Feature Selection and Non-Parametric Model. **Remote Sensing**, v. 15, n. 17, p. 4194, 25 ago. 2023.

HUO, L.; PERSSON, H. J.; LINDBERG, E. Early detection of forest stress from European spruce bark beetle attack, and a new vegetation index: Normalized distance red & SWIR (NDRS). **Remote Sensing of Environment**, v. 255, p. 112240, mar. 2021.

IBÁ. **Anuário Estatístico da Indústria Brasileira de Árvores**. [s.n.]. Brasília, 2024. Disponível em <https://iba.org/datafiles/publicacoes/relatorios/relatorio2024.pdf>. Acesso em: 31 jul. 2024.

ILES, K. **A Sampler of Inventory Topics: A Practical Discussion for Resource Samplers, Concentrating on Forest Inventory Techniques**. 1. ed. Nanaimo, B.C: Kim Iles & Associates Ltd., 2003.

JUNAID, M. et al. Mapping LULC Dynamics and Its Potential Implication on Forest Cover in Malam Jabba Region with Landsat Time Series Imagery and Random Forest Classification. **Sustainability**, v. 15, n. 3, p. 1858, 18 jan. 2023.

LA CECILIA, D. et al. Pixel-based mapping of open field and protected agriculture using constrained Sentinel-2 data. **ISPRS Open Journal of Photogrammetry and Remote Sensing**, v. 8, p. 100033, abr. 2023.

LACERDA, T. H. S. et al. Feature selection by genetic algorithm in nonlinear taper model. **Canadian Journal of Forest Research**, 20 jan. 2022.

LE MAIRE, G. et al. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. **Remote Sensing of Environment**, v. 115, n. 10, p. 2613–2625, out. 2011.

LE, T. S.; HARPER, R.; DELL, B. Application of Remote Sensing in Detecting and Monitoring Water Stress in Forests. **Remote Sensing**, v. 15, n. 13, p. 3360, 30 jun. 2023.

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, p. 1073, 25 nov. 2019.

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

LOPES, I. L. E et al. A comparative approach of methods to estimate machine productivity in wood cutting. **International Journal of Forest Engineering**, v. 33, n. 1, p. 43–55, 2022.

MA, M. et al. Tree Species Classification Based on Sentinel-2 Imagery and Random Forest Classifier in the Eastern Regions of the Qilian Mountains. **Forests**, v. 12, n. 12, p. 1736, 9 dez. 2021.

MA, T. et al. A Novel Vegetation Index Approach Using Sentinel-2 Data and Random Forest Algorithm for Estimating Forest Stock Volume in the Helan Mountains, Ningxia, China. **Remote Sensing**, v. 15, n. 7, p. 1853, 30 mar. 2023.

- MAIN-KNORN, M. et al. **Sen2Cor for Sentinel-2**. (L. Bruzzone, F. Bovolo, J. A. Benediktsson, Eds.) Image and Signal Processing for Remote Sensing XXIII. **Anais...SPIE**, 4 out. 2017.
- 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, jan. 2022.
- MOURET, F. et al. Toward an Operational Monitoring of Oak Dieback With Multispectral Satellite Time Series: A Case Study in Centre-Val De Loire Region of France. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 17, p. 643–659, 2024.
- MURA, M. et al. Exploiting the capabilities of the Sentinel-2 multi spectral instrument for predicting growing stock volume in forest ecosystems. **International Journal of Applied Earth Observation and Geoinformation**, v. 66, p. 126–134, abr. 2018.
- R CORE TEAM. **R: A language and environment for statistical computing**. ViennaR Foundation for Statistical Computing, , 2024.
- RAHIMZADEH-BAJGIRAN, P. et al. Forest Potential Productivity Mapping by Linking Remote-Sensing-Derived Metrics to Site Variables. **Remote Sensing**, v. 12, n. 12, p. 2056, 26 jun. 2020.
- REDDY, N.; GEBRESLASIE, M.; ISMAIL, R. A hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data. **South African Journal of Geomatics**, v. 6, n. 3, p. 377, 2017.
- SABATIA, C. O.; BURKHART, H. E. Predicting site index of plantation loblolly pine from biophysical variables. **Forest Ecology and Management**, v. 326, p. 142–156, ago. 2014.
- SANTOS, J. S. et al. Predicting eucalyptus plantation growth and yield using Landsat imagery in Minas Gerais, Brazil. **Ecological Informatics**, v. 75, p. 102120, jul. 2023.
- SCOLFORO, H. F. et al. Modeling dominant height growth of eucalyptus plantations with parameters conditioned to climatic variations. **Forest Ecology and Management**, v. 380, p. 182–195, nov. 2016.
- SCOLFORO, H. F. et al. Incorporating rainfall data to better plan eucalyptus clones deployment in eastern Brazil. **Forest Ecology and Management**, v. 391, p. 145–153, maio 2017.
- SCOLFORO, H. F. **Development of a modeling framework for clonal Eucalypt plantations subject to different environmental conditions in Brazil**. 2018. PhD Dissertation - North Carolina State University, Raleigh, 2018.
- SCOLFORO, H. F. et al. Eucalyptus growth and yield system: Linking individual-tree and stand-level growth models in clonal Eucalypt plantations in Brazil. **Forest Ecology and Management**, v. 432, p. 1–16, jan. 2019a.

- SCOLFORO, H. F. et al. Yield pattern of eucalypt clones across tropical Brazil: An approach to clonal grouping. **Forest Ecology and Management**, v. 432, p. 30–39, jan. 2019b.
- SCOLFORO, H. F. et al. Site index estimation for clonal eucalypt plantations in Brazil: A modeling approach refined by environmental variables. **Forest Ecology and Management**, v. 466, p. 118079, jun. 2020.
- SHI, C.; ZHANG, F. A Forest Fire Susceptibility Modeling Approach Based on Integration Machine Learning Algorithm. **Forests**, v. 14, n. 7, p. 1506, 24 jul. 2023.
- SILVA, C. A. et al. Predicting stem total and assortment volumes in an industrial Pinus taeda L. forest plantation using airborne laser scanning data and random forest. **Forests**, v. 8, n. 7, p. 1–17, 2017.
- SPITTLEHOUSE, D. L. Water Availability, Climate Change and the Growth of Douglas-Fir in the Georgia Basin. **Canadian Water Resources Journal / Revue canadienne des ressources hydriques**, v. 28, n. 4, p. 673–688, 23 jan. 2003.
- SRIVASTAVA, A. K. et al. A Day-Ahead Short-Term Load Forecasting Using M5P Machine Learning Algorithm along with Elitist Genetic Algorithm (EGA) and Random Forest-Based Hybrid Feature Selection. **Energies**, v. 16, n. 2, p. 867, 12 jan. 2023.
- WANG, L. et al. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. **Crop Journal**, v. 4, n. 3, p. 212–219, 2016.
- WATT, M. S. et al. Long-term effects of water stress on hyperspectral remote sensing indicators in young radiata pine. **Forest Ecology and Management**, v. 502, p. 119707, dez. 2021.
- ZHANG, N. et al. Forest Height Mapping Using Feature Selection and Machine Learning by Integrating Multi-Source Satellite Data in Baoding City, North China. **Remote Sensing**, v. 14, n. 18, p. 4434, 6 set. 2022.
- ZHOU, Y.; FENG, Z. Estimation of Forest Stock Volume Using Sentinel-2 MSI, Landsat 8 OLI Imagery and Forest Inventory Data. **Forests**, v. 14, n. 7, p. 1345, 29 jun. 2023.

Anexo A

Tabela 7 - Descrição dos índices orbitais de Sentinel-2.

Índice (Sigla)	Nome (Descrição)	Fórmula
ARI1	Anthocyanin Reflectance Index	$ARI1 = \frac{1}{B3} - \frac{1}{B5}$
ARI2	Anthocyanin Reflectance Index 2	$ARI2 = \left(\frac{B8}{B2}\right) - \left(\frac{B8}{B3}\right)$
ARVI	Atmospherically Resistant Vegetation Index	$ARVI = \frac{B8 - (2 * B4 - B2)}{B8 + (2 * B4 - B2)}$
BAI	Burn Area Index	$BAI = \frac{1}{((0.1 - B4)^2 + (0.06 - B8)^2)}$
BAI2	Burn Area Index 2	$BAI2 = \left(1 - \sqrt{\left(\frac{B6 * B7 * B8A}{B4}\right)}\right) * \left(\frac{(B12 - B8A)}{\sqrt{B12 + B8A}} + 1\right)$
CHLRE	Chlorophyll Red Edge Index	$CHLRE = \frac{B5}{B8}$
CRI1	Carotenoid Reflectance Index 1	$CRI1 = \left(\frac{1}{B2}\right) - \left(\frac{1}{B3}\right)$
CRI2	Carotenoid Reflectance Index 2	$CRI2 = \left(\frac{1}{B2}\right) - \left(\frac{1}{B5}\right)$
CRSWIR	Continuum Removal Short Wave Infrared	$CRSWIR = \frac{B11}{B8A + (1613.7 - 864.8) * \frac{(B12 - B8A)}{(2202.4 - 864.8)}}$
DSWI	Drought Stress Water Index	$DSWI = \frac{(B7 + B3)}{(B11 + B4)}$
EVI	Enhanced Vegetation Index	$EVI = EVI = \frac{2.5 * (B8 - B4)}{(B8 + 6 * B4 - (7.5 * B2 + 1))}$
EVI2	Enhanced Vegetation Index 2	$EVI2 = \frac{2.5 * (B8 - B4)}{(B8 + 2.4 * B4 + 1)}$
GNDVI	Green Normalized Difference Vegetation Index	$GNDVI = \frac{(B8 - B3)}{(B8 + B3)}$
GRVI	Green Red Vegetation Index	$GRVI = \frac{(B3 - B4)}{(B3 + B4)}$
GSAVI	Green Soil Adjusted Vegetation Index	$GSAVI = 1.5 * \left(\frac{B8 - B3}{B8 + B3 + 0.5}\right)$
IRECI	Inverted Red-Edge Chlorophyll Index	$IRECI = (B7 - B4) * \left(\frac{B6}{B5}\right)$
LAISAVI	Leaf Area Index Soil Adjusted Vegetation Index	$LAISAVI = \frac{-\log\left(\frac{0.371 + 1.5 * (B8 - B4)}{(B8 + B4 + 0.5)}\right)}{2.4}$
MCARI	Modified Chlorophyll Absorption in Reflectance Index	$MCARI = \frac{1 - 0.2 * (B5 - B3)}{(B5 - B4)}$
MIRBI	Mid-Infrared Burn Index	$MIRBI = (10 * B12) - (9.8 * B11) + 2$
mNDVI705	Modified NDVI 705	$mNDVI705 = \frac{(B6 - B5)}{(B6 + B5 - 2 * B2)}$
MSI	Moisture Stress Index	$MSI = \frac{B11}{B8A}$
MSR	Modified Simple Ratio	$MSR = \frac{(B7 - B2)}{(B4 - B2)}$
MTCI	Meris Terrestrial Chlorophyll Index	$MTCI = \frac{(B6 - B5)}{(B5 + B4)}$

Índice (Sigla)	Nome (Descrição)	Fórmula
NBR	Normalized Burn Ratio	$NBR = \frac{(B12 - B8A)}{(B12 + B8A)}$
NBR2	Normalized Burn Ratio 2	$NBR2 = \frac{(B11 - B12)}{(B11 + B12)}$
NBRRAW	Normalized Burn Ratio Raw	$NBRRAW = \frac{(B8 - B12)}{(B8 + B12)}$
NDI45	Normalized Difference Index 4/5	$NDI45 = \frac{(B5 - B4)}{(B5 + B4)}$
NDII	Normalized Difference Infrared Index	$NDII = \frac{(B8A - B11)}{(B8A + B11)}$
NDVIG	Normalized Difference Vegetation Index Green	$NDVIG = \frac{B3 * (B8 - B4)}{(B8 + B4)}$
NDVI2	Normalized Difference Vegetation Index 2	$NDVI2 = \frac{B8A - B4}{B8A + B4}$
NDVI705	Normalized Difference Vegetation Index 705	$NDVI705 = \frac{(B6 - B5)}{(B6 + B5)}$
NDMI	Normalized Difference Moisture Index	$NDMI = \frac{(B8 - B11)}{(B8 + B11)}$
NDSWIR	Normalized Difference SWIR Index	$NDSWIR = \frac{(B11 - B8A)}{(B11 + B8A)}$
NDWI	Normalized Difference Water Index	$NDWI = \frac{(B3 - B8)}{(B3 + B8)}$
PSSRA	Plant Senescence Reflectance Index	$PSSRA = \frac{B7}{B4}$
PSRI	Plant Senescence Reflectance Index	$PSRI = \frac{(B4 - B2)}{B5}$
PSRINIR	Plant Senescence Reflectance Index NIR	$PSRINIR = \frac{(B4 - B2)}{B8}$
OSAVI	Optimized Soil Adjusted Vegetation Index	$OSAVI = \frac{(1 + 0.16) * (B7 - B4)}{(B7 + B4 + 0.16)}$
RVI	Ratio Vegetation Index	$RVI = \frac{B4}{B8}$
RDVI	Renormalized Difference Vegetation Index	$RDVI = \frac{(B7 - B4)}{(B7 + B4)^{0.5}}$
RENDVI	Red-Edge Normalized Difference Vegetation Index	$RENDVI = \frac{(B8 - B6)}{(B8 + B6)}$
RENDWI	Red-Edge Normalized Difference Water Index	$RENDWI = \frac{B7 - B5}{B7 + B5}$
RGRI	Red Green Ratio Index	$RGRI = \frac{(B4 + B5)}{(B2 + B3)}$
S2REP	Sentinel-2 Red Edge Position	$S2REP = \frac{705 + 35 * (0.5 * (B8 + B5) - B6)}{(B7 - B6)}$
SAVI	Soil Adjusted Vegetation Index	$SAVI = \frac{1.5 * (B8 - B5)}{(B8 + B5 + 0.5)}$
SELI	Sentinel-2 Enhanced Vegetation Index	$SELI = \frac{(B8A - B5)}{(B8A + B5)}$
SIPI	Structure Insensitive Pigment Index	$SIPI = \frac{(B8 - B2)}{(B8 - B4)}$
SR	Simple Ratio	$SR = \frac{B8}{B4}$

Índice (Sigla)	Nome (Descrição)	Fórmula
TCARI	Transformed Chlorophyll Absorption in Reflectance Index	$TCARI = 3 * (B5 - B4) - 0.2 * (B5 - B3) * \left(\frac{B5}{B4}\right)$
TVI	Triangular Vegetation Index	$TVI = \frac{(120 * (B6 - B3)) - (200 * (B4 - B3))}{2}$